INNOVATION & RESEARCH CAUCUS

PREDICTING BUSINESS GROWTH:

A REVIEW OF BEST PRACTICE ECONOMETRIC AND MACHINE LEARNING APPROACHES

IRC Report No: 046

REPORT PREPARED BY

Rita Nana-Cheraa
Warwick Business School

Michalis Papazoglou
Oxford Brookes University

Stephen Roper
Warwick Business School

ERC Enterprise Research Centre

UKRI

Delivered with ESRC and Innovate UK

CONTENTS

Authors

The core members of the research team for this project were as follows:

- Rita Nana-Cheraa (PI) – Warwick Business School
- Michalis Papazoglou – Oxford Brookes University
- Stephen Roper – Warwick Business School

This document relates to IRC Project IRCP0041: Predictive models of business growth - Testing potential applications to the Cell and Gene Therapies Sectors

Acknowledgements

This work was supported by Economic and Social Research Council (ESRC) grant ES/X010759/1 to the Innovation and Research Caucus (IRC) and was commissioned by Innovate UK (IUK). We are very grateful to the project sponsors at UK Research & Innovation (UKRI) for their input into this research. The interpretations and opinions within this report are those of the authors and may not reflect the policy positions of IUK.

We would also like to acknowledge and appreciate the efforts of the IRC Project Administration Team involved in proofreading and formatting, for their meticulous attention to detail and support.

About the Innovation and Research Caucus

The Innovation and Research Caucus supports the use of robust evidence and insights in UKRI's strategies and investments, as well as undertaking a co-produced programme of research. Our members are leading academics from across the social sciences, other disciplines and sectors, who are engaged in different aspects of innovation and research systems. We connect academic experts, UKRI, IUK and the (ESRC), by providing research insights to inform policy and practice. Professor Tim Vorley and Professor Stephen Roper are Co-Directors. The IRC is funded by UKRI via the ESRC and IUK, grant number ES/X010759/1. The support of the funders is acknowledged. The views expressed in this piece are those of the authors and do not necessarily represent those of the funders.

About the Enterprise Research Centre

The Enterprise Research Centre (ERC) is an independent research centre based at Warwick Business School focusing on growth, innovation and productivity in small and medium-sized enterprises. The Centre is funded by the Economic and Social Research Council, The Department for Business and Trade, The Department for Science Innovation and Technology, Innovate UK, the British Business Bank and the Intellectual Property Office. The views expressed in this report are those of the authors and do not necessarily represent those of the funders.

Contact

You are also welcome to email us if you have any questions about this report or the work of the IRC generally: info@ircaucus.ac.uk

Cite as: Nana-Cheraa, R., Papazoglou, M. and Roper, S. January 2026. *Predicting business growth: A review of best practice econometric and machine learning approaches*. Oxford, UK: Innovation and Research Caucus

Executive Summary

The difficulty of forecasting business growth has long engaged economists and scholars in business and management. Yet, empirical research has found it hard to pinpoint consistent growth drivers, with most models showing very low predictive accuracy. This unpredictability stems from the ongoing heterogeneity of firms and the fact that variations across industries, technologies, and countries make generalisation challenging.

While some stylised facts exist, such as the tendency for younger and smaller firms to grow faster, conventional econometrics models often fail to explain future growth. ML techniques, however, offer new opportunities by processing high-dimensional and unstructured data sources (e.g., financial reports, web content) to uncover hidden relationships. In this review, we provide an accessible overview of the latest developments in modelling business performance.

Econometric models

Econometric models typically use a deductive approach, guided by theoretical or conceptual frameworks that develop testable hypotheses about what affects business growth. Most studies look at various growth indicators, including employment, sales, productivity, assets, exports, and profitability. Employment growth is the most frequently analysed metric in the 19 studies considered here (12 studies), followed by sales and labour productivity growth (6 studies each), and total factor productivity (TFP) and asset growth (3 studies each).

Growth predictors encompass a broad range of areas: human and knowledge capital, innovation, R&D, support, leadership and governance structures, financial resources and access to credit, market conditions, institutional environments, and policy and regulatory frameworks.

Although OLS and panel models are the most common methodological approaches, they generally exhibit lower predictive power than alternative methods. For example, OLS estimations often yield R-squared values below 0.09. Panel estimations show greater variation, with R-squared values ranging from 0.026 and adjusted R-squared values between 0.017 and 0.304, highlighting differences in explanatory power across various contexts and specifications.

When analysing individual studies, the Difference-in-Differences approach combined with Propensity Score Matching (PSM) provides the highest predictive power, with adjusted R-squared values ranging from 0.88 to 0.98 depending on the growth model specification. Quantile regression also demonstrates strong explanatory ability, with pseudo-R-squared values between 0.68 and 0.80.

Machine Learning (ML) and AI-based approaches

ML enables computers to learn from data and enhance their performance on specific tasks by recognising patterns and making predictions or decisions based on experience rather than fixed rules, with little or no human input and without explicit programming. ML algorithms generate predictions by searching data for complex associations between variables.

At a high level, ML is categorised into three main types: supervised learning, unsupervised learning, and reinforcement learning. While supervised learning relies on labelled data to predict outcomes, unsupervised learning detects hidden structures in unlabelled data, and reinforcement learning involves decision-making through feedback from an environment. Each category offers unique analytical and predictive capabilities that serve diverse applications, from financial forecasting to autonomous systems.

Supervised learning (SL) is the most popular ML method and involves training a model on labelled datasets to link input variables with known output variables (Maple et al., 2023). During this process, the model identifies patterns that allow it to predict future or unseen results accurately. SL has been used to forecast company performance, solvency, and overall success by pinpointing the most influential variables affecting outcomes.

For example, supervised ML algorithms have been used to predict which firms will achieve high growth alongside econometric approaches (i.e., Logistic Regression). When both approaches are employed, ML algorithms outperform econometric models in forecasting high-growth firms, demonstrating the predictive power of ML techniques.

Contrasting strengths

Both econometric models and ML techniques aim to learn from data, but they differ in philosophy and purpose. Econometric models, based on statistical theory and economic reasoning, are mainly used for hypothesis testing and causal inference. They rely on predefined theoretical frameworks and assumptions about data distribution, emphasising interpretability and formal inference through confidence intervals and significance tests.

In contrast, ML methods are driven by algorithms and are less limited by theoretical assumptions. Their main aim is predictive accuracy rather than inference, focusing on improving performance through computational learning. While econometrics aims to confirm or refute predefined hypotheses, ML seeks to identify complex, often non-linear patterns in large datasets without relying on assumptions about data distributions or model structures.

Despite its advantages, ML’s focus on predictive performance creates interpretability challenges often called the “black box” problem. (Huang et al., 2024; Valizade et al., 2024). Unlike econometric models, where coefficients give direct insights into the relationships between variables, ML algorithms usually offer limited transparency about how input features affect outcomes. This lack of clarity and interpretability raises concerns, especially for policymakers, managers, and investors, who care not only about prediction accuracy but also about the main factors driving a firm’s potential for high growth.

Essentially, econometric and ML paradigms are complementary rather than mutually exclusive. As both fields evolve, a more integrated, boundary-expanding methodological paradigm is emerging, capable of balancing interpretability with predictive power and blending econometric rigour with ML flexibility to generate more robust, generalisable, and theoretically meaningful insights.

Practical implications

Implementing either econometric or ML approaches involves several specific choices related to the goals of the predictive task, data availability, and transparency. These issues are summarised in the following table:

Criterion	Econometric Approach	ML/AI Approach	Real-World Example
Primary Goal	Hypothesis testing, causal inference	Predictive accuracy, pattern recognition	Econometric: Assessing impact of R\&D grants on SME growth (e.g., Vanino et al., 2019). ML: Predicting high-growth firms using Random Forest (e.g., Houle & Macdonald, 2025).
Interpretability	High (coefficients, significance tests)	Low to medium (often “black box”; explainable AI needed)	Econometric: Quantile regression showing R\&D effects at different growth quantiles (Coad et al., 2016). ML: Neural networks predicting revenue growth but hard to interpret (Houle & Macdonald, 2025).
Data Requirements	Structured, longitudinal/panel data; smaller datasets	Large, high-dimensional, possibly unstructured data	Econometric: Longitudinal Small Business Survey (UK). ML: Web-scraped financial and social media data for firm success prediction.
Assumptions	Strong (distributional,	Minimal; non-parametric, flexible	Econometric: OLS models assuming linearity (Murro et al., 2023). ML: Gradient Boosted Trees handling

Criterion	Econometric Approach	ML/AI Approach	Real-World Example
	linearity, independence)		non-linear interactions (Vuković et al., 2024).
Transparency	High (clear theoretical framework)	Lower (complex algorithms, harder to explain)	Econometric: DiD models for policy evaluation (Mulier & Samarin, 2021). ML: Deep learning for text-based growth prediction (Gangwani & Zhu, 2024).
Computational Demand	Low to moderate	High (requires significant computing resources)	Econometric: Panel regressions on survey data. ML: Neural networks trained on millions of observations.
Predictive Power	Generally low to moderate; better for causal insights	High for out-of-sample prediction	Econometric: R^2 often <0.1 for OLS models. ML: CatBoost achieving 86% accuracy for growth prediction (Vuković et al., 2024).
Theory Integration	Strong (based on economic reasoning)	Weak; primarily data-driven	Econometric: Testing Schumpeterian growth theory. ML: Inductive discovery of patterns without prior theory.
Handling Non-Linearity	Limited (requires transformations)	Strong (captures complex, non-linear relationships)	Econometric: Adding quadratic terms for size effects. ML: Random Forest capturing non-linear effects of age and leverage.
Adaptability to New Data	Limited; model structure fixed	High; models can retrain and adapt	Econometric: Static regression models. ML: Online learning algorithms updating predictions in real time.
Policy Usefulness	High (clear drivers of growth for policy design)	Lower (harder to justify decisions based on opaque models)	Econometric: Evaluating subsidy impacts for innovation policy. ML: Predicting which firms will become high-growth for investment targeting.
Sector-Specific Relevance	Strong if theory tailored	May require retraining for sector-specific patterns	Econometric: Sector-specific productivity models. ML: Industry-specific training for growth prediction in tech vs manufacturing.

Defining the aims of the predictive exercise is essential for selecting between ML and econometric approaches. ML methods may deliver superior predictive accuracy for a given dataset compared to purely econometric methods. However, all ML predictions are subject to the ‘black box’ problem, which means it may not be very clear how or why specific predictions are made. This complicates the use of these predictions to refine related policy initiatives or support measures. Conversely, econometric models— which establish a more explicit link between drivers and growth— offer more direct insights.

Other questions may also be important when examining growth within a specific group of businesses. In such cases, models trained on a broadly based database might be less relevant to particular sectors or firm size bands.

Predicting business growth with either an econometric or ML approach also demands substantial data resources, including growth metrics and potential explanatory or correlated variables for many companies, ideally spanning several years.

Finally, it is important to consider the transparency and persuasiveness of the two modelling approaches. ML methods may be seen as less transparent and possibly less reliable due to the ‘black box’ approach. Econometric methods may be more transparent but can also be challenging to communicate because of their complexity.

1. Introduction

The challenge of predicting business growth has long engaged economists and scholars in business and management. However, new data sources offering more comprehensive coverage of potential growth factors, combined with innovative analytical methods like machine learning, create new opportunities. In this review, we examine recent academic and grey literature that employs formal analytical techniques to forecast business growth. We focus on studies that include in-sample testing and cite other research illustrating the range of variables influencing growth and the various methods used to understand how growth occurs. Our goal is to provide an accessible overview of the latest advances in modelling business performance.

In broad terms, the studies we review fall into three main groups:

- Econometric models – mainly adopt a deductive approach to testing specific hypotheses about the determinants of business growth. Here, potential drivers such as skills, R&D, innovation, and investment are standard. Typically, these studies build on an underlying conceptual framework of the links between specific drivers and growth and often also examine potential moderators of these links.
- Unsupervised machine learning – which uses numerical simulation methods to identify relationships between growth and other business characteristics with no regard to the relationship between indicators or the causal mechanisms linking growth and its drivers — may be regarded as inductive, as it makes no prior assumptions about the interrelation between variables.
- Supervised machine learning, in which numerical simulation approaches are combined with human determination of input and outcome variables, is trained based on a known dataset. For example, ML approaches may be used to estimate a regression model once the dependent and independent variables are selected.

A key element of all these approaches is the quality and scope of the underlying data. Deriving causal mechanisms usually requires longitudinal or panel data on individual enterprises, with potential growth drivers observed before the business growth occurs. Administrative data, which may cover an entire population rather than a survey sample, can be a useful data source. However, these sources are often designed to minimise response burden and therefore frequently lack the key variables believed to influence growth most strongly.

For example, the Inter-Departmental Business Register (IDBR) provides information on the turnover and employment growth of UK firms but does not include data on exporting, skills, or business leadership. Even when data is collected with a clear aim to understand business growth, survey coverage may be limited to increase response rates. The Longitudinal Small Business Survey (LSBS), for instance, includes information on innovation, exporting, training, etc., but offers little insight into the leadership capabilities of respondent companies.

The remainder of this SOTA review is organised as follows:

- Section 2 focuses on econometric models that forecast business growth. Our review is necessarily selective due to the large number of potential studies. We emphasise recent research published in top-tier journals and prioritise those with some in-sample testing of growth forecasts.
- Section 3 explores predictive models that use machine learning, including both supervised and unsupervised approaches. Methodological papers are common and often take a comparative approach to assess the predictive performance of various machine learning techniques.
- Section 4 examines the strengths and limitations of each approach, evaluates the data and computational requirements necessary to implement each type of predictive model, and provides some recommendations for future research and application.

2. Econometric models

2.1 Econometric methodology – deductive/hypothesis testing

Econometric models typically adopt a deductive approach, guided by theoretical or conceptual frameworks that inform testable hypotheses about what affects business growth. These studies use econometric techniques to identify causal or correlational relationships, often supported by robustness checks and alternative model specifications to confirm their validity. Common methods include OLS, instrumental variables (IV), panel and quantile regressions, dynamic panel models (GMM), difference-in-differences (DiD), and propensity score matching (PSM).

Given our interest in predicting growth, we focus on econometric studies which have an element of in-sample predictive testing. Most studies examine various growth metrics, including employment, sales, productivity, assets, exports, and profitability. Employment growth is the most commonly analysed metric in the 19 studies considered here (12 studies), followed by sales and labour productivity growth (6 studies each), and total factor productivity (TFP) and asset growth (3 studies each).

Growth predictors span a wide range of domains:

- Human and knowledge capital (e.g., Loncan, 2025; Grillitsch et al., 2019)

- Innovation, R&D, and support (e.g., Gong et al., 2026; Davydiuk et al., 2024; Mulier & Samarin, 2021)
- Leadership and governance structures (e.g., Harutyun et al., 2025; Aguilera et al., 2024; Von Nitzsch et al., 2024)
- Financial resources and access to credit (e.g., Blickle & Santos, 2024; Murro et al., 2023; Bircan et al., 2020)
- Market conditions and institutional environments (e.g., Jiang et al., 2024; Ilzetzki, 2024)
- Policy and regulatory frameworks (e.g., Wang et al., 2024)

The empirical studies cover a wide range of academic fields, including innovation studies, economics, finance, business, and management. They appear in leading journals such as Research Policy (e.g., Mulier & Samarin, 2021; Vanino et al., 2019; Guarascio & Tamagni, 2019; Grillitsch et al., 2019; Di Cintio et al., 2017; Coad et al., 2016), Journal of International Economics, and American Economic Review (Gong et al., 2026; Ilzetzki, 2024), as well as prominent finance journals like Journal of Financial Intermediation, Journal of Financial Economics, Journal of Corporate Finance, Journal of Financial and Quantitative Analysis, and The Review of Financial Studies (Loncan, 2025; Blickle & Santos, 2024; Davydiuk et al., 2024; Murro et al., 2023; Bircan & De, 2020). Other top-tier outlets include Energy Policy (Wang et al., 2024), Journal of Management (Aguilera et al., 2024), Strategic Entrepreneurship Journal (Von Nitzsch et al., 2024), Journal of Business Venturing (Harutyunyan et al., 2025), Journal of Operations Management (Jiang et al., 2024), and Small Business Economics (Barba Navaretti et al., 2022).

2.3 Methodological approaches

The methodological approaches used by econometric studies are diverse but can be grouped into five main categories:

- **Panel fixed and random effects models** which use longitudinal data on individual firms. These are the most frequently used, appearing in seven studies (Harutyun et al., 2025; Loncan, 2025; Davydiuk et al., 2024; Wang et al., 2024; Blickle & Santos, 2024; Von Nitzsch et al., 2024; Grillitsch et al., 2019). specifications.
- **Ordinary least squares (OLS) regression models** which are readily interpretable but may provide biased results when data are skewed or non-normal in distribution. OLS is **u**sed as a baseline in six studies (Gong et al., 2026; Jiang et al., 2024; Aguilera et al., 2024; Murro et al., 2023; Barba Navaretti et al., 2022; Bircan et al., 2020). OLS models are often supplemented with robustness checks such as alternative specifications, split-

sample analyses, and instrumental variable techniques to address endogeneity and improve reliability.

- **Instrumental variable (IV) and endogeneity-corrected models** are used to assess causality in situations where longitudinal data is not available or endogeneity may create biased estimates. This category includes 2SLS, 3SLS, and GMM estimations, used in studies like Ilzetzki (2024), Bircan et al. (2020), Grillitsch et al. (2019), and Di Cintio et al. (2017). These models aim to correct for simultaneity and omitted-variable bias, with instrument validity and over-identification tests.
- **Quantile regression** allows different effect sizes at different values of a variable such as company size. Employed as the primary method in three studies (Guarascio & Tamagni, 2019; Grillitsch et al., 2019; Coad et al., 2016), to captures variation across the growth distribution by estimating effects at different quantiles.
- **Quasi-experimental and matching designs** exploit matching approaches to harmonise the characteristics of treatment and control groups. Techniques such as propensity score matching (PSM), difference-in-differences (diff-in-diff), and triple-differences (DDD) are used in studies evaluating policy impacts and funding interventions (e.g., Mulier & Samarin, 2021; Vanino et al., 2019). These designs enhance causal inference by balancing treatment and control groups and controlling for confounding factors**.**

2.4 Key lessons from econometric studies with in-sample testing

Although OLS and panel models are the most common methodological approaches, they generally display lower predictive power than other approaches. For example, OLS estimations often report R-squared values below 0.09 (see Jiang et al., 2024; Murro et al., 2023; Grillitsch et al., 2019; Bircan et al., 2020). Panel estimations show more variation, with R-squared values ranging from 0.026 (Von Nitzsch et al., 2024) to 0.626 (Davydiuk et al., 2024), and adjusted R-squared values between 0.017 (Wang et al., 2024) and 0.304 (Loncan, 2025), indicating differences in explanatory power across contexts and specifications.

When analysing individual studies, the Difference-in-Differences approach combined with Propensity Score Matching (PSM) delivers the highest explanatory power, with adjusted R-squared values between 0.88 and 0.98 depending on the growth model specification (Mulier & Samarin, 2021). Quantile regression also shows strong explanatory capability, with pseudo-R-squared values ranging from 0.68 to 0.80 (Coad et al., 2016). However, its effectiveness decreases when used alongside other methods. For example, Grillitsch et al. (2019) applied panel fixed effects, pooled OLS, GMM, and quantile regression to the same dataset. Among these, panel fixed effects produced the highest R-squared (0.337), while quantile regression yielded the lowest (0.02).

Most of the reviewed studies employ layered methodologies. Typically, they go beyond simple OLS models by including fixed effects, instrumental variable techniques, or matching methods, which enhance robustness and address concerns about omitted-variable bias or sample selection issues. Consistently across the studies are robustness and validity checks. Techniques such as endogeneity testing, placebo timing analyses, and alternative model estimation, along with sub-sample evaluations, are routinely used to bolster empirical credibility.

Table 1: Econometric studies with in-sample prediction testing

Study	Country	Growth metric(s)	Growth predictor(s)	Baseline Methodological approach	In-sample test(s) / Robustness check(s)	Main Findings
Loncan, T. (2025)	US	Employment growth	Employee welfare policies (EWPs) Industry sales growth	Panel fixed effect Theoretical framework and hypotheses testing	Significant F-statistics for baseline model; Adjusted R² = 0.283 - 0.304 Endogeneity checks. Robustness check with Dynamic GMM estimation, Alternative specifications, Subsample analysis and Sensitivity analysis.	EWPs effect on firm employment growth is significantly positive firm employment growth; Significant Positive sensitivity of firm employment growth to Industry sales growth; EWPs weakens the effect industry sales exerts employment growth. Thus, insuring workers against fluctuations in employment.

Gong et al. (2026)	China	Export growth	First successful US patent application	OLS regression
				2SLS IV regression
				Hypothesis testing
					Significant F-statistics for baseline models.
					Alternative specifications;
					Subsample analysis;
					Model validity testing.
						US patent approval improves export growth of Chinese firms by 17–21 percentage points over 3 years (with IV estimation) (sub-sample) and by 6-7 percentage points with naïve OLS estimation (full sample).

Harutyun et al. (2025)	Norway	Sales growth Employment growth	Outside board directors (OBDs) experience: industry and directorial experience	Panel fixed effects Theoretical framework and hypotheses testing	Within R² = 0.403 – 0.416 Between R² = 0.277-0.304 Overall R² = 0.037 – 0.069 Significant Coeff. estimates Coarsened Exact Matching estimation for robustness	OBDs with industry experience has immediate positive influence on sales growth, especially in volatile environments. Directorial experience has delayed but positive effects. The combination of industry and directorial experience yields the strongest growth effects.
Blickle & Santos (2024)	USA	Assets, capital Investment and employment growth	Debt Overhang	Panel fixed effect Hypothesis testing	Overall R² = 0.065 – 0.325 depending on specification. Alternative debt overhang measures Quasi-natural experiment using COVID-19	Firms with high debt overhang experience 5–10% lower growth in assets, investment, and employment. Effects persist even among firms with access to credit and investment opportunities.

Study	Country	Growth metric(s)	Growth predictor(s)	Baseline Methodological approach	In-sample test(s) / Robustness check(s)	Main Findings
Davydiuk et al. (2024)	USA	Employment growth Patenting activity	Access to Business Development Company (BDC) funding	Panel regressions with fixed effects Hypothesis testing	R^2 = 0.505 – 0.626 depending on model specification; Alternative models include: Diff-in-diff regression; Triple-difference regressions; Propensity score matching. Parallel trends testing, Placebo tests for timing, Multiple shocks testing, Sub-sample analysis.	BDC-funded firms experience +0.8%–1.2% employment growth and +2% per quarter (~10% increase) in patenting. Managerial assistance boosts employment growth by 0.3%–0.4%
Aguilera et al. (2024)	France, Germany, Italy, Spain and United Kingdom	Total Factor Productivity (TFP) and TFP Growth	Family ownership status Degree of shared control (ownership, leadership, governance) with non-family members	Multivariate regression model using intermediate inputs proxies Theoretical framework and hypotheses testing	Adjusted R^2 values: ~0.56 for TFP models, ~0.02 for TFP growth models; Robustness checks including: extended sample analysis, survival selection bias control, propensity score matching, and two-step GMM with \iv	Family firms are more labour-intensive and less capital-intensive than nonfamily firms. Family firms exhibit lower productivity and productivity growth. Sharing control with non-family members improves productivity and shifts input mix toward capital. A minimum threshold of shared control (~10%) yields significant productivity gains.
Von Nitzsch et al. (2024)	Germany	Sales growth	Owners' experience-based matching competence and Governance-based competences	Random-effects panel regression Theoretical framework and hypotheses testing	Overall R^2 values: ~0.026 Between R^2 values: ~0.069 Within R^2 values: ~0.022 Heckman selection correction; Alternative growth metrics, subsample analysis.	Owners' matching and governance competences positively influence firm growth, especially in younger firms. Governance competence effect is weaker in family firms.

Study	Country	Growth metric(s)	Growth predictor(s)	Baseline Methodological approach	In-sample test(s) / Robustness check(s)	Main Findings
Jiang et al. (2024)	41 Countries	Labour productivity growth	Public utility obstacles (power outages and transportation obstacles) Moderators: National culture indicators	OLS regression model with year and industry fixed effect Theoretical framework and hypotheses testing	Adjusted R² values: ~0.062 Several model validity test Robustness checks with PSM estimation, use of an alternative measure for transportation Obstacles, and industry-year joint fixed effect regressions.	Power outages and transportation obstacles negatively affect labour productivity growth; National culture moderates these effects: (a) Power distance (and uncertainty avoidance amplify the negative impact of power outages. (b) Long-term orientation mitigates the impact of power outages. (c) Individualism and masculinity mitigate the impact of transportation obstacle
Wang et al. (2024)	China	Employment growth	Energy conservation and emission reduction (ECER) targets	Panel data regression Theoretical framework and hypotheses testing	Adjusted R² values: ~0.017 Unit root and cointegration testing; Robustness checks with Diff-in-Diff and IV estimation, sub-group analysis	ECER increases firm-level fixed asset investment and tax burden, and reduces wages—leading to lower employment growth.
Murro et al. (2023)	Italy	Employment growth	Bank-firm relationships	Pooled OLS regression Theoretical framework and hypotheses testing	R² values: ~0.086. Placebo test using lagged employment growth, Matched sample analysis using PSM, IV estimation using historical bank branch distribution. Robustness check using alternative measures of main explanatory variable, Subgroup analysis.	Firms with durable bank relationships are less sensitive to negative sales shocks in employment decisions. Relationship lending acts as liquidity insurance, enabling labour hoarding during temporary downturns. Stronger effects observed in younger and smaller firms, sectors with higher human capital and regions with higher labour market rigidity.

Study	Country	Growth metric(s)	Growth predictor(s)	Baseline Methodological approach	In-sample test(s) / Robustness check(s)	Main Findings
Barba Navaretti et al. (2022)	Germany, France, Italy, Spain, United Kingdom, Austria, and Hungary	Sales, assets and profitability growth	CEO's age (Binary: <45 vs. ≥45)	Cross-sectional OLS regression Theoretical framework and hypotheses testing	Goodness-of-fit values: ~0.085, ~0.044 and ~0.0005 respectively for sales, assets and profit growth models. Alternative specifications using moderators.	Firms managed by young CEOs (under 45 years) grow faster in sales and assets, but not in profitability. Effect is stronger for firms in the higher percentiles of the growth rate distribution.
Di Cintio et al. (2017)	Italian SMEs	Employment growth rate, Hiring rate and separation rate	R&D intensity and exporting	3SLS regression Heckman two-stage selection model T. framework and hypotheses testing	R&D intensity failed exogeneity test under baselined model Quantile estimation using IV Tobit model were significant and passed model validity tests	R&D intensity positively affects employment growth, hiring, and reduces separations. R&D-induced exports negatively affect employment growth and hiring, increase separations.
Coad et al. (2016)	Spain	Sales growth, Productivity growth and Employment growth	R&D Investment	Quantile regression with fixed effects T. framework and hypotheses testing	Pseudo R^2 values: ~0.8016, ~0.6751 and ~0.7677 for sales, productivity, and employment growth model. Robustness checks using split sample analysis.	Larger growth gains at upper quantiles of the growth rate distribution, larger losses at lower quantiles.
Mulier & Samarin (2021)	27 European countries	Tangible and intangible assets growth; Turnover growth; Employment growth Patent stock	Innovation Subsidies (pan-European innovation funding program)	Diff-in-Diff model with propensity score matching. Theoretical framework and hypotheses testing	Adjusted R^2 values: 0.88 – 0.98 for all growth models. Alternative estimation includes: Dynamic Diff-in-Diff models with year-by-year effects and sectoral splits analysis. Several robustness checks	Subsidies increase investment, turnover and employment growth, and patenting; Effects grow over time, especially for intangible assets and patents; Stronger effects in R&D-intensive, knowledge-intensive, and less competitive sectors;
Guarascio & Tamagni (2019)	Spain Manufacturing firms	Sales growth Sales growth persistence	Innovation persistence indicators (R&D, patents, product/process innovation)	Quantile regression with year fixed effects	Robustness checks with GMM panel models and split sample analysis	Innovation persistence has significantly negative effect on sales growth and consistent sale growth; No significant growth premium for persistent innovators; Prior sales growth predicts consistent sales growths among firms in the q20 to q60 and q90 of the distribution.

Study	Country	Growth metric(s)	Growth predictor(s)	Baseline Methodological approach	In-sample test(s) / Robustness check(s)	Main Findings
Grillitsch et al. (2019)	Sweden SMEs	Employment growth	Knowledge base shares: analytical, synthetic, and symbolic)	Panel fixed effects, Pooled OLS, Panel (GMM) and quantile regressions T. framework and hypotheses testing	R^2 value (OLS) = ~0.034 R^2 within (FE) = ~0.337 Wald Chi^2 (GMM) = 64,273 R^2 value (quantile) = ~0.020 Robustness checks of curvilinear relationships	Combinations of all three knowledge bases exerts the strongest effect on growth, followed by a combination of any two and then individual knowledge base; GMM estimates lie between FE and OLS estimates; Stronger effects for high-growth firms; Curvilinear (inverted U-shape) relationship between knowledge base intensity and growth
Vanino et al. (2019)	UK	Employment and turnover growth	R&D Grant receipt, grant size, project characteristics	Propensity score matching Conceptual framework and hypotheses testing	Several robustness checks including kernel matching estimation, different time windows estimation, split-sample analysis, and continuous treatment estimation based on grant size	Public R&D grants positively affect employment and turnover growth; Stronger effects for SMEs and less productive firms; Larger relative grant size yields stronger growth effects; Collaborations with universities and industrially related partners enhance grant impact
Ilzetzki, E. (2024)	USA	Total Factor Productivity (TFP) growth Labour productivity growth	Government purchases, capacity utilization.	Dynamic panel IV regression Hypotheses testing	impulse response estimation. OLS regression Sub-sample analysis	1% government demand shock leads to 0.4% TFP growth. High-utilization plants see 0.28% additional growth. Plants adapt to surging demand by improving production methods, outsourcing, and combating absenteeism, primarily when facing tighter capacity constraints.
Bircan et al. (2020)	Rusia	Total Factor Productivity (TFP); labour productivity growth & Employment growth	local bank branch density	2SLS IV regressions OLS regression Hypotheses testing	F-stats on IVs = 99.3 for both TFP growth \7 Lab. productivity growth model R^2 values (IV and OLS) = 0.05 for both TFP growth \7 Lab. productivity growth; Sample split analysis	Credit access boosts innovation, TFP growth and labour productivity growth; Stronger effects in export-oriented, upstream, and low-agglomeration industries; Innovation and productivity gains concentrated in borrowing firms.

3. Machine Learning and AI-based approaches

Machine Learning (ML) is considered a subfield of artificial intelligence. Although the theoretical foundations of ML began in the 1950s, it has only recently experienced rapid growth, mainly driven by increasing computer processing power, data digitisation, and data storage (Buchanan, 2019; Shrestha et al., 2021).

ML allows computers to learn from data and improve their performance on specific tasks by recognising patterns and making predictions or decisions based on experience rather than fixed rules, with little or no human input and without explicit programming. ML algorithms aim to generate predictions by searching data for complex associations between variables that are unlikely to be random or simply coincidental and can be reproduced by anyone following the same methods. These complex and reliable associations discovered by ML algorithms result from procedures that create models that fit the data (i.e., reducing bias in prediction) while also preventing overfitting (i.e., reducing variance in predictions) (Shrestha et al., 2021).

The modern global economy has begun to recognise the effectiveness of ML techniques in uncovering reliable insights hidden in data and to adopt related technologies. However, among all sectors, it is within the financial sector that these techniques are most widely adopted and actively utilised. Private financial institutions early on embraced these methods, providing examples of how ML can improve financial processes. This is supported by numerous reports published by public or non-private institutions (such as national banks, OECD, and government authorities), which aim to keep pace with ML developments by monitoring the ML landscape in finance and establishing the foundation for regulatory frameworks (OECD, 2021, 2024; U.S. Department of the Treasury, 2024; World Economic Forum, 2025). Particularly in the UK, the Bank of England, the Alan Turing Institute, and the UK government regularly inform all interested parties about developments in artificial intelligence and machine learning within financial services (BoE, 2024; Buchanan, 2019; DSIT, 2023; Maple et al., 2023).

However, within academia, especially in business and management studies, finance is not the only field that has adopted ML techniques (Shrestha et al., 2021; Valizade et al., 2024). Research articles that focus on or rely on ML techniques are common in disciplines such as Operations Research, Organisation Science, Management Science, Innovation, Economics, Strategic Management, and Entrepreneurship. It seems that there is a consensus among researchers that ML techniques are capable of revealing knowledge hidden in data that is difficult to extract using conventional methods alone (i.e., econometrics).

Among the research questions where ML approaches have been applied are those related to firm dynamics, such as firm performance, failure, innovativeness, and growth (Gangwani & Zhu, 2024). In particular, firm growth remains a complex and largely unpredictable phenomenon despite decades of research (Bargagli-Stoffi et al., 2021). Empirical studies have struggled to identify consistent drivers of growth, with most models achieving very low predictive power (Chae, 2024). This unpredictability results from the persistent heterogeneity of firms and from the fact that differences across industries, technologies, and countries make generalisation difficult (Bargagli-Stoffi et al., 2021). While some stylised facts exist (such as the tendency for younger and smaller firms to grow faster), conventional econometrics models and firm-related data often fail to explain future growth (Hyytinen et al., 2023). ML techniques, however, offer new opportunities by processing high-dimensional and unstructured data sources (e.g., financial reports, web content) to uncover hidden relationships.

These approaches could assist investors and policymakers in more effectively identifying and supporting high-growth firms (HGFs). Essentially, investors and venture capitalists undertake the risk of investing in companies at very early stages, often relying on information that cannot accurately predict which ventures will succeed. For instance, Lyonnet and Stern (2024) demonstrated that venture capitalists tend to invest in companies that perform predictably poorly and dismiss those that perform predictably well. They applied machine learning techniques to French administrative data and discovered that factors such as being male, an graduate of an elite school, and based in Paris, tend to disproportionately influence VCs’ decisions compared to their actual significance in predicting venture success. Although predicting the success of start-up companies becomes less reliable as the company's age diminishes, creating natural limitations in predictive accuracy, ML techniques offer tools for the more effective and improved utilisation of available information in forecasting high-growth firms.

At a high level, ML is categorised into three main types: supervised learning, unsupervised learning, and reinforcement learning (Maple et al., 2023). While supervised learning relies on labelled data to predict outcomes, unsupervised learning detects hidden structures in unlabelled data, and reinforcement learning involves decision-making through feedback from an environment. Each category offers unique analytical and predictive capabilities that serve diverse applications, from financial forecasting to autonomous systems.

Supervised learning (SL) is the most popular ML method and involves training a model on labelled datasets to connect input variables to known output variables (Maple et al., 2023). During this process, the model learns patterns that enable it to predict future or unseen results accurately. Common SL techniques include Decision Trees, Random Forest, and Neural Networks, which are used across fields such as finance, marketing, and fraud detection. SL methods can handle both classification tasks, where the result is categorical (e.g., success/failure), and regression tasks, where the result is continuous (e.g., market trends) (Bargagli-Stoffi et al., 2021). In business, SL has been utilised to forecast company performance, solvency, and overall success by identifying the most influential variables that affect outcomes.

In particular, several studies examining the factors with the strongest predictive ability for identifying high-growth firms have used supervised learning techniques applied to datasets compiled from multiple sources. These studies have identified various key determinants of firm growth. Although the specific factors differ across studies — depending on the variables included in each model — the findings consistently highlight the importance of certain variables. These include financial and human capital (Garkavenko et al., 2023), productivity, personnel, and tangible assets (Hyytinen et al., 2023), high profits and investment, alongside low reserves and inventories (Coad & Srhoj, 2020), as well as revenue growth, managerial efficiency, asset investment, and human resource management (Chae, 2024). Overall, these results suggest that, subject to data availability, supervised learning methods can reveal valuable patterns that improve our understanding of the most influential input variables linked to different growth-related outcomes.

Beyond prediction, supervised learning algorithms provide significant methodological benefits. Their nonparametric, data-driven approach allows them to identify complex and nonlinear relationships in large datasets that traditional statistical models might overlook. These algorithms learn decision rules from a training sample and test them on a separate sample, ensuring dependable performance and reducing biases. SL models are particularly effective for predictive analytics, as they optimise accuracy by balancing bias and variance. They offer considerable value in forecasting and decision-making, making them vital tools for organisations aiming for data-driven insights into performance, profitability, and risk.

Table 2 presents the benefits and common applications of the most popular ML algorithms, discussing their suitability for regression (a continuous real-number dependent variable) or classification (a categorical dependent variable) problems, as well as their level of interpretability.

Table 2: Supervised ML algorithms (source: Choudhury et al., 2021)

Algorithm	Regression or Classification	Interpretability	Advantages	Common Usages
Decision tree	Both	High	Highly interpretable due to visualization of tree and variable importance.	Useful for quick understanding of important features and partitions in data
Random forest	Both	Medium	Versatile and generally performs better than decision tree. It is easy to tune and has a low memory footprint. Can also estimate trees in parallel.	General purpose
Neural network	Both	Low	Highly flexible functional form; difficult to tune. More reliable and useful with big data. Generally harder to interpret.	Image recognition, language processing, forecasting
K-nearest neighbors (KNN)	Both	Medium	Lazy nonparametric estimation based entirely from values of K neighboring observations; high memory requirements.	Useful when little is known about the distribution and structure of the data
Gradient boosted tree	Both	Medium	Estimates trees sequentially; often outperforms random forest but harder to tune, slower, and more memory needed.	General purpose high performance; especially good for unbalanced data
Support vector machine (SVM)	Both	Medium	Good for drawing optimal boundaries between linearly separable classes; reliable with relatively few observations and many features.	Image recognition (for example, character recognition) and text categorization
LASSO or ridge	Both	High	Easy to understand and interpret for those with econometrics background. Highly interpretable coefficients.	Simple methods for reducing overfitting and complexity for linear models
Naïve bayes	Classification	Medium	Minimal structure; strongly assumes independence of features so cannot exploit interactions; scalable for large data and reliable with few observations.	Multiclass classification; text classification, such as assigning emails to “spam” or “not spam”

Additionally, in Table 3, we present the main findings of some studies that used supervised ML algorithms to predict high-growth firms, that is, firms classified as high-growth if they achieve at least 20% growth per year over three years (Chae, 2024). In the column Performance Metrics, various performance measures of the algorithms employed by each study are displayed. In classification tasks (e.g., high-growth or non-high-growth firms), the most common performance metrics of the ML algorithms relate to the comparison of correct predictions (positive or negative) versus false ones. For example, the Accuracy indicator measures how often the model correctly predicts the outcome, calculated as the ratio of correct predictions to the total number of predictions. Sensitivity (or True Positive Rate, or Recall) assesses the algorithm's ability to correctly identify true positives (proportion of true positives to actual positives), while Specificity (or True Negative Rate) concentrates on the prediction of true negatives (proportion of true negatives to actual negatives). Finally, Precision indicates how often of the positive predictions are correct (ratio of true positives to total predicted positives) (Bargagli-Stoffi et al., 2021; Vuković et al., 2024).

It is interesting to note that in the first three studies of Table 3, an econometric model (i.e., Logistic Regression) was included together with the ML models to predict high-growth firms. Comparing the models' performances, in all three studies, there was an ML algorithm that performed better than the econometric model in forecasting high-growth firms, revealing the predictive power of ML techniques.

Table 3: Evaluation metrics of the models predicting growth

<table>
<tr><th>Study</th><th>Country</th><th>Growth metric(s)</th><th>Growth predictor(s)</th><th>Methods</th><th>Performance Metrics</th><th>Main Findings</th></tr>
<tr><td>Houle &amp; Macdonald (2025)</td><td>Canada</td><td>High-growth firms based on employment or revenue (HGF)</td><td>Firm characteristics (e.g, size, age, foreign ownership)
Industry
Geography</td><td>Logistic Model
Random Forest
Neural Network</td><td>
<table>
<tr><th><i>Employment</i></th><th><i>Accuracy</i></th><th><i>Sensitivity</i></th><th><i>Specificity</i></th></tr>
<tr><td>Neural Network</td><td>0.719</td><td>0.706</td><td>0.721</td></tr>
<tr><td>Logistic Model</td><td>0.693</td><td>0.702</td><td>0.692</td></tr>
<tr><td>Random Forest</td><td>0.714</td><td>0.677</td><td>0.719</td></tr>
</table>
<table>
<tr><th><i>Revenue</i></th><th><i>Accuracy</i></th><th><i>Sensitivity</i></th><th><i>Specificity</i></th></tr>
<tr><td>Neural Network</td><td>0.698</td><td>0.610</td><td>0.712</td></tr>
<tr><td>Logistic Model</td><td>0.635</td><td>0.672</td><td>0.629</td></tr>
<tr><td>Random Forest</td><td>0.740</td><td>0.651</td><td>0.754</td></tr>
</table>
</td><td>For employment high-growth firms, the neural network performs best with an Accuracy of 71.9 percent.
For revenue high-growth firms, the random forest performs best with an Accuracy of 74 percent.
Industry variables are clearly important for prediction, as are variables that indicate smaller and younger firms.</td></tr>
<tr><td>Hyytinen et al. (2023)</td><td>Finland</td><td>High-growth firms based on revenue</td><td>Firm characteristics (e.g, size, age, productivity, foreign ownership, patents, CEO, export, rating)
Industry characteristics
Geography</td><td>Random Forest
Linear Regression</td><td>
<table>
<tr><th><i>Long-Term Growth</i></th><th><i>Precision</i></th></tr>
<tr><td>Random Forest</td><td>0.386</td></tr>
<tr><td>Linear Regression</td><td>0.279</td></tr>
</table>
</td><td>Random forest approach outperforms linear regression in terms of Precision.</td></tr>
</table>

<table>
<tr><th>Study</th><th>Country</th><th>Growth metric(s)</th><th>Growth predictor(s)</th><th>Methods</th><th>Performance Metrics</th><th>Main Findings</th></tr>
<tr><td>Vuković et al. (2024)</td><td>Russia</td><td>Long-term sales growth
Fast sales growth</td><td>Firm characteristics (e.g., size, age, leverage, ROA)</td><td>Logit Regression
Random Forest
Light GBM
Cat Boost</td><td>
<table>
<tr><th><i>Long-Term Growth</i></th><th><i>Accuracy</i></th><th><i>Precision</i></th></tr>
<tr><td>Cat Boost</td><td>0.8697</td><td>0.667</td></tr>
<tr><td>Logit Regression</td><td>0.868</td><td>0.5</td></tr>
<tr><td>Random Forest</td><td>0.865</td><td>0.4545</td></tr>
<tr><td>Light GBM</td><td>0.8651</td><td>0.4839</td></tr>
</table>
</td><td>Cat Boost achieves the highest scores in evaluation metrics.
Younger firms and those with higher leverage are more likely to grow.</td></tr>
<tr><td>Chae (2024)</td><td>South Korea</td><td>High-growth firms based on employment or revenue</td><td>Firm characteristics (e.g, financial performance, innovation, expansion, strategic alliance, size)
Geography
Industry</td><td>LASSO
Adaptive LASSO
Random Forest</td><td>
<table>
<tr><th><i>Employment</i></th><th><i>Accuracy</i></th><th><i>Sensitivity</i></th><th><i>Specificity</i></th></tr>
<tr><td>Adaptive LASSO</td><td>0.721</td><td>0.7075</td><td>0.7339</td></tr>
<tr><td>LASSO</td><td>0.705</td><td>0.6852</td><td>0.7242</td></tr>
<tr><td>Random Forest</td><td>0.642</td><td>0.6620</td><td>0.6219</td></tr>
</table>
<table>
<tr><th><i>Revenue</i></th><th><i>Accuracy</i></th><th><i>Sensitivity</i></th><th><i>Specificity</i></th></tr>
<tr><td>Adaptive LASSO</td><td>0.704</td><td>0.7164</td><td>0.6907</td></tr>
<tr><td>LASSO</td><td>0.695</td><td>0.6897</td><td>0.7012</td></tr>
<tr><td>Random Forest</td><td>0.687</td><td>0.6897</td><td>0.6877</td></tr>
</table>
</td><td>The best predictive models for the employment and revenue target variables among the three algorithms are adaptive LASSOs.
The study shows the significance of revenue growth, efficiency management, asset investment, and human resource management skills in increasing the chances of becoming an HGF.</td></tr>
</table>

In contrast, unsupervised learning (UL) concentrates on discovering hidden patterns and relationships in unlabeled datasets (Gangwani & Zhu, 2024). UL techniques can group similar data, detect unusual behaviours, or reduce high-dimensional data while preserving important information (Maple et al., 2023). It is commonly used for clustering, association rule mining, outlier detection, and dimensionality reduction. These methods are particularly valuable in business applications such as customer segmentation, fraud detection, portfolio optimisation, and market trend analysis.

For example, clustering methods are popular for grouping entities with similar features into clusters, such as customers with similar behaviours, products with similar profiles, and companies with comparable growth or failure histories. Additionally, association rule mining techniques are preferred for analysing customer purchase behaviour by revealing conditional statement patterns such as frequently purchased items together (“If customers buy the X, then they will buy the Y”), providing valuable business insights for product placement, pricing, and promotional strategies (Gangwani & Zhu, 2024). Another common application of an unsupervised technique is outlier detection in real-time identification of irregularities in the banking industry (usually related to fraud), by detecting among billions of transactions those that deviate significantly from others in the same category and flagging them for further investigation (Maple et al., 2023). Finally, the dimensionality reduction method is often used to improve the performance of predictive business failure models by excluding redundant data with little information about financial features (Gangwani & Zhu, 2024).

Although unsupervised models may be less interpretable and require more computational power than supervised ones, they excel at uncovering hidden structures and dependencies that can inform strategic decisions. Their ability to identify unknown patterns without prior labels makes them a valuable complement to supervised methods in modern data analysis. Table 4 summarises different categories of unsupervised learning approaches, describing common applications and the main algorithms in each category.

Table 4: Unsupervised ML techniques (sources: Gangwani & Zhu, 2024; Maple et al., 2023)

Technique	Usages	Algorithms
Dimensionality reduction	A technique used for dimensionality reduction that transforms high-dimensional data into a lower-dimensional space, while retaining as much of the original information as possible.	Principal Component Analysis (PCA), Isometric Feature Mapping (ISOMAP), Kernal-based Self-organizing Map (KFSOP)
Association rule mining	Used to discover relationship and patterns among variables in large datasets.	Apriori, FP-Growth, Partition Algorithm
Outlier detection	Used to identify irregularities.	Local Outlier Factor (LOF), Fuzzy Logic-based Outlier Detection
Clustering techniques	Used for finding similar features.	k-means clustering, Partition based clustering, Density based clustering, Hierarchical clustering, Model based clustering
Autoencoders	Useful for reconstructing the input data. They are used for tasks such as image denoising and anomaly detection.	Variational Autoencoder (VAE), Convolutional Autoencoder
Generative models	Used to generate new data resembling the input data.	Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs)

Besides supervised and unsupervised methods, machine learning also includes approaches such as reinforcement learning and deep learning. Reinforcement learning involves an agent interacting with its environment to maximise cumulative rewards, often modelled as a Markov Decision Process. It is particularly useful for problems where optimal actions are unknown, such as trading execution and dynamic pricing. Reinforcement learning can be model-based, building an internal environment model for efficient learning, or model-free, which is more flexible and easier to implement. (Maple et al., 2023). Deep learning techniques have increasingly been utilised to predict business success, especially for analysing complex data like text from social media, news, and financial sources. These methods, including CNN (convolutional neural network), LSTM (long short-term memory), and DNN (deep neural networks), automatically learn relevant features without requiring extensive domain expertise, allowing for adaptable prediction models (Gangwani & Zhu, 2024).

4. Summary and conclusions

4.1 Contrasting approaches

Both econometric models and ML techniques share a common goal of learning from data, but they differ in philosophy and purpose (Buchanan, 2019; Valizade et al., 2024). Econometric models, rooted in statistical theory and economic reasoning, are primarily used for hypothesis testing and causal inference. They rely on predefined theoretical frameworks and assumptions about data distribution, emphasising interpretability and formal inference through confidence intervals and significance tests.

In contrast, ML methods are driven by algorithms and are less limited by theoretical assumptions. Their main focus is on predictive accuracy rather than inference, aiming to improve performance through computational learning. While econometrics aims to validate or refute pre-existing hypotheses, ML seeks to identify complex, often non-linear patterns in large datasets, without requiring assumptions about data distributions or model frameworks (Bargagli-Stoffi et al., 2021). This enables an ML model to describe situations it has not previously encountered (Buchanan, 2019). Advances in computing power, data access, and algorithms have sped up the adoption of ML methods across various disciplines, including finance and management.

The difference between econometrics and machine learning becomes especially clear in how each manages data complexity and model evaluation (Valizade et al., 2024). Traditional econometric methods, such as linear regression, are parametric and assume linear or monotonic relationships between variables. In contrast, machine learning is better at capturing non-linear and high-dimensional interactions among features. Techniques like random forests, support vector machines, and neural networks can model complex patterns without requiring explicit distributional assumptions. While econometric models assess validity through statistical significance and in-sample fit, machine learning models are judged based on out-of-sample predictive performance. This emphasis on generalisability allows machine learning to perform well on unseen data, as demonstrated in applications like credit risk prediction, where they have achieved notable improvements in classification accuracy and cost savings.

While econometric models are evaluated based on statistical significance and in-sample fit, machine learning models are assessed according to their out-of-sample predictive performance. For example, in econometric models, the primary metric that indicates how well the model fits the data (goodness of fit) is R-squared (R^2). R^2 measures the proportion of variation in a dependent variable explained by the independent variables, ranging from 0 (no explanation) to 1 (perfect explanation).

Conversely, the most common metrics used in machine learning to evaluate performance relate to their ability to make correct predictions (positive or negative) compared to incorrect ones. Specifically, metrics such as Accuracy (the ratio of correct predictions to the total number of predictions), Sensitivity (the proportion of true positives among actual positives), Specificity (the proportion of true negatives among actual negatives), and Precision (the ratio of true positives to total predicted positives) are frequently employed for assessing ML models (Bargagli-Stoffi et al., 2021; Vuković et al., 2024). Selecting the most suitable metric depends entirely on the context of the analysis, particularly the business problem and the significance of different error types (False Positives versus False Negatives). For example, in fraud detection, where avoiding the blocking of legitimate transactions is vital, achieving high Specificity is essential. Conversely, in scenarios where identifying all actual positives is crucial—such as disease detection—Sensitivity becomes more important.

Despite its advantages, ML's focus on predictive performance presents interpretability challenges often referred to as the "black box" problem (Huang et al., 2024; Valizade et al., 2024). Unlike econometric models, where coefficients provide direct insights into the relationships between variables, ML algorithms typically offer limited transparency regarding how input features influence outcomes (Shrestha et al., 2021). This lack of clarity and interpretability raises concerns, particularly for policymakers, managers, and investors, who care not only about prediction accuracy but also about the key factors driving a firm's potential for high growth (Huang et al., 2024). However, recent advances in interpretable ML and explainable AI have begun to bridge this gap, providing tools to evaluate and visualise variable importance. Furthermore, ML's ability to uncover complex, non-monotonic relationships and identify patterns that traditional models might overlook offers valuable opportunities for theory development and inductive reasoning in management and economics.

In essence, econometric and ML paradigms are complementary rather than mutually exclusive (Shrestha et al., 2021). Econometrics excels in theory-driven, causal explanation, whereas ML offers data-driven, predictive insights. ML can enhance traditional statistical modelling by improving variable selection, managing non-linearity, and supporting algorithm-based induction to identify patterns that inform new theories. Through pattern recognition, data reduction, and inductive reasoning, ML can reveal unexpected or counterintuitive findings that challenge existing theoretical assumptions, fostering a more balanced methodological approach where exploratory quantitative studies stand alongside deductive, hypothesis-testing research (Valizade et al., 2024).

For instance, the design of an econometric model is shaped by deductive reasoning, according to which prior theoretical knowledge and developed hypotheses dictate what kind of variables should be included and how those variables should be used into the model. Therefore, econometric analysis seeks validity for relations among variables in predesigned models. However, interesting relations among variables could be passed unnoticed, if not dictated by prior theoretical knowledge or hypotheses; a scenario whose probability increases in cases where there are many variables and datasets are huge. For example, Choudhury et al. (2021), exploring factors associated with the likelihood of employees to leave their company, demonstrated an interesting relation found by ML techniques between the training performance and the time being in the company, which econometrics techniques failed to capture. Then, this specific knowledge generated by ML techniques led to a rearrangement of the econometric model. In other words, prior theoretical knowledge in combination with the knowledge emerged from ML analysis dictated the most appropriate modelling for explaining the likelihood of employees to leave their company.

Furthermore, algorithm-supported induction can bolster the reproducibility and generalisability of results through out-of-sample validation and the use of non-parametric techniques (Choudhury et al., 2021). However, ML remains an associative rather than causal tool, requiring researchers to interpret results within theoretical frameworks (Choudhury et al., 2021; Garkavenko et al., 2023). As both fields develop, a more integrated, boundary-expanding methodological paradigm is emerging, capable of balancing interpretability with predictive power and merging econometric rigour with ML flexibility to produce more robust, generalisable, and theoretically meaningful insights.

4.2 Practical implications

Implementing either econometric or ML approaches involves several specific choices related to the goals of the predictive task, data availability, and transparency.

Defining the aims of the predictive exercise is essential for selecting between ML and econometric approaches. As some of the studies discussed earlier indicate, ML methods may provide marginally better predictive power for a given dataset than purely econometric methods. However, all ML predictions are prone to the ‘black box’ issue, which means it can be unclear how or why specific predictions are produced. This complicates the use of these predictions to refine related policy initiatives or support measures. Conversely, econometric models— which establish a more explicit link between drivers and growth— offer more direct insight.

Table 5: Decision criteria in choosing between econometric and ML/AI approaches

Criterion	Econometric Approach	ML/AI Approach	Real-World Example
Primary Goal	Hypothesis testing, causal inference	Predictive accuracy, pattern recognition	Econometric: Assessing impact of R\&D grants on SME growth (e.g., Vanino et al., 2019). ML: Predicting high-growth firms using Random Forest (e.g., Houle & Macdonald, 2025).
Interpretability	High (coefficients, significance tests)	Low to medium (often “black box”; explainable AI needed)	Econometric: Quantile regression showing R\&D effects at different growth quantiles (Coad et al., 2016). ML: Neural networks predicting revenue growth but hard to interpret (Houle & Macdonald, 2025).
Data Requirements	Structured, longitudinal/panel data; smaller datasets	Large, high-dimensional, possibly unstructured data	Econometric: Longitudinal Small Business Survey (UK). ML: Web-scraped financial and social media data for firm success prediction.
Assumptions	Strong (distributional, linearity, independence)	Minimal; non-parametric, flexible	Econometric: OLS models assuming linearity (Murro et al., 2023). ML: Gradient Boosted Trees handling non-linear interactions (Vuković et al., 2024).
Transparency	High (clear theoretical framework)	Lower (complex algorithms, harder to explain)	Econometric: DiD models for policy evaluation (Mulier & Samarin, 2021). ML: Deep learning for text-based growth prediction (Gangwani & Zhu, 2024).
Computational Demand	Low to moderate	High (requires significant computing resources)	Econometric: Panel regressions on survey data. ML: Neural networks trained on millions of observations.
Predictive Power	Generally low to moderate; better for causal insights	High for out-of-sample prediction	Econometric: R² often <0.1 for OLS models. ML: CatBoost achieving 86% accuracy for growth prediction (Vuković et al., 2024).
Theory Integration	Strong (based on economic reasoning)	Weak; primarily data-driven	Econometric: Testing Schumpeterian growth theory. ML: Inductive discovery of patterns without prior theory.
Handling Non-linearity	Limited (requires transformations)	Strong (captures complex, non-linear relationships)	Econometric: Adding quadratic terms for size effects. ML: Random Forest capturing non-linear effects of age and leverage.
Adaptability to New Data	Limited; model structure fixed	High; models can retrain and adapt	Econometric: Static regression models. ML: Online learning algorithms updating predictions in real time.
Policy Usefulness	High (clear drivers of growth for policy design)	Lower (harder to justify decisions based on opaque models)	Econometric: Evaluating subsidy impacts for innovation policy. ML: Predicting which firms will become high-growth for investment targeting.
Sector-Specific Relevance	Strong if theory tailored	May require retraining for sector-specific patterns	Econometric: Sector-specific productivity models. ML: Industry-specific training for growth prediction in tech vs manufacturing.

A key uncertainty is how well findings from general sector datasets transfer to different sectors or sub-sectors of the original sector. This may limit the predictive power of models when focusing on the growth potential of a specific group of businesses. Similar considerations might also apply to the age of the firm and forecasting growth. Notably, repeated observations indicate that growth is more variable among younger and smaller firms, which are also more vulnerable to closure. This variability likely makes predicting growth in these firms more challenging than in larger, more established firms.

Predicting business growth using either an econometric or ML approach also requires substantial data resources that include both growth metrics and potential explanatory or correlated variables for a large number of companies, ideally over several years. Suitable business growth data can be obtained from three main sources. First, business survey data enables detailed exploration of specific growth drivers, an approach underlying many econometric models of firm performance. This type of data, usually collected through interviews or business panels, can be expensive to gather, especially when covering multiple years. For example, the Longitudinal Small Business Survey has been conducted annually by BEIS since 2015, covering around 11,000 to 15,000 firms each year with an annual budget of approximately £0.3m. Second, administrative data from Companies House, HMRC, or other government departments can also be used. While such data often covers fewer variables, it may be a more cost-effective alternative to survey data. Some administrative sources provide data on the entire population of firms rather than a smaller survey sample. However, administrative data may also be subject to restrictions that hinder access or use. Finally, data can be unstructured and derived from web scraping. Automatically collected from company websites, this data may offer insights that are otherwise inaccessible, but its completeness and validity are often difficult to verify.

Finally, it is important to consider the transparency and persuasiveness of the two modelling approaches. ML methods may be seen as less transparent and possibly less reliable due to the ‘black box’ approach. Econometric methods may be more transparent but can also be challenging to communicate because of their complexity.

References

Aguilera, R.V., Crespi-Cladera, R., Martín-Oliver, A., Pascual-Fuster, B., 2024. Ownership, control, and productivity: family firms in comparative perspective. Journal of Management, 01492063241259964.

Barba Navaretti, G., Castellani, D., Pieri, F., 2022. CEO age, shareholder monitoring, and the organic growth of European firms. Small Business Economics, 59(1), 361-382.

Bargagli-Stoffi, F.J., Niederreiter, J., Riccaboni, M., 2021. Supervised learning for the prediction of firm dynamics, Data science for economics and finance: Methodologies and applications. Springer International Publishing Cham, 19-41.

Bircan, Ç., De Haas, R., 2020. The limits of lending? Banks and technology adoption across Russia. The Review of Financial Studies, 33(2), 536-609.

Blickle, K., Santos, J.A., 2024. The costs of corporate debt overhang. Journal of Financial Intermediation, 60, 101118.

BoE (Bank of England), 2024. Artificial intelligence in UK financial services. Bank of England.

Buchanan, B.G., 2019. Artificial intelligence in finance. The Alan Turing Institute.

Chae, H.-C., 2024. In search of gazelles: Machine learning prediction for Korean high-growth firms. Small Business Economics 62, 243-284.

Choudhury, P., Allen, R.T., Endres, M.G., 2021. Machine learning for pattern discovery in management research. Strategic Management Journal 42, 30-57.

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

Coad, A., Segarra, A., Teruel, M., 2016. Innovation and firm growth: does firm age play a role?. Research policy, 45(2), 387-400.

Davydiuk, T., Marchuk, T., Rosen, S., 2024. Direct lenders in the US middle market. Journal of Financial Economics, 162, 103946.

Di Cintio, M., Ghosh, S., Grassi, E., 2017. Firm growth, R&D expenditures and exports: An empirical analysis of Italian SMEs. Research Policy, 46(4), 836-852.

DSIT (Department for Science, Innovation and Technology), 2023. A pro-innovation approach to AI regulation.

Gong, R.K., Li, Y.A., Manova, K., Sun, S.T., 2026. Tickets to the Global Market: First US Patent Award and Chinese Firm Exports. Journal of International Economics.

Gangwani, D., Zhu, X., 2024. Modeling and prediction of business success: a survey. Artificial Intelligence Review 57, 44.

Garkavenko, M., Beliaeva, T., Gaussier, E., Mirisaee, H., Lagnier, C., Guerraz, A., 2023. Assessing the factors related to a start-up's valuation using prediction and causal discovery. Entrepreneurship Theory and Practice 47, 2017-2044.

Grillitsch, M., Schubert, T., Srholec, M., 2019. Knowledge base combinations and firm growth. Research policy, 48(1), 234-247.

Guarascio, D., Tamagni, F., 2019. Persistence of innovation and patterns of firm growth. Research Policy, 48(6), 1493-1512.

Harutyunyan, T., Timmermans, B., Frederiksen, L., 2025. Outside board director experience and the growth of new ventures. Journal of Business Venturing, 40(3), 106484.

Houle, S., Macdonald, R., 2025. Identifying and Predicting Nascent High-Growth Firms Using Machine Learning. Canadian Public Policy 51, 41-60.

Hyytinen, A., Rouvinen, P., Pajarinen, M., Virtanen, J., 2023. Ex ante predictability of rapid growth: a design science approach. Entrepreneurship Theory and Practice 47, 2465-2493.

Huang, Y., Xu, S., Lü, L., Zaccaria, A., Mariani, M.S., 2024. Uncovering key predictors of high-growth firms via explainable machine learning. arXiv preprint arXiv:2408.09149.

Ilzetzki, E., 2024. Learning by necessity: Government demand, capacity constraints, and productivity growth. American economic review, 114(8), 2436-2471.

Jiang, S., He, H., Liu, X., Huo, B., 2024. Public utility obstacles and labor productivity growth: The moderating effect of national culture. Journal of Operations Management, 70(6), 904-932.

Lyonnet, V., Stern, L., 2024. Machine Learning About Venture Capital Choices.

Loncan, T., 2025. Can employee welfare policies insure workers against fluctuations in employment? Journal of Corporate Finance, 94, 102849.

Maple, C., Szpruch, L., Epiphaniou, G., Staykova, K., Singh, S., Penwarden, W., Wen, Y., Wang, Z., Hariharan, J., Avramovic, P., 2023. The AI revolution: Opportunities and challenges for the finance sector. The Alan Turing Institute.

Mulier, K., Samarin, I., 2021. Sector heterogeneity and dynamic effects of innovation subsidies: Evidence from Horizon 2020. Research Policy, 50(10), 104346.

Murro, P., Oliviero, T., Zazzaro, A., 2023. Relationship lending and employment decisions in firms' bad times. Journal of Financial and Quantitative Analysis, 58(6), 2657-2691.

OECD, 2021. Artificial Intelligence, Machine Learning and Big Data in Finance.

OECD, 2024. Regulatory approaches to Artificial Intelligence in finance.

Shrestha, Y.R., He, V.F., Puranam, P., von Krogh, G., 2021. Algorithm supported induction for building theory: How can we use prediction models to theorize? Organization Science 32, 856-880.

U.S. Department of the Treasury, 2024. Artificial Intelligence in Financial Services.

Valizade, D., Schulz, F., Nicoara, C., 2024. Towards a paradigm shift: How can machine learning extend the boundaries of quantitative management scholarship? British Journal of Management 35, 99-114.

Vanino, E., Roper, S., Becker, B., 2019. Knowledge to money: Assessing the business performance effects of publicly-funded R&D grants. Research policy, 48(7), 1714-1737.

Von Nitzsch, J., Bird, M., Saiedi, E., 2024. The strategic role of owners in firm growth: Contextualizing ownership competence in private firms. Strategic Entrepreneurship Journal, 18(3), 553-581.

Vuković, D.B., Spitsin, V., Bragin, A., Leonova, V., Spitsina, L., 2024. Forecasting firm growth resumption post-stagnation. Journal of Open Innovation: Technology, Market, and Complexity 10, 100406.

Wang, H., Chen, H., Zhu, L., Yin, J., 2024. “Hidden price”: Energy conservation and emission reduction targets and employment growth. Energy Policy, 189, 114135.

World Economic Forum, 2025. Artificial Intelligence in Financial Services.

Now that you have read our report, we would love to know if our research has provided you with new insights, improved your processes, or inspired innovative solutions.

Please let us know how our research is making a difference by completing our short feedback form via this link.

You are also welcome to email us if you have any questions about this report or the work of the IRC generally: info@ircaucus.ac.uk

Thank you

The Innovation & Research Caucus

www.ircaucus.ac.uk

Email info@ircaucus.ac.uk Twitter @IRCaucus

INNOVATION &
RESEARCH
CAUCUS

Delivered with
ESRC and
Innovate UK