Machine Learning About Venture Capital Choices

Victor Lyonnet Léa H. Stern[*]

18th November, 2024

Abstract

We study early-stage venture capitalists' (VCs) decisions through the lens of a predictive model of venture success. Using French administrative data on VC-backed and non-VC-backed companies, we find that VCs invest in some companies that perform predictably poorly and pass on others that perform predictably well. VCs tend to select entrepreneurs whose features are representative of success – such as being male, graduates of elite schools, and based in Paris. Although entrepreneurs with these characteristics exhibit higher success rates, VCs exaggerate the importance of these features relative to their impact on performance, contributing to the narrowness of the VC industry.

Keywords: Venture Capital, Machine Learning, Stereotypes, Representativeness.

JEL Classification: G11, G24, G41, M13, D83, D8.

[*] For helpful comments, we thank Ajay Agrawal (discussant), Will Cong, Isil Erel, Murray Frank, Nicola Gennaioli, Avi Goldfarb, Jorge Guzman (discussant), Camille Hébert (discussant), Johan Hombert, Chris Hrdlicka, Hyeik Kim, Arthur Korteweg (discussant), Josh Lerner, Clément Mazet-Sonilhac, Isabelle Méjean, Ramana Nanda (discussant), Giang Nguyen (discussant), Gordon Phillips, Matthias Qian (discussant), David Robinson, Daniel Rock, Francesco Saita, Amin Shams, Andrei Shleifer, Michael Weisbach, Karen Wruck, and conference and seminar participants at Boston College, the University of Colorado Boulder, Ohio State, University of Washington, the 2022 RFS-GSU Fintech conference, 2022 Frontiers in Finance Conference, the University of Calgary, the University of Chicago Booth, the University of Maryland, the University of Michigan, 2022 Private Equity Research Oxford Symposium, AI & Big Data in Finance Research Forum, 2022 NBER SI Entrepreneurship, HEC Paris, the University of Mannheim, Bocconi University, the 2022 Carey Finance Conference, CKGSB, and the Virtual Corporate Finance seminar series. Samin Jalali, George Nurisso, Boyu Wang, Daisy Wang, and especially Danial Salman provided excellent research assistance. We are grateful to Wendy Hu, Greg Brown, and IPC/PERC for providing access to MSCI-Burgiss data. We also thank Matthieu Brun and Alexandre Gazaniol for providing access to Bpifrance data. Victor Lyonnet is with the University of Michigan (Email: vlyonnet@umich.edu). Léa H. Stern is with the University of Washington (Email: leastern@uw.edu). A previous version of this paper was distributed with the title "Venture Capital (Mis)Allocation in the Age of AI". All errors are our own.

1 Introduction

Venture capital is the dominant source of financing for high-growth startups (Lerner and Nanda, 2020). Out of hundreds of thousands of new ventures founded each year in the U.S., the average venture capital firm (VC) considers two hundred ventures and invests in only four (Gompers et al., 2020). In an intensely competitive market where success hinges on their ability to evaluate companies' potential by predicting future performance, VCs consider deal selection a crucial determinant of their investment returns. This task is particularly challenging for new ventures due to the scarcity of historical data and the complex set of entrepreneur and startup characteristics to consider (Kerr, Nanda and Rhodes-Kropf, 2014). Which startups have the highest chance of success? Do VCs back the most promising entrepreneurs, or do their choices reveal patterns of systematic errors?

A key obstacle to answering these questions and analyzing VCs' choice behavior is the difficulty of observing VCs' choice set and outcomes for non-selected investments. To overcome this obstacle, we use French administrative survey data on 121,936 new companies from four cohorts of entrepreneurs who founded a new company between 1998 and 2010, which we merge with administrative tax data. Unlike typical VC research datasets, these data allow us to observe detailed information on over one hundred features of entrepreneurs and venture characteristics at creation, both for VC-backed and non-VC-backed companies, and to track all companies' performance over time.

Given the heterogeneous determinants of venture success, theory cannot guide the choice of a model for ex ante predictions of new venture performance. The importance of some covariates may vary across cases, and their interactions are often nonlinear. To avoid relying on parametric assumptions, we use supervised machine learning (ML) methods to predict venture performance. We use entrepreneur and venture characteristics at inception as input features to train a gradient boosting algorithm. The algorithm is trained to predict various measures of companies' future performance in the first three cohorts of entrepreneurs and is evaluated out-of-sample in the last cohort, which is our test set. Input features are restricted to those readily available to a VC who would conduct a first-pass evaluation of the company. All results pertain exclusively to the algorithm's predictive accuracy in the test set, which is left untouched during training. We also evaluate the algorithm's predictive accuracy in a "pure hold-out" sample (the 2014 cohort of entrepreneurs).[1]

The model successfully predicts the distribution of outcomes out of sample, including in the right tail of the distribution, which is of particular interest to VCs. We show that the algorithm's

[1] We obtained access to the realized performance of the 2014 cohort of entrepreneurs only after training the model and completing a first draft of the paper.

predictive capacity does not depend on which venture performance measure the model predicts (e.g., revenue at age 5 or 7, an indicator of the right tail of the outcome distribution at age 5 or 7, an IPO or M&A exit, or imputed exit multiple and valuation), which cohorts the algorithm is trained on, or using random instead of cohort-based split to construct the training and test samples.[2] Even in the upper tail of predicted performance, the distribution of companies' realized performance aligns with our algorithm's prediction.

We observe significant alignment between VCs' early-stage investment decisions and the algorithm's evaluations of new ventures.[3] For example, the median VC-backed company in our test set ranks at the 83rd percentile of the distribution of predicted exit probability. This alignment suggests that the model's objective function largely captures what VCs are optimizing for. However, despite this concurrence in evaluations, our analysis shows that VCs invest in some new companies that perform predictably poorly and pass on others that perform predictably well. We refer to these patterns as investment mistakes or errors. Under some assumptions, we quantify the costs associated with these two types of errors for VCs. Consistent with Strebulaev and Dang (2024), we find that the cost of the first type of error is dwarfed by the cost of missing out on top performers – "home runs matter; strikeouts don't". Using imputed multiples on invested capital (MOIC), we find that eliminating the bottom 10% of portfolio companies in terms of predicted performance would increase investors' imputed MOIC by 9%. More strikingly, if VCs were to invest exclusively in the top 1% of companies with the highest predicted performance, they would increase their imputed MOIC by almost 200%.

A natural explanation for why some predictably good performers were passed on by VCs is that these companies were not suitable for VCs for supply and/or demand reasons. Since the data does not indicate whether the best predicted performers sought and were denied venture capital, we cannot completely rule out that VCs were unable to select them. However, we leverage the richness of the data to show that our results remain robust after accounting for supply and demand considerations. First, the entire analysis is restricted to companies in industries that receive VC-backing in the data. Second, there exist high predicted performers without VC-backing even among those companies that are most prone to seeking and receiving VC backing: new ventures that are

[2] In our main analysis, our algorithm identifies the best predicted performers in terms of revenue, which allows us to train on *all* new companies regardless of their VC-backing status. Return measures used by VCs (e.g., MOIC, TVPI, IRR) cannot be observed for ventures that are not VC-backed. See the framework in Section 2 and Section 4.1 for a detailed discussion on the choice of objective function.

[3] Failure to secure early-stage financing often forces entrepreneurs to cease operations, highlighting the critical role of early-stage venture capital in supporting the entrepreneurial ecosystem, innovation, and startup growth (Kerr, Nanda and Rhodes-Kropf, 2014).

financially constrained, have growth expectations, bring an innovation or a novel idea, want to hire, and/or operate in the same industries and locations as VC-backed companies. The realized performance gap between the best predicted performers and VC-backed companies narrows when the best predicted performers are restricted to VC-prone companies, but it remains sizable. For instance, investing in the best predicted performers that operate in the same industries and locations as VC-backed companies would increase VCs' imputed MOIC by 50%. Finally, we find that the algorithm predicts the distribution of outcomes well even within the set of VC-backed companies. Even among this set, for which by revealed preference supply and demand concerns are absent, the algorithm can identify predictably good and bad new companies.

While designing an algorithmic decision aid for VCs would require addressing concerns such as feature manipulation or the use of restricted variables such as gender or race (see, for example Fuster et al., 2022), this is not our objective. Instead, we use predictive methods to understand VCs' observed decisions, focusing on how early-stage VCs, who often have access to very limited "hard" information (Mullainathan, 2002) and frequently rely on "gut investment decisions" (Gompers et al., 2020), form expectations of entrepreneurial success.

We explore whether VCs' decision-making process explains the two types of errors we identify. We start by constructing a separate algorithm that predicts whether each venture is VC-backed. This model of VCs' decisions achieves an AUC of .77 and as high as .6 when restricted to just three founder demographic features (gender, age, and education). In a regression framework, we show that the strong predictability in VCs' decisions remains after controlling for the venture's predicted performance. This finding suggests that VCs' decisions diverge from those based solely on predicted performance.

A key novelty of our analysis is showing that VCs do not correctly weigh all entrepreneur characteristics in their decisions. Building on the methodology developed in Mullainathan and Obermeyer (2022), our approach departs from Becker-like tests that would interpret ex-post realized performance between VC-backed and non-VC-backed entrepreneurs as ex-ante patterns of errors in VCs' decisions. Instead, we leverage ML predictions to account for the predictive power of entrepreneur and venture characteristics on future performance, capturing potentially complex, interacted, and non-linear relationships. Controlling for predicted performance, we isolate the role of each individual characteristic on VCs' decisions over and above their effect on predicted performance. Crucially, this methodology enables us to assess and quantify how much certain characteristics disproportionately influence VCs' choices relative to their impact on predicted performance. Our analysis reveals that VCs tend to overweight several entrepreneur features. For instance, we find that compared to the unconditional backing rate, male founders are 1.4 times more likely to be VC-backed, graduates of elite schools are three times more likely, and founders based in Paris are two times more likely to receive VC than justified by the actual impact of these characteristics on exit predictions (again, accounting for interactions and non-linearities).

Why do VCs overweight certain founder features despite their strong financial incentives to identify and invest in the best companies? We know that the right tail of the outcome distribution is of particular interest to VCs (Mallaby, 2022). Consistent with this focus on home runs rather than average performance, we find that VCs overweight the features that are representative of success, that is, features that occur more frequently among the best performing entrepreneurs relative to others (Tversky and Kahneman, 1974; Bordalo et al., 2016). Entrepreneurs with these characteristics do exhibit higher success rates than others, but less so than would justify the rates at which they receive VC backing.

Finally, we test whether our ML-derived measure of feature exaggeration is rooted in distortions in estimated odds of entrepreneur success (Bordalo et al., 2023; Rambachan, 2024). Although suggestive, our findings suggest that VCs' exaggeration of certain entrepreneur features is linked to distorted beliefs about the conditional distribution of outcomes. The results are consistent with the belief formation literature and further support the idea that VCs mispredict success rates for certain types of entrepreneurs due to their beliefs being driven by the representativeness heuristic.

This paper contributes to the literature that seeks to understand VCs' investment decisions, building on studies of VCs' investment analyses (Kaplan and Strömberg, 2004), surveys (Gompers et al., 2020, 2021), deal flows (Jang and Kaplan, 2023), and voting decisions (Malenko et al., 2021). While we acknowledge that the French VC market differs from the US market in many ways, the French administrative data provide an unparalleled opportunity to observe detailed information on over one hundred features of entrepreneurs and venture characteristics for a large number of new companies and to track all companies' performance over time. Combined with our ML approach, these data enable us to predict venture performance for all new companies without relying on parametric assumptions, capturing potentially complex, interacted, and non-linear relationships between entrepreneur and venture characteristics and realized performance. This approach not only identifies but also quantifies and explains errors in VC investment decisions, providing insights that are difficult to obtain with more limited datasets typically used in VC research. Our paper fills a gap in the literature emphasized in Lerner and Nanda (2020): "We understand that early-stage investors rely heavily on signals of entrepreneur quality (Bernstein, Korteweg and Laws, 2017), but know very little as to whether the emphasis on these signals is efficient."

Recent U.S. studies help alleviate external validity concerns by showing that our findings align with broader patterns in VC decision-making. Davenport (2022) examines U.S. Pitchbook data and finds that some VC-backed companies have predictably bad performance. Our paper shows that VCs also pass on predictably good performers, who are also well-suited for VC, and that VCs' decisions are consistent with stereotypical thinking. Similarly, Jang and Kaplan (2023) shows that an early-stage U.S. VC firm is skilled at identifying strong startups but overweights the founding team in its investment decisions. Our study extends these findings by showing whether, how much, and why each entrepreneur and startup characteristic influences VCs' early-stage investment decisions. We thoroughly examine the French institutional context, highlighting both its unique features and the comparability of key metrics – such as the share of female VC-backed founders and elite-educated founders between France and the US. Finally, data from MSCI-Burgiss confirm that U.S. VCs tend to favor stereotypical success factors, such as California-based or IT companies.

The existing literature has documented evidence consistent with discrimination or biased preferences (Ewens, 2023), such as differential VC-backing rates and outcomes across entrepreneurs of different gender (e.g., Raina, 2019; Balachandra et al., 2019; Ewens and Townsend, 2020; Gornall and Strebulaev, 2020; Hu and Ma, 2021; Calder-Wang and Gompers, 2021; Hebert, 2023), race (Cook, Marx and Yimfor, 2023; Fairlie, Robb and Robinson, 2022), network (Hochberg, Ljungqvist and Lu, 2007; Howell and Nanda, 2019; Gompers et al., 2020), and location (Chen et al., 2010; Bryan and Guzman, 2021). Our paper contributes to this literature in two fundamental ways. First, we leverage ML methods to account for the complexity of the relationship between attributes and outcomes. Our methodology relies on an ex-ante approach to identify potential errors, which accounts for the uncertainty at the time of decision-making and prevents inferring biases solely based on realized outcomes, which would be appropriate if the decision-maker had perfect foresight. Second, ML methods allow us to show that VCs do not correctly weigh all entrepreneur characteristics in their decisions (Mullainathan and Obermeyer, 2022). Controlling for their effect on predicted performance, we find that certain characteristics disproportionately influence VC investment decisions, such as gender, attendance at elite schools, and being in Paris. Despite their incentives to maximize returns, VCs are human and make decisions based on imperfect information under high uncertainty. This leads to "kernel of truth" stereotypes that overemphasize traits associated with success (Bordalo et al., 2016). Such cognitive bias can hinder experimentation in the VC industry

(Kerr, Nanda and Rhodes-Kropf, 2014) and contribute to its overall narrowness (Lerner and Nanda, 2020).

Finally, our paper is related to recent work that studies the adoption of data-driven approaches by VCs. Röhm, Bick and Boeckle (2022) find that VCs use data-driven methods but have not yet adopted artificial intelligence (AI) methods in their investment decisions. Bonelli (2023) finds that data-driven VCs are better at avoiding startups that fail, though less likely to pick home run deals. Our findings also contribute to efforts that use ML tools to predict new companies' potential (e.g., Ferrati, Muffatto et al., 2021; Te et al., 2022; Żbikowski and Antosiuk, 2021).

2 Framework

We propose a simple two-period model of VCs' investment decisions to formalize our approach. At $t = 0$, each new company i is created and is characterized by a set of features (x_i, z_i). VCs observe (x_i, z_i), but only x_i is recorded in the data, so that features z_i represent private information or characteristics unobservable to the econometrician. Upon observing (x_i, z_i), VCs form conditional expectations of company outcomes at $t = 1$. Denoting y_i as the unknown company outcome at $t = 1$, VCs' expectation at $t = 0$ is $E[y_i|x_i, z_i]$.

Investment Policy. At $t = 0$, VCs choose an investment policy h that specifies for each potential portfolio company i in their investable pool, $\mathcal{D}$, whether to invest or not. The policy h also determines the total number of ventures the VCs will invest in:

$$h \in \{0,1\}^{|\mathcal{D}|} \; and \; \|h\|_0 = N. \tag{1}$$

VCs chose an investment policy h in (1) to maximize their expected payoffs $\pi(h)$,

$$\pi(h) = \sum_{i \in \mathcal{D}} h_i E[r_i|h], \tag{2}$$

where r_i represents VCs' investment returns from investing in company i and is a function of the company's outcome, y_i, which materializes at $t = 1$ when the VC exits.

VCs' Optimal Policy. We define VCs' optimal policy h^* as the investment policy that maximizes their payoff by investing in the top $s\%$ of companies:

$$h_i^* = 1 \text{ iff } R(x_i, z_i) > 1 - s, \tag{3}$$

where $R(x_i, z_i)$ is the percentile rank of company i in the distribution of conditional expected returns for companies in $\mathcal{D}$. VCs do not invest below the percentile threshold $1 - s$ such that $\|h\|_0 = N$.[4] Denoting $\Delta(x_i, z_i)$ as the wedge between the optimal policy (3) and VCs' actual policy:

$$h_i = 1 \text{ iff } R(x_i, z_i) > 1 - s + \Delta(x_i, z_i). \tag{4}$$

To test whether VCs' observed policy deviates from the optimal policy (such that $\Delta(x_i, z_i) \neq 0$), we approximate the percentile rank of rational predictions $R(x_i, z_i)$ by estimating a benchmark percentile rank $M(x_i)$ for each company using the performance predictions $\hat{m}(x_i)$ of a supervised machine learning algorithm that takes characteristics x_i as its input vector. We design an algorithmic investment policy:

$$\alpha_i = 1 \text{ iff } M(x_i) > 1 - s. \tag{5}$$

We denote $\mathcal{A}_s$ the set of companies for which $\alpha_i = 1$, i.e. the best predicted performers as identified by the predictive model. We denote $\mathcal{V}_s$ the set of VC-backed companies. We ask whether companies i and j exist such that

$$\begin{cases} i \in \mathcal{A}_s \ , \ i \notin \mathcal{V}_s \\ j \in \mathcal{V}_s \ , \ j \notin \mathcal{A}_s \\ r_i > r_j. \end{cases} \tag{6}$$

In words, we ask whether companies identified as the best predicted performers are not selected by VCs and yet outperform VC-backed companies not identified as best predicted performers, and whether companies not identified as best predicted performers are selected by VCs and yet underperform those identified as better predicted performers.

[4] The threshold s is determined outside our model and depends on VCs' financing and operational constraints.

3 Data

This section describes the data used in this paper. We construct our dataset using a representative survey of entrepreneurs conducted by the French Statistical Office (INSEE) merged with two other administrative datasets on firm creation and operational performance, and nine commercial datasets on M&As, IPOs, and VC investment returns. Parts of the analysis also use MSCI-Burgiss data on US VC investment returns.

3.1 SINE Survey of Entrepreneurs

Our main dataset is a large-scale survey of French entrepreneurs called *Système d'Information des Nouvelles Entreprises*, or *SINE*. The French Statistical Office administers this survey every four years. The questionnaire is sent to entrepreneurs who registered a new company or took over a business in the first semester of the survey year.[5] Our analysis focuses on new businesses, which represent approximately 80% of the surveyed entrepreneurs. Companies are sampled from the exhaustive firm registry using stratified sampling.[6] The business owner is responsible for filling out the survey. The response rate to the SINE survey is very high because the French statistical office oversees the distribution of questionnaires. This high response rate helps ensure that the sample is representative of new businesses in the French economy.[7]

Our sample comprises 121,936 entrepreneurs from four cohorts of entrepreneurs (1998, 2002, 2006, 2010). The survey typically comprises 47 detailed questions (some questions vary slightly across survey waves) about the founder's personal information, including sociodemographics, motivations for starting the business, and future expectations, as well as investing and financing activities, among others. After encoding survey responses, we obtain 462 covariates for each new venture.[8] The questionnaire includes questions about sources of financing, which we use to determine companies' VC-backed status. Since the SINE survey is sent to entrepreneurs who have newly registered a company, our analysis focuses on early stage financing such as seed or series A.[9]

[5] The French Statistical Office included companies created throughout the entire year for the 1998 survey wave to ensure that surveying from the set of entrepreneurs who created a business in the first semester only did not introduce biases.

[6] The strata are defined using the company's headquarters region, industry, and whether it employs salaried staff.

[7] The Statistical Office includes non-respondents in the data, tags them through meta-data, and uses either cold-deck or imputation methods to fill in responses. Our results are robust to excluding data not directly obtained from survey respondents.

[8] Most questions are multiple-choice, and commuting zone locations and two-digit industries generate numerous one-hot encoded variables, leading to 462 covariates from the 47 survey questions. Excluding locations and industries, the dataset contains over one hundred covariates. We provide a description of a subset of these variables in Appendix E.

[9] The 2006 survey wave does not allow for the identification of VC financing. Therefore, we exclude the 2006 cohort

The SINE survey has been used in the existing literature to study entrepreneurship and external financing, exploring the effects of entrepreneurial optimism on financial contracting and corporate performance (Landier and Thesmar, 2008), the role of unemployment insurance in business creation (Hombert et al., 2020), and the gender gap in external financing across male vs female-dominated industries (Hebert, 2023). Landier and Thesmar (2008) were the first, to the best of our knowledge, to use the SINE data and provide details on how the French Statistical Office administers the survey.

The survey design ensures that sampled companies are largely representative of all new companies (excluding the agricultural sector), attenuating important selection concerns. In contrast to most of the VC literature, our sample includes both VC-backed and non-VC-backed companies, allowing for an examination of VCs' investment decisions within a set that is not limited to startups that have successfully raised VC.[10,11] Because some industries are not suitable for VC, we limit the analysis to industries where at least two companies in the training sample received VC funding.

3.2 Other Data Sources

Accounting data. We match data from the SINE survey with accounting data (balance sheet and income statements) extracted from the tax files used by the Ministry of Finance for corporate tax collection purposes. The accounting information is therefore available for virtually all French firms from 1998 to 2015.[12] We observe firm performance at different ages in the tax files.

Firm registry. We use data from the firm registry (*SIRENE*) for the period 1998 to 2015.[12] For each newly created firm, the registry contains the industry the firm operates in based on a four-digit classification system similar to the four-digit SIC. It also provides the firm's legal status

from tests that require identification of VC-backed status.

[10] VC commercial data sets have been shown to be subject to severe reporting biases (see, for example Gompers and Lerner, 2001, for a discussion of how VCs often underreport poorly-performing deals). While most of the literature focuses on commercial datasets of VC-backed companies, exceptions include Jang and Kaplan (2023) who use proprietary data from a venture capital firm to observe funded and non-funded startups, Ewens and Townsend (2020) who use data from crowdsourcing platform AngelList to study early stage investors' biases against women, as well as Chemmanur, Krishnan and Nandy (2011) and Puri and Zarutskie (2012) who both use the Longitudinal Business Database (LBD), a panel data set collected by the US Census Bureau, to identify companies that do and do not receive VC financing. Hebert (2023) uses the SINE survey to study whether the gender gap in entrepreneurs' external financing varies depending on whether the firm's industry is male or female-dominated and finds evidence of context-dependent gender stereotypes. A few other studies examine smaller hand-collected samples of private VC-backed and non-VC-backed companies, though they are limited to certain geographies, time periods, industries, and firm outcomes (e.g., Hellmann and Puri, 2000, 2002).

[11] The survey identifies nonrespondents, providing us with a unique opportunity to mitigate selection concerns, such as the possibility that successful VC-backed outliers might not respond to the survey. We find that our main results are robust to excluding all imputed responses for nonrespondents and that they are consistent in the 2014 and 2018 cohorts of entrepreneurs (unavailable to the algorithm at the time it was designed), alleviating such selection concerns.

[12] Our sample ends in 2015 because our preferred predicted outcome is firm success at age 5, so that we need data until 2015 to compare our predictions for the 2010 cohort to observed realizations.

(e.g., Sole Proprietorship, Limited Liability Corporation, Corporation), the official creation date and geographical location. We use the firm registry to construct a failure dummy equal to one if a firm disappears from the registry, that is, if it does not survive past a given year.

M&A and IPO exits. To identify exits in our sample, we match the French administrative data with data from Pitchbook, CBInsights, Preqin, SDC, VentureXpert, CapitalIQ, Orbis, and Crunchbase. A company outcome is considered an exit if the company was either acquired or went public. Consistent with the limitations of VC commercial datasets reported in Kaplan, Strömberg and Sensoy (2002); Kaplan and Lerner (2016), we find limited overlap among these various datasets. This underscores the importance of incorporating information from multiple sources.

Pitchbook data on exit valuations. Because deal-level returns data is unavailable for the French VC-backed companies in the SINE survey, we use data on exit valuations from Pitchbook. The Pitchbook data is unique in that it reports both the operating performance and exit valuation of startups. This allows us to test the correspondence between the companies' revenue – the measure of performance our algorithm predicts – and their valuations at exit. Keeping French and US VC-backed exits available in Pitchbook results in 350 French companies and 7,593 U.S. companies in our sample.

MSCI-Burgiss data on deal-level returns to investments. The MSCI-Burgiss data often are described as the gold standard for fund-level VC returns (e.g., Kaplan and Lerner, 2016; Brown et al., 2020). MSCI-Burgiss gathers data from the financial reports of general partners whose investors are MSCI-Burgiss clients to provide data on the underlying deal-level information. We use the two measures of returns available in the MSCI-Burgiss deal-level dataset: multiple of invested capital (Total Value to Paid-in capital, or TVPI) and internal rate of return (IRR) at the investment level. Because the MSCI-Burgiss data do not yet make French deals available separately, we use their sample of U.S. deals that are realized and for which the firm location is available. The resulting sample comprises 26,626 deals with an available TVPI, and 19,793 deals with an available IRR.

Bpifrance data on deal-level returns for French deals. We use proprietary data from the Banque Publique d'Investissement (also known as Bpifrance) to validate the SINE data and dampen selection concerns. Bpifrance is a French public investment bank that supports businesses through various financial solutions, including equity investments. It partners with VCs to foster innovation, growth, and competitiveness among French startups. Bpifrance usually represents 15 to 20% of VC funds' subscriptions. The Bpifrance data is constructed from GP reports that contain information on deal-level returns (MOIC). Deal-level MOIC data is available for 357 French deals.

3.3 Descriptive Statistics

Table 1 presents summary statistics for the main outcome measure and a subset of input features using data from the SINE survey, tax files, and the firm registry. Our sample includes all companies in industries that received venture capital,[13] comprising 84,583 entrepreneurs in the training set (1998, 2002, and 2006 cohorts) and 37,353 in the test set (2010 cohort). In the test set, the 5-year survival rate is 66%, with a mean revenue of 160k euros and a 99th percentile of 2.1 million euros at age 5, including failed companies (assigned zero revenue).[14] The typical entrepreneur is 40 years old, and 28% are female. 15% hold graduate degrees, and 6% graduated from elite schools (Grande École). Most entrepreneurs (61%) start ventures in their previous industry, with about one-fifth leveraging customer and/or supplier relationships from prior jobs. New ideas serve as the primary motivation for 16% of entrepreneurs in starting their companies, while about a quarter anticipate expanding their workforce within the next twelve months. 8% of companies are located in Paris, with roughly one-third operating in B2B sectors and two-thirds in B2C markets. On average, new ventures employ 1.6 individuals.

4 Algorithmic Predictions

4.1 Empirical Implementation: Underlying Assumptions

Recall that Equation (3) defines VCs' optimal investment policy h^*. This policy aims to maximize expected returns by investing exclusively in the top s% of companies in terms of predicted performance. Our objective is to identify deviations from this optimal policy by benchmarking VCs' observed choices ($\mathcal{V}_s$) against those made by a predictive model ($\mathcal{A}_s$). This section delineates the assumptions and empirical decisions underlying our analysis.

[13] To avoid potential classification errors, we include only industries with at least two companies in the training sample (1998, 2002, and 2006 cohorts) that received VC backing.

[14] Table B.1 presents additional summary statistics for various company performance measures.

4.1.1 Choice of performance measure

Selective labels. The most widely used performance metrics among funds and investors are the Internal Rate of Return (IRR), the Total Value Paid In (TVPI), and the Multiple on Invested Capital (MOIC) (Harris, Jenkinson and Kaplan, 2014; Gompers et al., 2020; Gompers and Kaplan, 2022). Ideally, we would observe these metrics for all companies – VC-backed and non-VC-backed – to compare returns from VCs' actual portfolio companies against those from the best predicted performers. However, we face a missing data or selective labels problem (Kleinberg et al., 2018; Rambachan, 2024): we cannot observe VCs' counterfactual returns on companies that were not VC-backed. However, while VCs aim to maximize investment returns, the performance of their portfolio companies is a first-order determinant of their investment returns. Therefore, we seek to train our predictive model to learn the conditional distribution of outcomes using a measure of venture success $y_i(x_i, z_i)$ that is highly correlated with VC returns and that is available for all new companies, not just the VC-backed ones.

A key advantage of our data is that we observe all new companies' operational performance from the tax files, regardless of their VC-backed status.[15] We use companies' revenue as the outcome measure in our main analysis for several reasons: exit valuations are routinely calculated as revenue multiples, the probability of an exit or a future financing round is increasing in revenue, and VCs use revenue forecasts to understand how ventures ultimately monetize their product or service (Gompers et al., 2020).[16] Company revenue, therefore, represents a useful quasi-label for our exercise. In our main analysis, we task the algorithm with predicting new companies' revenue 5 years after creation.[17]

[15] The dataset is not subject to survivorship bias. So long as a company is registered, it is in the sample, and tax files are available up until the firm goes out of business. We take several steps to address the concern that some VC-backed companies may have left France, which would lead us to wrongly interpret their lack of sales in France as poor performance. First, we manually checked online that the 2010 VC-backed companies in our sample did not leave France. Second, we verify in Pitchbook that such exits from France are extremely rare and that the companies in Pitchbook that do leave France are not in our data. Finally, we study new companies that vanish from the tax files and find that they are not more likely to be internationally oriented than others.

[16] Besides the impact of revenue on exit valuations, indirect pay for performance from future fundraising motives should incentivize VCs to care about their portfolio companies revenue. Chung et al. (2012) show that both the likelihood of raising a follow-on fund and the size of that fund, if one is raised, is strongly positively related to performance in the current fund and that indirect pay for performance from future fund flows is substantial relative to their direct pay for performance. When general partners (GPs) raise a new fund while the current fund has not yet closed, limited partners (LPs) use information, including current portfolio companies' revenue, as an interim performance signal to evaluate the GP's skills and decide whether to allocate capital to his or her next fund. Ensuring that current LPs invest in their new funds is especially important for GPs due to the informational holdup problem documented in Hochberg, Ljungqvist and Vissing-Jørgensen (2013). Therefore, because portfolio companies' revenue provides a signal to LPs to assess their skills, fund managers are incentivized to care about portfolio companies' revenue beyond the direct relation between portfolio companies' revenue and fund manager compensation.

[17] This horizon matches existing estimates in Gompers et al. (2020) and Brown et al. (2020), and is consistent with the 6.6 average years between the seed and exit rounds we find in the Pitchbook data.

Revenue and Valuation. We use Pitchbook data to build confidence in the fact that a company's revenue is highly correlated with VCs' returns, specifically for high-return companies. Table E.1 focuses on companies at the top of the revenue distribution at exit. Rows 1, 2, and 3 focus on companies in the top 1%, 5%, and 10% of the revenue distribution, respectively. For each set of companies, column 1 shows the percentile rank of the average firm in the distribution of exit valuations, and column 2 shows the percentile rank of the median firm in the distribution of exit valuations. Because exit valuations can differ across sectors, all percentile ranks are calculated at the sector level and then averaged across sectors. Row 1 implies that the average firm in the top 1% of revenues ends up in the top decile in terms of exit valuation (column 1) and that half of the companies in the top 1% of revenue end up above the 96th percentile of exit valuation. Although the percentile ranks of the average and median companies decrease in rows 2 and 3, these companies remain at the top of the distribution of exit valuation.

Exit is endogenous to VC. VCs are incentivized to realize profits for their limited partners by guiding portfolio companies toward initial public offerings or acquisitions. Consequently, the exit measure, reflecting a firm's likelihood of achieving these outcomes, is endogenous to being VC-backed. Using this measure as a prediction target for our algorithm would bias the selection process towards firms typically favored by VCs and poised for exit. This could hinder the identification of high-performing firms that are valuable investment opportunities but are typically overlooked by VCs. Therefore, while we predict company exits in tests that aim at understanding VCs' choices, using exits as the performance measure in our model to identify the best predicted performers would bring back the missing data problem.

Deal terms and exit revenue multiples. As previously explained, the selective labels problem prevents us from predicting returns. Our choice to use company operational performance instead implicitly implies that we abstract away from deal terms considerations. To see why, consider the MOIC (see Davenport, 2022):

$$MOIC_i = \frac{\delta_i * M_s * y_i}{k_i}, \tag{7}$$

where δ_i represents the percentage ownership (accounting for dilution), M_s is the sector level revenue multiple (such that $M_s * y_i$ is the venture's post valuation when y_i is the venture's exit revenue), and k_i is the initial investment. δ_i and k_i together capture deal terms such that VCs' returns are a function of venture performance (size of the pie) and deal terms (split of the pie). Using company revenue y as the performance measure is equivalent to assuming constant deal terms. Practitioners often report that pricing considerations, or the difference between "cheap" vs "expensive" deals in early-stage investing, is not first order: What drives investment returns is investing in the right companies.[18] Using venture revenue as the outcome measure also assumes constant exit revenue multiples. We test the sensitivity of our results to relaxing the constant deal terms assumption in Section 5.1, and to relaxing the constant revenue multiples assumption in Table 2 and Appendix E.2.

4.1.2 Private information and the VC treatment effect

Even though we choose revenue as a predicted outcome measure, which is available for all companies, a version of the selective labels problem persists. When a company i is in $\mathcal{A}_s$ but not in $\mathcal{V}_s$, we do not observe its revenue had it received VC-backing: $(y_i|h_i = 1)$ is missing. A key advantage of our dataset is that we observe a useful quasi-label (Erel et al., 2021): its realized revenue without VC backing $(y_i|h_i = 0)$. This quasi-label offers valuable insights into the best predicted performers who did not receive VC backing. Notably, realized company revenue reflects relevant private information (unobservables z_i) that VCs may observe at the time they invest and that affect outcomes, such as entrepreneur skills not directly captured in our data.

What revenue as quasi-label does not account for, however, is the treatment effect of VC investment; that is, the difference between the missing label $(y_i|h_i = 1)$ and the observed quasi-label $(y_i|h_i = 0)$. To evaluate whether VCs follow the optimal policy outlined in Section 2, we identify non-VC-backed best predicted performers $i \in \mathcal{A}_s$ (but not in $\mathcal{V}_s$) that have higher realized revenue than VC-backed companies $j \in \mathcal{V}_s$. This approach assumes that the (unobserved) VC treatment effect for these best predicted performers is greater than the opposite of the performance gap between VC-backed companies and the best predicted performers:

$$\underbrace{(y_i|h_i = 1) - (y_i|h_i = 0)}_{\text{treatment effect}} > -\underbrace{((y_i|h_i = 0) - (y_j|h_j = 1))}_{\text{performance gap}(>0)}. \tag{8}$$

Since the performance gap on the right-hand side is positive, the assumption implies that the VC treatment effect must not be too negative. The large and positive VC treatment effect documented

[18] See for example https://www.angellist.com/blog/do-startup-valuations-matter-for-investment-returns and https://www.signatureblock.co/articles/do-valuations-matter.

in the literature (e.g., Puri and Zarutskie, 2012; Chemmanur, Krishnan and Nandy, 2011) suggests that this is a weak assumption.

4.1.3 Market-level assessment and omitted payoff bias

Because we identify the VC-backed status for each company using the SINE survey, we observe the aggregate outcome of VCs' decisions rather than individual VCs' choices, $h_i = \{0, 1\}$. For each new company i, we observe $H_i = max[h_{ij} | j \in J]$, where J is the set of VCs. This market-level assessment of the company's potential allows us to abstract from several complexities inherent to the VC decision-making process. It helps mitigate the omitted payoff bias (Kleinberg et al., 2018), addressing concerns that VCs may have preferences not captured when using operational performance to identify the best companies. By observing a company's VC-backed status in the aggregate, we are not limited to observing whether a firm matched with one particular investor. This approach alleviates concerns related to negotiations, the two-sided matching process subject to negotiations (Cong and Xiao, 2021), VCs' portfolio considerations, and idiosyncratic preferences or constraints driving investment decisions unrelated to a company's potential (e.g., personal relationships, timing constraints, concerns about peer comparisons). Abstracting from individual match-level complexities allows us to focus our analysis on broader patterns of VC decision-making and their alignment with investing in the best companies.

While VCs are ultimately judged on their ability to generate returns, which are closely tied to the performance of their portfolio companies, we acknowledge that the potential for omitted payoff bias remains a concern. We explore this issue in two ways. First, in Section 5.1, we analyze the extent to which our results depend on the specific outcome measure used to train and evaluate the model. Second, in Section 5.2, we embed several constraints and preferences that VCs may have and test how accounting for these affects our results.

4.2 Algorithm Design

To assess whether VCs allocate investments to the companies deemed most promising (for which $M(x_i) > 1 - s$), we design an algorithm that takes characteristics of company i as its input vector x_i to predict company performance y_i. In this section, we describe how we train and evaluate the algorithm. We then show the algorithm's predictive ability across various performance metrics.

Algorithm class and train/test sets. We use Gradient Boosting Trees (*XGBoost*) to generate performance predictions (Chen and Guestrin, 2016). The algorithm is trained on three cohorts of entrepreneurs (1998, 2002, and 2006) representing 69% of our data (84,583 observations) using 10-fold cross-validation. The test set is always left untouched during training. The model's predictions are evaluated out-of-sample on the test set comprised of entrepreneurs in the 2010 cohort (37,353 observations, or 31% of our data).[19] We follow standard practice in the machine learning literature and split our sample into a training and a test sample to prevent the algorithm from appearing to do well because it is being evaluated on data it has already seen. We verify the robustness of the model's predictive ability across several train/test splits and report results in Appendix Table E.8.

Input features. To generate predictions of future venture performance, the algorithm uses a set of 47 covariates (462 covariates once one hot-encoded) that include the entrepreneur's demographics (gender, age, nationality, education), work experience, as well as answers to the administrative survey (e.g., what motivated the founder to start her venture, whether this is the first company she founded, and what her growth expectations are). Examples of firm-level covariates include industry and number of employees. Because our objective is to study VCs' decision making, we avoid look-ahead bias by ensuring that all input features are *ex-ante* covariates and that the information used by the algorithm would easily be accessible to any VC during a first-pass evaluation of the venture.[20] Table 1 reports summary statistics for a subset of input features. We report these statistics separately for the training set and the test set.

[Insert Table 1 here]

Although most input features (i.e., entrepreneur and firm characteristics) are similar across the training and test sets, we observe that average realized performance is slightly higher in the test set compared to the training set, and some founder characteristics, such as the entrepreneur's age and

[19] Our train/test split is based on cohorts rather than a random split for three reasons. First, this approach avoids using outcomes of companies created in the future to make performance predictions. Second, it sets a level playing field for the predictive model against VCs, ensuring that both would only be able to observe past new companies' performance before selecting new ones. Third, it allows us to examine whether the underlying data-generating process that links firm characteristics to firm performance has changed over time, such that different combinations of characteristics might predict success in 2010 and in earlier cohorts.

[20] As a result, we omit several covariates that are available in the survey that we deem not readily available to a potential investor (e.g., bank loans). These omitted variables do not count toward the 47 covariates described earlier.

education are somewhat larger in the test set.[21,22]

4.3 Predictive Accuracy

All companies in test set. We compare our performance predictions $\hat{m}(x_i)$ to the observed realized performance y_i, revenue at age 5 (in log) for the 37,353 observations in our test set. Figure 1 plots a binned scatterplot depicting the relationship between algorithmic predictions and the observed outcome among *all* new companies in our representative sample, that is, both VC-backed companies and non-VC-backed companies. Each point represents the average realized performance for new companies grouped in bins according to their predicted performance. Figure 1 illustrates the algorithm's ability to predict the distribution of new companies' success reliably.[23]

[Insert Figure 1 here]

Most promising companies in the test set. Only about 0.3% of new companies receive VC funding in our test set, so that $s \simeq 0.3\%$ in Equation (3).[24] We focus on the set of ventures in $\mathcal{A}_s$ with various cutoffs for s in Figure 2.

[Insert Figure 2 here]

First, we find that the average performance of companies in $\mathcal{A}_s$ increases as s decreases – i.e., as selectivity increases. Second, despite the large variance in the outcome variable in the data, Figure 2 shows that the algorithm is able to reliably rank new companies, regardless of the number of companies in $\mathcal{A}_s$. Our interpretation of Figure 2 is that even within the subset of most promising companies in the test set, the algorithm is able to produce a useful ex ante ranking of companies. In other words, the algorithm demonstrates predictive ability along the entire distribution, even in the right tail.

[21] Liebersohn and Lyonnet (2024) study the time-series evolution of entrepreneurship quality over time, showing that the quality of entrepreneurs has increased over the years. For the purpose of our analysis, long-term changes in entrepreneur characteristics and in the relationship between entrepreneur characteristics and new company performance make it more difficult for an algorithm trained on the earlier training set to be successful at predicting the performance of new companies in the later test set.

[22] We report input features statistics for the 2014 and 2018 cohorts of entrepreneurs in Appendix Table E.7.

[23] Appendix Figure D.1 displays the top SHAP values for this model.

[24] This fraction is slightly lower than that in the US (Puri and Zarutskie, 2012; Lerner and Nanda, 2020). The effective threshold $1 - s$ depends on VCs' financing and operational constraints. In the U.S., $1 - s \simeq 0.995$, so that VCs invest in about 0.5% of all new companies in a typical year. Our main analysis focuses on the 2010 cohort of French entrepreneurs, for which $s \simeq 0.3\%$ and $N = 120$.

Omitted Payoffs. The results presented so far report the model's predictive accuracy when predicting company revenue, for which we document a strong correspondence with exit valuations in Section 4.1. To mitigate the potential for omitted payoff bias discussed in Section 4.1.3 (Kleinberg et al., 2018), we now explore the robustness of the results by training and evaluating the model on different outcome measures.

Table 2 reports the performance of several predictive models trained on several company outcomes measures, shown in the first column. When the model is trained to predict companies' (log) revenue at age 5 (age 7), for example, the observed average revenue at age 5 (age 7) of companies in $\mathcal{A}_s$ is 6.05 (5.62). For comparison, the average revenue of VC-backed companies at age 5 (age 7) is 2.82 (2.46). We also train the model to predict imputed valuations, revenue growth, as well as two measures of home runs, to account for VCs' focus on the right tail: whether the venture is acquired or goes through an IPO and an indicator variable for those in the top 5% of the revenue distribution. Appendix B describes the data source and construction of these outcome measures. We find that companies selected by a predictive model outperform VC-backed companies, not only in the specific success measure the model was trained on, but also across other performance metrics.[25] The robust predictive performance exhibited by the models across diverse outcome measures provides reassurance that our results do not depend on a specific performance measure.

[Insert Table 2 here]

Robustness. The results in Appendix Table E.8 show that the algorithm's predictive accuracy is robust to dropping one of the training cohorts from the training set (Panel A), and to using a random split across cohorts. In unreported results, we also find that our algorithm is able to predict performance in a "pure hold-out" sample of new companies founded in 2014 and 2018 that were not accessible when we first developed our algorithm.

5 Do VCs Make Mistakes?

5.1 The performance of VC-backed companies vs. best predicted performers

Differences in Operational Performance. We begin by comparing the realized performance of VC-backed companies in $\mathcal{V}_s$ to that of the best predicted performers in $\mathcal{A}_s$, in terms of (log)

[25] The only exception is when evaluating models on exits when they were trained on alternative outcomes. This is not surprising considering that exits are largely endogenous to being VC-backed (see Section 4.1). The model trained on exits does identify the same number of exits as VCs.

revenue at age 5.[26] Figure 3 reports the distribution of realized outcomes for all new companies, for the set of VC-backed companies, $\mathcal{V}_s$, and for the companies in $\mathcal{A}_s$. For comparison purposes, we set $|\mathcal{A}_s| = |\mathcal{V}_s| = 120$.

[Insert Figure 3 here]

Figure 3 illustrates several interesting facts. First, the average VC-backed company performs better than the average company: the average log of revenue is 2.82 for VC-backed companies, whereas it is 2.43 for the entire sample.[27] This operating performance gap confirms VCs' ability to identify, invest in, and help promising new ventures. Second, we confirm the *Babe Ruth Effect* in our data: VCs bet on magnitude over frequency, and outcomes tend to follow a power law distribution (Mallaby, 2022; Strebulaev and Dang, 2024). Third, the realized average performance of the best predicted performers (in red) is greater than that of VC-backed companies. Crucially, this is not just an average effect: The best predicted performers include fewer companies that fail within 5 years and more top performers among surviving companies.

These distributions imply that VCs invest in some companies that perform predictably poorly and pass on others that perform predictably well. These two types of errors in VCs' decisions represent a first indication that the process by which VCs acquire and aggregate signals about a venture's prospects may be inefficient. VCs' policy in Equation (4) may differ from the optimal policy in Equation (3), such that $\Delta(x, z) \neq 0$.

Sensitivity Analysis. We assess outcomes using venture revenue, which we have shown in Section 4.2 to be highly correlated with VCs' returns in Pitchbook, even in the right tail of the distribution. However, the above results come with the caveat that we do not observe the deal size and deal terms, so we cannot directly infer that higher revenue would translate in higher investment returns for the best predicted performers. To assess the sensitivity of our results to these pricing considerations, we relax our implicit assumption of constant deal terms and analyze how expensive the deal terms for companies in $\mathcal{A}_s$ would need to be to invalidate our results.

We leverage Pitchbook data to estimate imputed multiples on invested capital (MOIC) across the empirical distribution of French deal terms. We first estimate imputed post valuations by multiplying observed y_i (revenue at age 5) by M_s, the industry-specific median revenue multiples at

[26] Recall that we drop companies that operate in industries that never receive VC funding during our sample period to focus on companies that are more suited to receive VC funding. Our results remain qualitatively similar without this filter.

[27] Companies that fail by age 5 are included and assigned zero revenue.

exit (post valuation/exit revenue).[28] For each deal, we impute the MOIC given by Equation (7).[29] We empirically estimate the distribution of deal terms $\frac{\delta}{k}$ using French early VC deals in Pitchbook during 2009-2011 and assume dilution to be 75% (as in Davenport, 2022). For each deal, we thereby obtain imputed MOICs that span the empirical distribution of deal terms. We then compute the estimated portfolio MOIC as the average of estimated company-level MOIC in $\mathcal{A}_s$, ($MOIC_\alpha$), and for VC-backed companies in the 2010 cohort ($MOIC_h$).

[Insert Figure 4 here]

Figure 4 shows the differential MOIC returns for VCs vs. best predicted performers for different assumptions on $\frac{\delta}{k}$ in the two portfolios. Common deal terms secured by VCs on their actual investments are represented by the median of the observed deal terms in the data (x-axis). With such terms, Figure 4 shows that VCs would have had to get, on average, very unfavorable terms – below the 8th percentile of the distribution for companies in $\mathcal{A}_s$ – in order for the average difference in MOIC to be negative. Alternatively, assuming VCs could secure median deal terms for companies in $\mathcal{A}_s$ (y-axis), they would have needed to negotiate exceptionally favorable deal terms for the VC-backed companies – above the 93rd percentile of the empirical distribution – for the difference in MOIC to be negative. Appendix Table E.2 reports this percentile pair (8th and 93rd) under various revenue multiple assumptions for companies in $\mathcal{A}_s$.

The key takeaway from this sensitivity analysis is that the companies in $\mathcal{A}_s$ do not, of course, surpass the selections made by VCs under any and all circumstances. If VCs had to pay extremely high prices for these companies or received extremely low revenue multiples for these deals, it could rationalize why they decided to pass.[30] The results in the next sections, however, cast further doubt on the possibility that deal terms fully account for our main findings.

Quantifying VCs' Errors. To assess the potential cost associated with the two types of errors identified in Figure 3, Table 3 Panel A reports the imputed portfolio MOIC resulting from dropping a subset of VC-backed companies with low predicted performance. In Panel B, we report the average imputed MOIC from investing in the best predicted performers. While this analysis relies

[28] We compute industry-level revenue multiples using US deals to circumvent data limitations on French deals.

[29] Although the MOIC metric does not account for VCs' investment holding periods, our approach offers an implicit control for investment duration. By calculating the difference in imputed portfolio MOIC using venture revenue at the 5-year mark, we effectively standardize the investment horizon across investments in the two compared portfolios. Moreover, the best predicted performers outperform the VC-backed portfolio in terms of revenue throughout the entire horizon (up to seven years out), not only at the five-year mark.

[30] This seems unlikely, as practitioners often report that pricing considerations in early-stage investing are not key drivers of investment returns: What matters is picking the right companies (e.g., see AngelList and SignatureBlock).

on assumptions for imputing MOICs, we find that the cost of picking predictably bad performers is dwarfed by the cost of missing out on predictably best performers. While dropping the bottom 10% of VC-backed companies with the lowest $\hat{m}(x_i)$ increases the imputed portfolio MOIC by 9%, investing exclusively in companies in the top 1% of predicted performance $\hat{m}(x_i)$ increases imputed portfolio MOIC by about 200%. This is consistent with reports that VCs care much less about investing in bad companies than missing out on winners. As Strebulaev and Dang (2024) note: "home runs matter; strikeouts don't."

5.2 Counterfactual models of VC allocation.

Our results so far show the performance of the best predicted performers in $\mathcal{A}_s$ when the investable pool $\mathcal{D}$ includes all new companies, as long as they are in an industry that has received VC in our sample period. This approach implicitly assumes that VCs could have selected any company in $\mathcal{A}_s$. However, supply and demand factors can explain why VCs may not invest in the best predicted performers. On the supply side for example, VCs often specialize in specific industries where they possess expertise. On the demand side, not all new companies seek VC funding. This section analyzes the performance of the best predicted performers, considering various restrictions to the investable pool that stem from these supply and demand factors.

To quantify the importance of supply and demand factors, we create counterfactual models that sequentially drop VC-backed companies with the lowest $\hat{m}(x_i)$ and replace them with the best predicted performers (i.e., companies in $\mathcal{A}_s$ with the highest $\hat{m}(x_i)$) selected from various investable pools $\mathcal{D}$, ensuring the total number of portfolio companies stays constant at $|\mathcal{V}_s| = 120$. We first run the counterfactual model without any constraints on $\mathcal{A}_s$. The top (red) line in Figure 5 shows the performance of the counterfactual model as the number of VC-backed companies replaced with high predicted performers increases (along the x-axis). The leftmost point shows the performance of VC-backed companies in $\mathcal{V}_s$ and the rightmost point shows the performance of companies in $\mathcal{A}_s$.

[Insert Figure 5 here]

Next, we introduce several sets of constraints or preferences that simulate those of VCs. As before, our algorithm ranks companies in the test set by predicted performance using the function $M(x_i)$. However, unlike the previous approach where companies in $\mathcal{A}_s$ were selected based on $M(x_i) > 1 - s$, we now also require that these companies match VC-backed companies on one or more criteria. We start in Figure 5 with a first set of constraints that pertains to the companies' industry and location. If a VC-backed company is excluded, the counterfactual model identifies the best predicted performer (with the highest $\hat{m}(x_i)$) within the same industry and/or geographical location to replace it.

This analysis yields several interesting results. First, all counterfactual models outperform VCs' selections, indicating that industry and location constraints do not fully account for the performance gap. Second, we can quantify the shadow cost of a given constraint as the difference in average portfolio company performance between the unconstrained counterfactual model and the constrained counterfactual model subject to that constraint. As expected, the more restrictive the set of constraints, the lower the portfolio performance of the counterfactual model. Figure E.2 in the Appendix shows similar results when the companies' performance is evaluated using imputed MOICs instead of revenue. We find that investing in the best predicted performers that operate in the same industries and locations as VC-backed companies would increase VCs' imputed MOIC by 50%. Overall, our results are not driven by the algorithm selecting companies in industries or locations where VCs typically do not invest, or in sectors characterized by low revenue multiples.

In Figure 6, we introduce additional constraints to focus on companies that closely resemble those backed by VCs, ensuring that the best predicted performers identified by the model are realistic candidates both in terms of desirability and attractiveness to investors. The grey line in Figure 6 replaces each VC-backed company with a company in $\mathcal{A}_s$, reflecting constraints that limit the pool $\mathcal{D}$ of investable companies to those whose founders' responses to financial difficulty questions in surveys match those of VC-backed founders, resulting in a significant narrowing of the performance gap. The yellow line also constrains the company in $\mathcal{A}_s$ to match the industry and growth prospects of the replaced VC-backed company. Finally, the orange line constrains companies in $\mathcal{A}_s$ to match VC-backed companies on sets of criteria related to entrepreneur characteristics, including growth aspirations, innovation, idea, and hiring prospects.[31]

[Insert Figure 6 here]

Even for very restricted investable pools, companies in $\mathcal{A}_s$ outperform VC-backed companies, indicating that neither VC constraints nor entrepreneurial and company characteristics fully account for the performance gap. In Table E.3, we report the average portfolio company performance using a menu of constrained counterfactual models that restrict the investable pool $\mathcal{D}$. These counterfactuals account for a range of VC constraints and preferences. This analysis suggests that our findings

[31] The growth related questions in the entrepreneur survey are: "do you expect to grow?","do you expect to hire?","is a new idea the key motivation for starting your business?", and "do you consider your business to bring an innovation?"

are unlikely to result from the model identifying promising companies that VCs would inherently overlook due to their investment criteria, thereby alleviating concerns about omitted payoffs.

The objective of restricting the investable pool is to account for observed founders' and VCs' constraints and preferences. Taking this revealed preference argument to the extreme, we examine the algorithm's predictive ability within the set of VC-backed companies only, for which by revealed preference, supply and demand concerns are absent. Figure 7 shows that the model still generates a useful ex-ante ranking $M(x_i)$ among the set of companies that have received VC-backing, identifying companies with low $\hat{m}(x_i)$ that end up performing badly as well as companies with high $\hat{m}(x_i)$ that perform very well.

[Insert Figure 7 here]

5.3 How Do the Best Predicted Performers Grow?

Our findings beg the question of how the best predicted performers that VCs pass on end up doing better. According to Catalini, Guzman and Stern (2019), companies with growth potential are similar to each other, irrespective of whether they are VC-backed. This suggests that the best predicted performers may not have a fundamentally different growth path from the VC-backed ones. To provide a better understanding of the growth opportunities of highly promising non-VC-backed firms, we analyze founders' responses to two specific survey questions. To ensure that we analyze companies similar to VC-backed firms, we focus on the best predicted performers in $\mathcal{A}_s$ who match VC-backed firms in terms of financial constraints, industry, and growth prospects (yellow line in Figure 6).

The first question inquires about external sources of financing. We report the founders' answers in Figure E.3. The main source of outside funding upon firm creation for these companies is through bank loans in the company's name (more than half of companies). Personal bank loans represent another significant source of outside financing, as are external grants. Approximately 10% of founders are self-funded. The second survey question provides valuable insight into the challenges encountered by the best predicted performers. It asks founders to identify the main hurdle they faced when starting their business. We report the founders' answers in Figure E.4. Obtaining outside funding is the primary obstacle reported by these founders, followed by administrative hurdles and difficulties in hiring. This is consistent with the best-predicted performers, like their VC-backed counterparts, pursuing strategies requiring substantial outside funding. The challenge of hiring is a common pain point shared with VC-backed companies having high growth aspirations.

We view these results as consistent with Catalini, Guzman and Stern (2019), suggesting that the set of best predicted performers we identify face similar constraints as VC-backed firms. Of course, many firms with high growth aspirations likely never receive outside funding and therefore do not succeed.[32] Not every firm in the set of best predicted performers becomes successful, perhaps due to lack of VC-backing, but many do very well and outperform VC-backed firms. The question, then, is why don't VCs invest in such firms early on. This is the topic of the next section.[33]

6 Analyzing VCs' choices

6.1 The characteristics of VC-backed companies vs. best predicted performers

This section examines how various demographic measures of entrepreneurs differ between VC-backed entrepreneurs ($\mathcal{V}_s$) and the best predicted performers ($\mathcal{A}_s$). It is important to note that while informative, these statistics do not indicate whether certain characteristics disproportionately influence VCs' decisions. Such an analysis is the topic of Section 7, which leverages ML prediction methods. Instead, this section is a preliminary exploration of the profiles of entrepreneurs and whether VC-backed entrepreneurs and the best-performing ones systematically differ from the population and each other. We start with Figure 8, which reports the probability densities of founders' age, gender, education level, and geographic location for VC-backed and the best-predicted performers in the test set.

[Insert Figure 8 here]

Age. Although the average founder age of VC-backed and the best predicted performers is approximately the same, VCs select a larger fraction of young entrepreneurs than there are in $\mathcal{A}_s$. This result is in line with findings in Azoulay et al. (2020) that investors overemphasize youth as a key trait of successful entrepreneurs.

[32] Although some firms that never received funding might not have been included in the data if they never got started, all firms that registered are in the data. Registration is a typical first step before seeking outside funding.

[33] We do not rule out the possibility that these new companies may receive VC funding later, once VCs observe additional tangible signals. Recall that our focus is on early-stage financing decisions, when hard information is scarce or non-existent, and uncertainty about which startups have the best chance of success is highest. These early-stage investments, often based on "gut investment decisions" (Gompers et al., 2020; Hu and Ma, 2021) are crucial for the survival of startups (Kerr, Nanda and Rhodes-Kropf, 2014), impacting the entrepreneurial ecosystem and innovation. Using Pitchbook, we find that 10% of the best predicted performers eventually receive VC, confirming the suitability of the identified best predicted performers for VC investment.

Gender. Panel B examines differences in founders' gender. While female entrepreneurs represent 28% of entrepreneurs in our test set, only 9% of VC-backed companies are female-led. We find slightly more female entrepreneurs among the best predicted performers, at 14%.

Elite School Education. While only 6% of entrepreneurs are elite school (Grande Ecole) graduates, panel C shows that both VC-backed and the best predicted performers in $\mathcal{A}_s$ have more founders who graduated from elite schools. However, VCs select more than twice as many elite school entrepreneurs (27%) relative to the companies in $\mathcal{A}_s$ (11%).

Geography. Finally, Panel D shows the fraction of Paris-based entrepreneurs among VC-backed entrepreneurs and the best predicted performers. In our test set, only 8% of new companies are located in the Paris region. Interestingly, 21% of VC-backed companies are in Paris, while only 7% of the best-predicted performers in $\mathcal{A}_s$ are located there.

To provide a more complete picture of the VC-backed and best predicted performer entrepreneurs and their differences, we report in Table 4 the summary statistics for a larger set of features as well as t-tests for the difference between these groups. In Table 4, the number of companies in $\mathcal{A}_s$ matches the number of VC-backed companies in the test set (120 companies), but the basic patterns of discrepancy between VC-backed companies and the best predicted performers are consistent across various top levels of selectivity. We report these statistics for two cutoffs, $s = 0.5\%$ and $s = 1\%$, in Appendix Table E.4.

[Insert Table 4 here]

6.2 Predicting VCs' Decisions

To better understand why VCs make investment mistakes, we develop a separate estimator, denoted $\widehat{h}(\cdot)$, that predicts for each company whether it is VC-backed. We train this classification algorithm on a random split of 70% of the observations in the 1998, 2002, and 2010 cohorts, and test it out-of-sample on the remaining 30% of observations.[34]

[34] We exclude the 2006 cohort in this test because our prediction exercise is to predict which companies are VC-backed, but the 2006 entrepreneur survey does not allow for the determination of VC-backed status. We use a random split for this exercise for two reasons. The first is technical due to the limited number of VC-backed companies in these three cohorts. The second reason is that we are not comparing the performance of VC-backed companies to the best predicted performers in this exercise. Thus, we do not need to ensure a level playing field for the algorithm against VCs, where both would observe the performance of past new companies. We verify that our results are unchanged when we use a random split on the 2010, 2014 and 2018 cohorts of entrepreneurs.

Model performance. Our predictive model predicts VCs' investment decisions well. Figure 9 shows that our model has an area under the curve (AUC) of .77. This implies that if we randomly select a VC-backed company and a non-VC-backed company, our model will assign a higher probability of being VC-backed to the company that is truly VC-backed with an 77% chance.[35]

[Insert Figure 9 here]

One striking result is that if restricted to three founder demographic features, our predictive model of VCs' decisions produces an AUC of .60. Therefore, much of the signal to predict VCs' decisions is captured by these three demographic features. In contrast, when the predictor of venture performance $\widehat{m}(\cdot)$ takes only these three features as its input, the performance of companies in $\mathcal{A}_s$ decreases dramatically. The model's much lower predictive performance when restricted to these three input features implies that the signal to predict venture performance lies elsewhere, and VCs appear to put disproportionate weight on these three demographic features when making investment decisions.

Signal beyond venture performance. We follow the approach in Ludwig and Mullainathan (2024) and test in a regression framework whether there exist factors beyond predicted performance that can predict VCs' decisions. We first regress VCs' actual decisions, $\textit{VC-backed}_i$, on our algorithmic predictions of VCs' decisions:

$$\textit{VC-backed}_i = \beta_0 + \widehat{h}(X_i)\beta_1 + \epsilon_i \tag{9}$$

Table 5 confirms that our model of VCs' decisions indeed performs well. The model's estimates are correlated with VCs' actual decisions (column 1) and imply that a company in the third quartile of our VC-backed predictions is 1.02 p.p. more likely to be VC-backed compared to a firm in the first quartile, a 123% increase relative to the mean.[36] We then regress VCs' actual decisions on our performance predictions $\widehat{m}(X_i)$ using our two home run measures ($top5rev_5$ and successful exits) as well as revenue at age 5 (log):

$$\textit{VC-backed}_i = \beta_0 + \widehat{m}(X_i)\beta_1 + \epsilon_i \tag{10}$$

[35] Appendix Figure D.2 displays the top SHAP values for this model.

[36] There are 26,440 observations in this regression, which is the number of observations in our test set when the algorithm is trained using a random split using the 1998, 2002, and 2010 cohorts (this is due to VC-backed status not being available for the 2006 cohort).

If VCs were not concerned about portfolio companies' revenue, we would expect revenue predictions not to load significantly ($\beta_1 = 0$ in Equation (10)). This is not the case. In Column 2, we observe a strong correlation between VC-backed status and predictions for being in the top 5% of revenue performers within a cohort. A one standard deviation increase in our performance predictor $\widehat{m}(X)_{top5rev_5}$ is associated with a 59% increase in the probability of being VC-backed, relative to the mean. Column 3 (Column 4) shows that exit predictions (revenue predictions) also correlate strongly with VCs' decisions. A one standard deviation increase in $\widehat{m}(X)_{exit}$ ($\widehat{m}(X)_{LogRevenue_5}$) is associated with a 59% (47%) increase in the probability of receiving VC relative to the mean. These estimates show that VC decision-making closely aligns with algorithmic home run predictions, consistent with expert knowledge in the field.[37]

We test whether our predictions of VCs' decisions $\widehat{h}(X_i)$ remain significant once we control for performance predictions by estimating:

$$VC\text{-}backed_i = \beta_0 + \widehat{h}(X_i)\beta_1 + \widehat{m}(X_i)\beta_2 + \epsilon_i \tag{11}$$

Columns 5 through 8 of Table 5 show that there remains significant predictability in VCs' behavior even when controlling for algorithmic predictions of venture performance. The coefficient on our predictions of VCs' decisions, β_1, remains virtually unchanged from column 1 to column 8 where we add venture performance predictions. This result implies that our VC-backing predictor picks up signals above and beyond venture performance and suggests that there remains strong predictability in VCs' behavior beyond what we would expect if VCs were only interested in future venture performance and could predict this performance accurately.

[Insert Table 5 here]

7 Do VCs Overweight some Characteristics?

Lerner and Nanda (2020) write that while we know that VCs "rely heavily on signals of entrepreneur quality, we know very little about whether the emphasis on these signals is efficient". A key novelty of our analysis is demonstrating that VCs do not correctly weigh all entrepreneur characteristics in their decisions. Following the methodology developed in Mullainathan and Obermeyer (2022),

[37] For example, Andreessen Horowitz (a16z), the largest U.S. VC firm as of May 2024, articulates its investment strategy as "far more about how big the outcome will be if a deal succeeds than all the ways that it can fail" (link: a16z). The famous VC Peter Thiel wrote, "The biggest secret in venture capital is that the best investment in a successful fund equals or outperforms the entire rest of the fund" (Masters and Thiel, 2014).

we regress VCs' decisions on exit predictions $\widehat{m}(X)_{exit}$ as in Table 5, as well as on exit predictions from simple models, $\widehat{m}_{simple}(\cdot)$ that depart from the estimator $\widehat{m}(\cdot)$ by restricting the set of input features:[38] We regress VCs' decisions on our full model predicting which companies are most likely to exit, as well as our simple models:

$$\textit{VC-backed}_i = \beta_0 + \widehat{m}(X_i)\beta_1 + \widehat{m}_{simple}(X_i)\beta_2 + \epsilon_i \tag{12}$$

We interpret the sign of the coefficient β_2 as in Mullainathan and Obermeyer (2022). Intuitively, if $\beta_2 = 0$, the variables used in $\widehat{m}_{simple}(\cdot)$ do not matter for VCs' decisions over and above their effect on companies' performance. Alternatively, $\beta_2 \neq 0$ implies that the variables used in $\widehat{m}_{simple}(x_i)$ contain signal to predict VCs' decisions beyond their effect on predicted performance $\widehat{m}(x_i)$.[39] If $\beta_2 > 0$, VCs *overweight* the feature used in the simple model, that is, this feature disproportionately influences VCs' decisions relative to its actual impact on predicted venture performance. Instead, if $\beta_2 < 0$, VCs *underweight* the feature used in the simple model. Table 6 contains the results of Equation (12) for several simple models. We start with Panel A, in which entrepreneur features are used as inputs.

This approach leverages ML predictions to account for the predictive power of each characteristic on future performance – capturing potentially complex, interacted, and non-linear relationships. Controlling for predicted performance, the coefficient β_2 isolates the role of each characteristic on VCs' decisions over and above their effect on predicted performance. Crucially, this methodology enables us to assess whether certain characteristics disproportionately influence VCs' choices relative to their impact on predicted performance. Our analysis reveals this is the case.

Personal Characteristics. Since potential investors are highly responsive to information about the founding team (Bernstein, Korteweg and Laws, 2017; Gompers et al., 2020), our first sim-

[38] For simplicity of notation, we refer to the predictions of the simple models as $\widehat{m}_{simple}(X_i)$ for each entrepreneur i even though these models are restricted to a limited set of features in X_i. In the first part of the paper, we investigate whether VCs invest in the best predicted performers using venture revenue as our success measure. This metric, available for both VC-backed and non-VC-backed companies, allows us to circumvent the selective labels problem, under the assumptions described in Section 4.1. We now focus on understanding VC decision-making behavior and therefore use the performance predictor that best explains VCs' decisions. Consistent with the evidence in Gompers et al. (2020) that anticipated exits are the most important factor in VCs' evaluations, as well as the evidence in Table 5 and expert knowledge, our estimators $\widehat{m}(\cdot)$ and $\widehat{m}_{simple}(\cdot)$ effectively predict exits.

[39] $\beta_2 \neq 0$ would imply

$$\frac{\mathrm{Cov}(M_{\widehat{m}}\,\textit{VC-backed}, M_{\widehat{m}}\widehat{m}_{simple})}{\mathrm{Var}(M_{\widehat{m}}\,\textit{VC-backed})} \neq 0, \tag{13}$$

where $M_{\widehat{m}}\,\textit{VC-backed}$ and $M_{\widehat{m}}\widehat{m}_{simple}$ are the vectors of residuals from the regression of *VC-backed* and $\widehat{m}_{simple}(\cdot)$ on the columns of $\widehat{m}(\cdot)$, respectively (Frisch and Waugh, 1933).

ple model uses the personal characteristics of the entrepreneur as input features: age, gender, education, nationality, and whether the entrepreneur has relatives who are entrepreneurs. In Column 1, we regress VCs' decisions on $\widehat{m}(x_i)$, our full estimator that predicts exits. Column 2 adds our first simple model based on personal characteristics. $\widehat{\beta}_2$ is significant, which means that $\widehat{m}_{simple}(\text{personal features}_i)$ is *additionally* predictive of VCs' decisions, and it is positive, so that VCs *overweight* personal characteristics of entrepreneurs in their investment decisions. The interquartile range of $\widehat{m}_{simple}(\text{personal features}_i)$ is 0.0032, translating to a shift of 0.081 p.p. in the probability of being VC-backed, a 25% increase relative to the baseline VC-backing rate.[40] In Columns 3 to 8, we test other simple models focusing on one personal characteristic in isolation.

[Insert Table 6 here]

Gender. We find that VCs exaggerate several entrepreneur features. Column 4 shows that VCs put too much weight on the entrepreneur's gender in their decision to back a firm. Although female entrepreneurs do have fewer exits than male entrepreneurs, we find that controlling for exit performance predictions, male entrepreneurs are 0.14 p.p. more likely to be VC-backed than they would if VCs' decisions were solely based on the effect of gender on exit performance, a 43% increase relative to the mean. This finding is consistent with existing evidence that VCs pass up promising female-founded new ventures (e.g., Kanze et al., 2018; Howell and Nanda, 2019; Hebert, 2023; Calder-Wang and Gompers, 2021).

Education. VCs exaggerate the entrepreneur's education in their decisions (see, e.g., Queiró, 2021, on the importance of education in new companies' performance). In Column 5, we find that VCs overweight the fact that an entrepreneur has a graduate degree. Therefore, having a graduate degree increases an entrepreneur's likelihood of receiving VC-backing to a greater extent than justified by its effect on predicted performance. Column 6 shows that VCs overweight the "Elite School" feature, a dummy equal to one if the entrepreneur graduated from an elite French school.[41] Being an elite school graduate shifts the likelihood of being VC-backed by 0.68 p.p., or a 212% increase over the baseline. VCs are therefore three times more likely to back an entrepreneur who graduated from an elite school, even when controlling for exit performance predictions.

[40] Approximately 0.32% of companies are VC-backed in our test set.

[41] The Elite School variable is only available in the data starting in 2006, which prevents us from using it in our main analysis. It is equal to one if the entrepreneur graduated from a *Grande École* or an engineering school.

Other Entrepreneur Characteristics. We do not find evidence that VCs exaggerate the entrepreneur's age in their decisions (column 3), that they make systematic prediction mistakes when assessing the entrepreneur's nationality (column 7), or entrepreneurs with entrepreneurial families (column 8). Similarly, we do not find that VCs overweight optimism (column 9). They do, however, put too much weight on whether the founder is a serial entrepreneur (column 10). Our estimates imply that VCs back serial entrepreneurs 1.5 times more than if they were basing their decisions solely on exit predictions.

Venture Characteristics. We now examine simple models that focus on venture characteristics. As shown in columns 2 and 3 of Panel B in Table 6, VCs place significant emphasis on innovative ventures based on novel ideas. Column 4 explores proxies for a venture's traction, including the total number of workers, number of clients, and client locations. The interquartile range of $\widehat{m}_{simple}$(traction) is 0.003, which translates to approximately a 0.07 p.p. increase in the probability of receiving VC backing. This represents a 22% increase relative to the baseline average, suggesting that investors overweight venture traction in their decision-making process. Columns 5 and 6 focus on industries that are most frequently VC-backed in our data. Controlling for exit predictions, we do not find that VCs put disproportionate weight on the high-tech industry, yet, they do for the scientific R&D sector. Finally, our analysis also reveals that Paris-based companies are 0.31 p.p. more likely to receive VC backing compared to what exit predictions would suggest. This finding implies that controlling for exit predictions, ventures in Paris are almost twice as likely to be VC-backed compared to those elsewhere. This observation aligns with the concern raised by Lerner and Nanda (2020) regarding the high concentration of the VC industry.

8 What is Driving VCs' Bias?

8.1 Memory and Representativeness of Success

The disparities between VC-backed and best predicted performers documented in the previous section raise the question of the origins of such discrepancies. To understand these patterns, we explore how early-stage VCs, who have access to little to no "hard" information (Mullainathan, 2002), and who declare often making "gut investment decisions" (Gompers et al., 2020; Hu and Ma, 2021), form expectations of entrepreneurial success.

The growing belief formation literature shows that beliefs and decision-making are intimately bound up with memory and selective recall (e.g., Mullainathan, 2002; Wachter and Kahana, 2023; Bordalo, Gennaioli and Shleifer, 2020; Bordalo et al., 2023; Conlon and Patel, 2022): What comes to mind drives beliefs and probability estimates of competing hypotheses or scenarios. We conjecture that when a VC considers a raising entrepreneur, she scrolls through her mental database of past entrepreneurs to determine whether the raising entrepreneur resembles past successful entrepreneurs. This "pattern-matching" exercise fits the investment process often described by VCs.[42]

Bordalo et al. (2023) show that the way memory affects the formation of beliefs provides a microfoundation for Tversky and Kahneman (1974)'s representativeness heuristics and the emergence of stereotypes (Bordalo et al., 2016). The assessed probability of success for a type of entrepreneur increases when many examples of successful entrepreneurs of that type come to mind and when it is easier to recall instances of failure among entrepreneurs who do not share the entrepreneur's type. As a result, this assessed probability increases when the entrepreneur type is very diagnostic – or "representative" – of success, meaning the type is more prevalent among successful entrepreneurs than among the rest. For features along which VCs and best predicted performers differ, we calculate their *representativeness* of success for a percentile P of the performance distribution relative to the rest of the distribution $-P$ as the ratio:

$$\frac{Pr(f_i \mid P)}{Pr(f_i \mid -P)} \tag{14}$$

Table 7 contains the results. Column 1 shows the fraction of entrepreneurs with specific characteristics among the top-performing companies, defined as those in the top 1% of the revenue distribution at age 5. Column 2 provides the same information for companies in the bottom 99% of the distribution. Column 3 reports the ratio of these fractions, with values greater than one indicating that the characteristic is more prevalent among top-performing companies. These results confirm that certain entrepreneur characteristics are indeed representative of the best-performing companies, and in particular, the ones that are overweight in VCs' decisions. Because VCs tend to select companies that are representative of the most successful companies, their decisions are based on "kernel of truth" stereotypes (Bordalo et al., 2016).

[42] For example, Paul Graham, founder of the Y Combinator remarked: "I can be tricked by anyone who looks like Mark Zuckerberg. There was a guy once who we funded who was terrible. I said: How could he be bad? He looks like Zuckerberg!" in the New York Times; and Bruce Dunlevie, General Partner at Benchmark Capital: "Pattern recognition is an essential skill in venture capital... while the elements of success in the venture business do not repeat themselves precisely, they often rhyme. In evaluating companies, the successful VC will often see something that reminds them of patterns they have seen before." in AV VC Blog.

[Insert Table 7 here]

In Appendix Table E.5, we complement our analysis of French VC deals with U.S. VC deal-level returns from MSCI-Burgiss to (1) verify if identified stereotypes persist in the U.S. context and (2) confirm the existence of stereotypical thinking in VC using deal-level VC returns rather than company operating performance. We calculate representativeness ratios for company location and industry, focusing on the four largest U.S. states and industries by deal number. We define success as ventures in the top 5% or 1% of deal-level TVPI distribution. Our results indicate that companies in locations and industries that receive the most VC-backing – e.g., California and the I.T. sector – consistently have representativeness ratios above one. Appendix Table E.6 confirms these findings using IRR as a performance measure, suggesting that stereotypes of successful ventures are robust across performance measures and not specific to the French context.

8.2 Connecting Feature Exaggeration and Overestimation of Success Forecast in VCs' Decisions

Consistent with the belief formation literature (e.g. Bordalo et al., 2023), Section 7 shows that certain characteristics disproportionately influence VCs' decisions relative to their impact on predicted performance and Section 8.1 shows that VCs tend to overweight characteristics representative of the best-performing entrepreneurs. In this section, we aim to explain why certain characteristics appear more or less overweighted in VCs' decisions. To achieve this, we explore the relationship between feature exaggeration and distortion in the estimated odds of success (Rambachan, 2024).

Using the 2010 cohort of entrepreneurs and exits as the measure of entrepreneur success, we calculate the true odds of success for entrepreneurs with feature f as the probability of success conditional on f over the probability of failure conditional of f. Assuming that VCs' beliefs about success probabilities for different types of entrepreneurs are reflected in their backing rates,[43] we proxy VCs' estimated odds of success for entrepreneurs with feature f by the ratio of the number of VC-backed entrepreneurs with f to the number of non-VC-backed entrepreneurs with f. We denote Ψ as the distortion in estimated odds of success, calculated as the ratio of estimated to true odds.

Figure 10 plots the coefficients $\widehat{\beta}_2$ from Equation (12), which represent ML-based feature exaggeration, and Ψ, which captures the distortion in estimated odds of success. Focusing on features that VCs overweight, Figure 10 illustrates that the degree to which VCs exaggerate certain en-

[43] Using a survey administered to a large panel of wealthy retail investors, Giglio et al. (2021) show that beliefs are reflected in portfolio allocations.

trepreneur features in their choices increases with their overestimation of the odds of success for entrepreneurs exhibiting these features. Although suggestive, the evidence in Figure 10 nicely ties our ML-derived exaggeration measures with the belief formation literature's insights on stereotype formation through memory processes. Overall, our findings are consistent with the interpretation that overestimation of success probabilities, stemming from memory processes and the representativeness heuristic, is associated with inefficient decision-making by VCs. The role of cognitive biases in shaping investment choices provides a novel and internally consistent narrative that resonates with the empirical VC literature and the literature on human behavioral biases across various domains.

9 External Validity

Our analysis uses data from France; hence, our results raise questions about external validity. The French VC industry might operate differently from those in other countries, and certain specificities of the French VC industry and/or our sample period might influence some of our findings. In this section, we outline specific concerns that may challenge external validity and summarize our analysis of these issues. Appendix C contains our complete analysis, including a detailed description of the French VC industry along with a comparison between the French and U.S. VC industries.

French VC Market Analysis. The VC industry in France is less mature than in the US and operates on a much smaller scale. This raises the concern that French investors may lack experience and make mistakes that more sophisticated VCs would avoid. However, during our sample period, France was the second-largest VC market in Europe, behind the UK and followed by Germany.[44] Moreover, the data suggests a relatively strong international investor presence in the French VC ecosystem, particularly from US investors. Figure C.3 shows that 52% of funds (by count) making early-stage investments in French companies during our sample period are French, and nearly a quarter are US-based. Interestingly, when considering the total amounts raised by French companies during this period, US VC funds have contributed slightly more than French funds, primarily due to the larger size of US funds. That a significant fraction of investments in French companies originate from US investors partly alleviates the concern that our results would arise from French VCs' inexperience.

[44] In 2010 (our test set cohort year), Pitchbook reports a total of 6,464 VC deals in the US with a median invested capital of USD 1.5 million. In comparison, there were 595 (868) VC deals in France (the UK), with a median capital invested of USD 0.80 (0.82) million.

If French VCs make mistakes that more experienced VCs would avoid, we would expect them to generate worse investment returns. Although we caveat that Pitchbook data are subject to selection and reporting biases (which is especially the case for France during our sample period), they are nonetheless useful to examine whether French VCs generate worse returns. First, we measure ”success rates” by the proportion of companies with seed funding that either received subsequent VC funding, were acquired, or underwent an IPO. In Table C.4, we do not find significant differences in success rates between the two countries (32% in the US vs. 31% in France). Second, in Table C.5, we do not find evidence that French investors perform worse than US investors when investing in French companies. Pitchbook reports that the mean (median) fund TVPI is 1.7 (1.17) for funds located in the US, while it is 1.6 (1.46) for funds in France. The difference is not statistically significant. Figure C.5 using the median TVPI for funds in the two countries confirms these results.

Government Involvement. We investigate whether weak financial incentives due to government involvement could explain French VCs' suboptimal decisions. An important difference between the French and US VC ecosystems is the absence of long-term private investors in France, such as pension funds and university endowments, and a weaker network of angel investors (Ekeland, Landier and Tirole, 2016). As a result, public sources constitute a large share of funds raised by French VCs. In particular, Bpifrance, the French public investment bank, has played a significant role over the past two decades by making direct investments in startups and through funds of funds. Bpifrance's government ties raise concerns that its partners may lack incentives to select the best companies.

Several pieces of evidence suggest that weak financial incentives are unlikely to cause French VCs to behave differently from others. First, the above evidence does not show that French funds underperform US funds, mitigating this concern. Second, we use external data to investigate whether government-sponsored funds make worse investments than private investors, which would indicate misaligned incentives (see Section 3). Using independent data from Pitchbook and Bpifrance, we do not find evidence that government-sponsored funds make worse investment decisions than the most active private French VCs (see Table B.2 and Appendix Tables C.3, C.4, and C.5). For instance, Table B.2 uses data on 357 deals from Bpifrance. The top row shows that the average deal-level MOIC is 1.28, but returns are heavily skewed, with the median deal returning less than the invested amount and the top 1% returning ten times the invested amount. While these deals are not limited to early-stage investments, we find this return distribution consistent with Bpifrance partners investing according to an objective function similar to their US counterparts.[45]

French-Specific Characteristics. We investigate whether some of our findings on characteristics that disproportionately influence VCs' decisions could be driven by the French context. For instance, France is often regarded as an elitist country (Lamont, 2002), which may explain why VCs overweight elite school graduates in their choices. While the French context surely plays some role, with around a quarter of French VC-backed founders having graduated from one of the top three French elite schools, we find this fraction comparable to the roughly 25% of VC-backed founders in the US who graduated from an Ivy League school. Additionally, the proportion of female VC-backed founders during our sample period is similar in both countries, 9% in France and 11.5% in the US.

External U.S. Evidence. Finally, several of our results are consistent with external U.S. evidence by Davenport (2022), who also finds that some VC-backed companies have predictably bad performance using ML predictions, and Jang and Kaplan (2023), who find that VC choices overweight personal characteristics of companies' founding teams. Using U.S. data from MSCI-Burgiss, we also find evidence in Appendix Tables E.5 and E.6 that companies in locations and industries that receive the most VC-backing – e.g., California and the I.T. sector – consistently have representativeness ratios above one.

In summary, while we acknowledge that the French VC market differs from the US market in many ways and that not all our results may be fully generalizable, we have not found any systematic differences between early-stage investors in France and those in more mature markets like the US that would imply different decision-making by VCs. Our analysis indicates that the fundamental factors influencing VC decisions may be comparable across markets and does not suggest that our results are unique to the French context. Other studies further reinforce the view that stereotypical VC allocation is not specific to the French context.

10 Conclusion

We use machine learning techniques to study how venture capitalists (VCs) make early-stage investment decisions. By leveraging these techniques and representative administrative data, where we

[45] Several studies of Bpifrance's investment returns confirm our findings that these returns are above average, further dampening concerns about their VC partners' incentives (e.g., Bpifrance, 2021; Gilles, L'Horty and Mihoubi, 2023; Cour des comptes, 2023).

observe VCs' full choice set and realized outcomes, our approach allows us to identify, quantify, and explain two types of errors in VCs' early-stage investment decisions: VCs invest in some companies that perform predictably poorly, and pass on others that perform predictably well.

We show that VCs do not correctly weigh all entrepreneur characteristics in their decisions. They tend to select entrepreneurs whose features are representative of success – such as being male, graduates of elite schools, and based in Paris. Controlling for the extent to which a given characteristic is predictive of future performance (potentially in highly complex, interacted, and non-linear ways), we show that these representative features disproportionately influence VCs' decisions relative to their actual impact on predicted venture performance. Our approach prevents inferring biases solely based on realized outcomes, which would be appropriate if the decision-maker had perfect foresight.

Overall, our results shed light on the root cause for the VC investment patterns we document. Our results contribute to our understanding of the underlying reasons for the narrowness of the VC industry, as discussed in Lerner and Nanda (2020). Errors arising from the emphasis on representative features are likely prevalent in domains similar to venture capital, where magnitudes and tails matter more than averages, and where the resulting misallocation may have significant consequences.

Figures and Tables

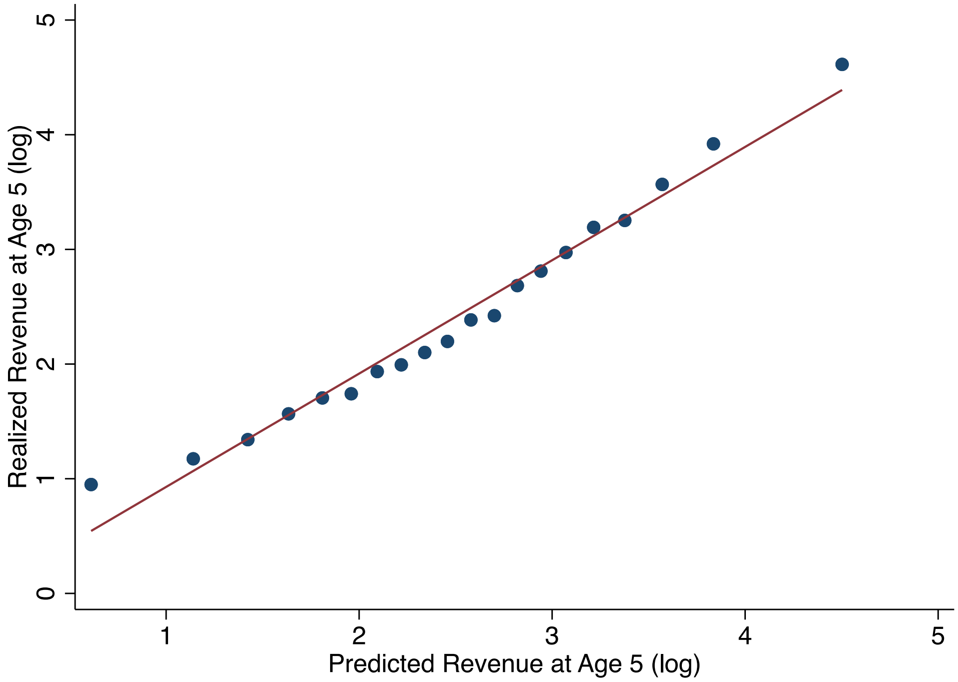


Figure 1: Algorithm's Predictive Accuracy in Test Set: All Companies. This figure shows the average observed performance (y-axis) across 20 bins of predicted performance (x-axis) among the 37,353 companies in the test set (the 2010 cohort). The company performance measure is the log revenue at age 5. Realized and predicted revenue are in thousands of euros. The predictive model was trained using 10-fold cross-validation on the sample of all companies in the 1998, 2002, and 2006 cohorts (84,583 observations).

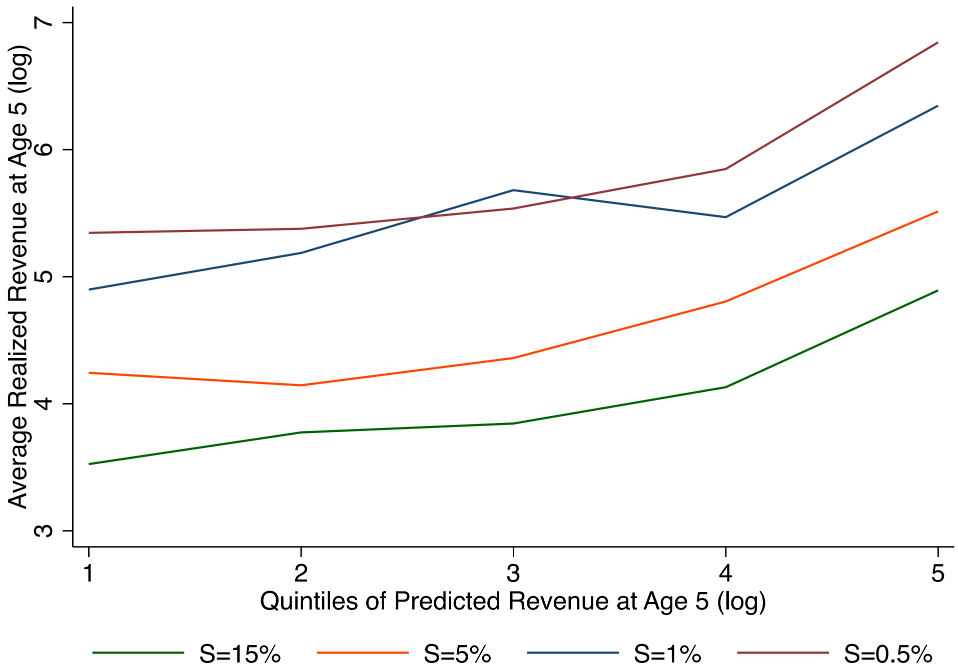


Figure 2: Algorithm's Predictive Accuracy in Test Set: Right Tail of the Predicted Performance Distribution. This figure shows the average observed performance (y-axis) across five quintiles of predicted performance (x-axis) for the companies in the top $s\%$ of the predicted performance distribution in the test set (the 2010 cohort), for four selectivity thresholds s. The performance measure is the log revenue at age 5. Realized and predicted revenue are in thousands of euros. The predictive model was trained using 10-fold cross validation on the sample of all companies in the 1998, 2002 and 2006 cohorts (84,583 observations).

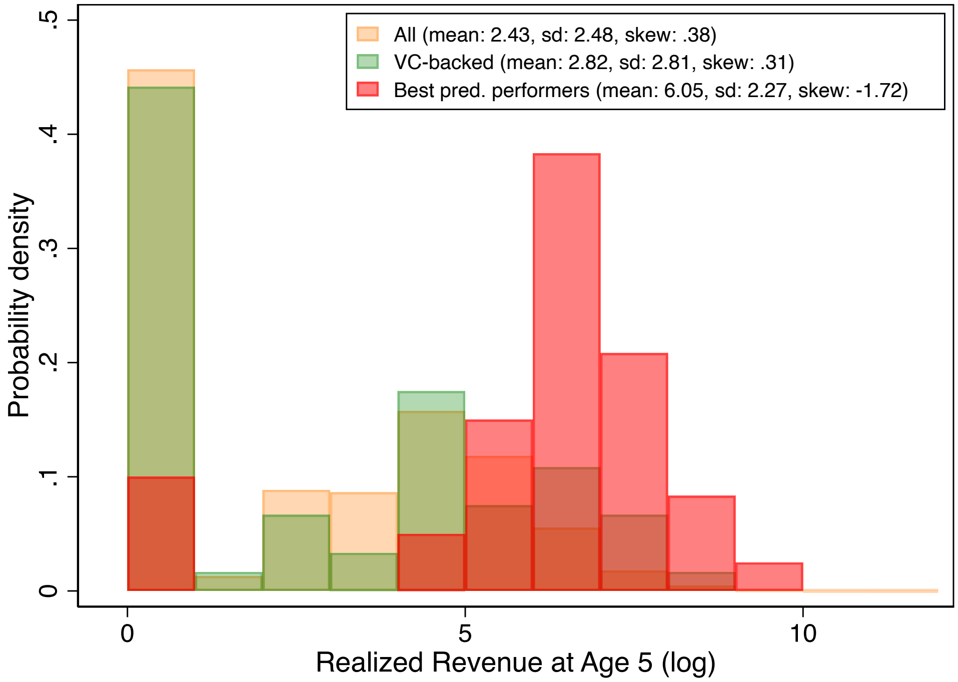


Figure 3: Performance Distribution of Companies in Test Set: All Companies, VC-backed Companies, and Companies in the Top of the Predicted Performance Distribution. This figure shows the distribution of realized performance (the log of revenue at age 5) for all 37,353 companies in the 2010 cohort (our test set in orange) as well as for the 120 VC-backed companies (in green), and for the 120 companies with the highest predicted performance (in red). We report the mean, standard deviation and skewness of revenue at age 5 (log) for each set of companies. Realized and predicted revenue are in thousands of euros. The predictive algorithm is trained on the sample of all new companies in the 1998, 2002 and 2006 cohorts using 10-fold cross validation. The predictive algorithm is "unconstrained": the top predicted performers are not restricted to a subset of companies in the test set.

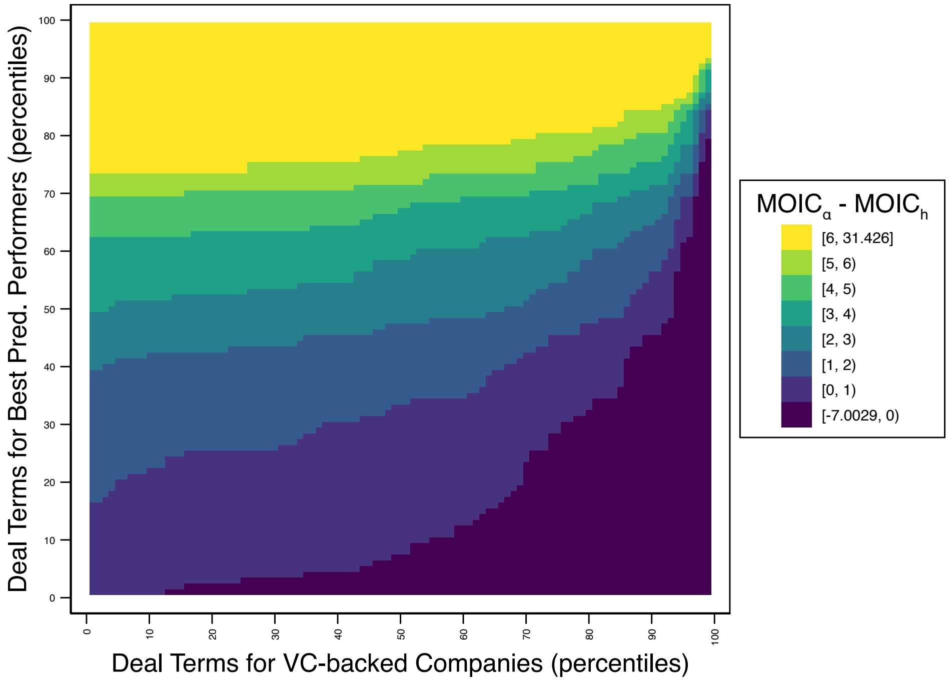


Figure 4: Deal Term Sensitivity. This figure compares the imputed portfolio multiple on invested capital (MOIC) for the best predicted performers selected by the unconstrained algorithm ($MOIC_\alpha$) with the imputed portfolio MOIC for VC-backed companies in the 2010 cohort ($MOIC_h$) under varying deal terms assumptions. The x-axis represents the percentiles of deal terms for VC-backed companies, while the y-axis represents the percentiles of deal terms for the best predicted performers. Company-level imputed MOIC is defined as in Equation (7): $MOIC_i = \frac{\delta_i * M_s * y_i}{k_i}$, where we compute the median revenue multiple at exit for each sector, M_s, using US Pitchbook data; dilution, δ_i, is assumed to be equal to 75%; and we compute the median deal terms $\frac{\delta}{k}$ using Pitchbook data on French early VC deals during 2009-2011. We obtain imputed MOICs that span the empirical distribution of deal terms. We then compute the portfolio-level imputed MOIC as the average company-level imputed MOIC for companies in $\mathcal{A}_s$, ($MOIC_\alpha$), and for VC-backed companies in the 2010 cohort ($MOIC_h$). We report the difference between the two imputed portfolio MOICs, with varying deal terms for the VC-backed companies shown on the x-axis, and for the companies in $\mathcal{A}_s$ shown on the y-axis. The color gradient represents the difference between $MOIC_\alpha$ and $MOIC_h$, with darker colors indicating smaller or negative differences.

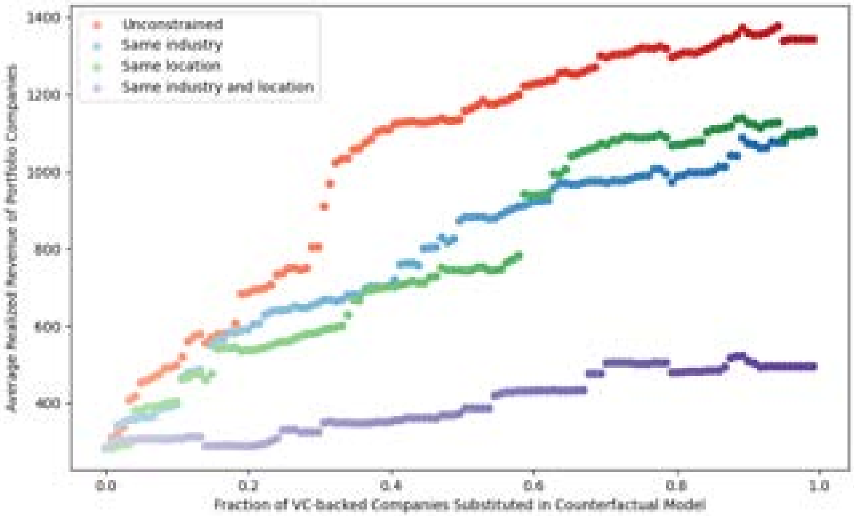


Figure 5: Counterfactual Models. This figure shows the average performance (average realized revenue at age 5 in thousands of euros) of companies selected by several counterfactual models. The counterfactual models sequentially drop VC-backed companies with the lowest $\hat{m}(x_i)$ and replace them with the best predicted performers (i.e., companies in $\mathcal{A}_s$ with the highest $\hat{m}(x_i)$) selected from various investable pools $\mathcal{D}$, ensuring the total number of portfolio companies stays constant at $|\mathcal{V}_s| = 120$. The x-axis shows the fraction of VC-backed companies replaced. The y-axis reports the average performance of the companies in the portfolio. The red line shows the performance of the unconstrained counterfactual model, that is, the best predicted performers are not constrained within a specific set of companies. Other lines represent the performance of a counterfactual model constrained to replace each VC-backed company with a company in the same industry (in blue), the same location (in green), or both the same industry *and* location (in purple).

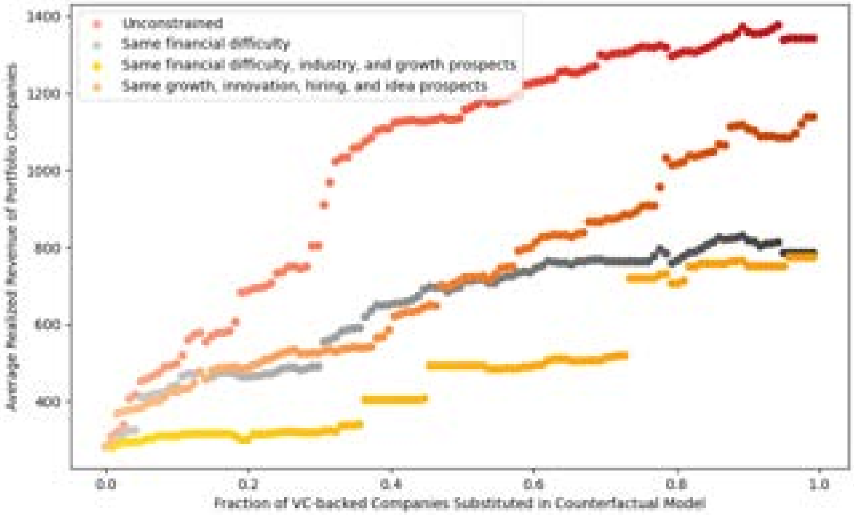


Figure 6: Counterfactual Models Constrained to VC-prone Ventures. This figure shows the average performance (average realized revenue at age 5 in thousands of euros) of companies selected by several counterfactual models. The counterfactual models sequentially drop VC-backed companies with the lowest $\hat{m}(x_i)$ and replace them with the best predicted performers (i.e., companies in $\mathcal{A}_s$ with the highest $\hat{m}(x_i)$) selected from various investable pools $\mathcal{D}$, ensuring the total number of portfolio companies stays constant at $|\mathcal{V}_s| = 120$. The investable pools $\mathcal{D}$ are designed to focus on companies that closely resemble those backed by VCs, ensuring that the best predicted performers identified by the model are realistic candidates both in terms of desirability and attractiveness to investors. The red line shows the performance of the unconstrained counterfactual model. The orange line represents the performance of a counterfactual model constrained to replace each VC-backed company it excludes with a company whose founder's responses to growth-related questions in the entrepreneur survey match those of the founder whose VC-backed company was excluded by the counterfactual model (same growth prospects, expectation to hire, innovate and motivated by a new idea). The grey line shows the performance of portfolio companies when the counterfactual model is restricted to selecting companies from the set of companies whose founder listed "securing financing" as a major difficulty in the 2010 entrepreneur survey. The yellow line shows the performance of portfolio companies when the counterfactual model is subject to the same constraint of selecting from financially constrained companies, and two additional restrictions are introduced: first, each VC-backed company excluded by the counterfactual must be replaced with a company whose founder's responses to growth related questions in the entrepreneur survey match those of the founder of the excluded VC-backed company (same growth prospects, expectation to hire, innovate and motivated by a new idea). Second, the replacement company must operate in the same industry as the excluded VC-backed company.

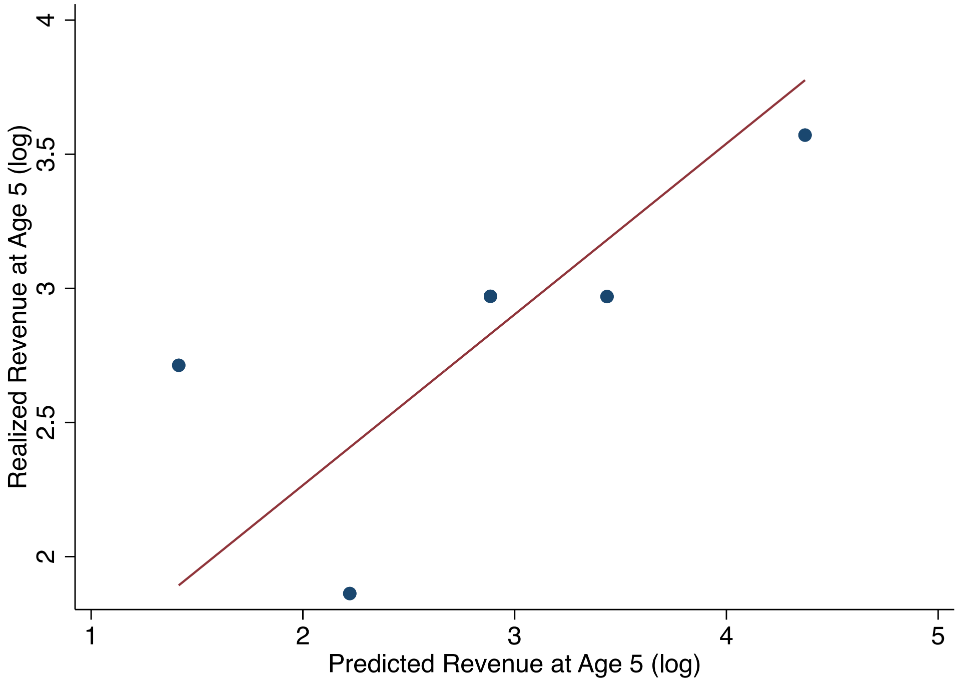


Figure 7: Algorithm Performance: VC-backed Companies in Test Set. This figure shows the average observed performance (y-axis) across 5 bins of predicted performance (x-axis) for the VC-backed companies in our 2010 cohort (our test set). Realized and predicted revenue are in thousands of euros. The predictive model is trained on the sample of all new companies in the 1998, 2002, and 2006 cohorts using 10-fold cross validation.

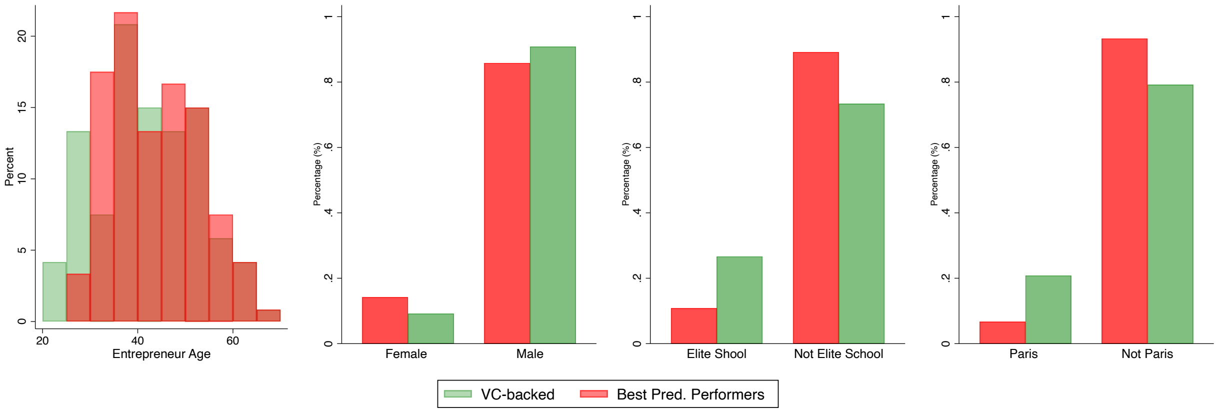


Figure 8: Entrepreneur Demographics for VC-backed companies and Best Predicted Performers. This figure shows the probability densities of founders' ages as well as the breakdown of entrepreneurs' gender, elite school attendance, and geographic location in the 2010 cohort (our test set) for VC-backed companies and predicted top performers. The predictive algorithm is trained on the sample of all new companies in the 1998, 2002 and 2006 cohorts using 10-fold cross validation. The predictive algorithm is unconstrained, that is, the top predicted performers are not restricted to a subset of companies in the test set.

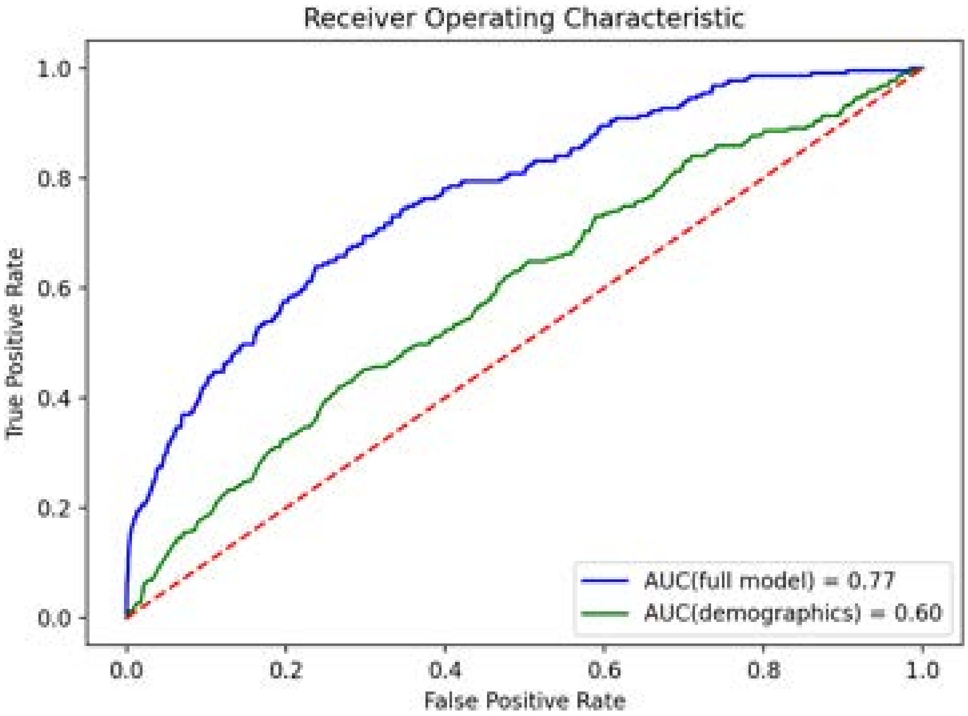


Figure 9: Area Under the Curve (AUC) of a Predictive Model of VCs' Decisions. This figure reports the AUC of an XGBClassifier model of VCs' decisions. The model was trained using a random split (70/30) over the 1998, 2002 and 2010 cohorts of entrepreneurs using 5-fold cross-validation. The AUC of .77 for the full model implies that for two randomly picked ventures, one VC-backed and one not, the odds that our model assigns a higher probability of being VC-backed to the one that is indeed VC-backed is 77%. We also report the AUC of a model that only includes entrepreneurs' demographic features (age, gender, and education level).

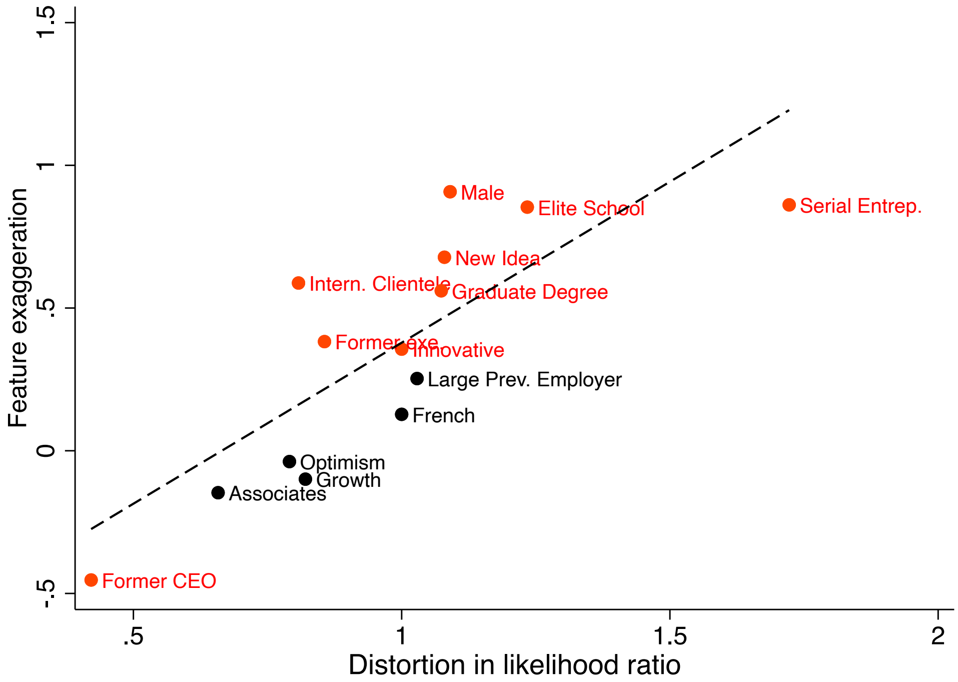


Figure 10: Overestimation of Success Forecast and Feature Exaggeration. This figure tests the relation between the distortion in estimated odds of success, (Ψ on the x-axis) and feature exaggeration on the y-axis. We measure feature exaggeration using coefficient estimates β_2 from Equation 12 (using the ML-based approach in Section 7). Ψ, the distortion in estimated odds of success is calculated as estimated odds over true odds. To estimate Ψ, we calculate the true odds of success for entrepreneurs with feature f as the probability of success conditional on f over the probability of failure conditional of f. We proxy VCs' estimated odds of success for entrepreneurs with feature f as the number of VC-backed entrepreneurs with f over the number of non-VC-backed entrepreneurs with f.

Table 1: Summary Statistics: Entrepreneur and Venture Characteristics. This table reports summary statistics for the outcome measure (Revenue at Age 5) and a subset of features in our training (Panel A) and test (Panel B) sets. We assign a zero as the (log) revenue at age 5 of firms that do not survive. The number of industries, based on a classification system similar to the two-digit SIC, and the number of regions are listed. The data come from the entrepreneur survey (SINE) conducted by the French Statistical Office, tax files from the Ministry of Finance and the firm registry (SIRENE). Appendix E describes the variables in the entrepreneur survey.

		Training						Test					
	Variable	Mean	SD	p50	p90	p99	N	Mean	SD	p50	p90	p99	N
Outcomes													
	Revenue at Age 5 (log), k euros	2.31	2.46	2.20	5.67	7.68	84,583	2.43	2.48	2.08	5.78	7.64	37,353
	Revenue at Age 5, k euros	157.55	1,420.22	8.00	289.00	2,155.00	84,583	160.09	1,168.25	7.03	322.38	2,084.73	37,353
	Alive at Age 5	0.62	0.48	1.00	1.00	1.00	84,583	0.66	0.48	1.00	1.00	1.00	37,353
Demographics													
	Entrepreneur's Age	37.78	10.00	37.00	52.00	63.00	84,583	39.72	10.66	39.00	54.00	66.00	37,353
	Female	0.28	0.45	0.00	1.00	1.00	84,583	0.28	0.45	0.00	1.00	1.00	37,353
	Entrepreneur's Nationality (FR)	0.90	0.30	1.00	1.00	1.00	84,583	0.92	0.27	1.00	1.00	1.00	37,353
	Entrepreneurial Family	0.69	0.46	1.00	1.00	1.00	84,583	0.71	0.46	1.00	1.00	1.00	37,353
Professional Background													
	Self-employed	0.37	0.48	0.00	1.00	1.00	84,583	0.32	0.47	0.00	1.00	1.00	37,353
	Previously Employed	0.51	0.50	1.00	1.00	1.00	84,583	0.55	0.50	1.00	1.00	1.00	37,353
	Part-time Entrepreneur	0.18	0.39	0.00	1.00	1.00	84,583	0.21	0.41	0.00	1.00	1.00	37,353
	Same Prior Industry	0.54	0.50	1.00	1.00	1.00	84,583	0.61	0.49	1.00	1.00	1.00	37,353
	Serial Entrepreneur	0.04	0.19	0.00	0.00	1.00	84,583	0.03	0.17	0.00	0.00	1.00	37,353
	Previously Employed in Small Firm	0.45	0.50	0.00	1.00	1.00	84,583	0.61	0.49	1.00	1.00	1.00	37,353
	Previously Inactive	0.10	0.30	0.00	0.00	1.00	84,583	0.00	0.00	0.00	0.00	0.00	37,353
	Below High School Degree	0.38	0.48	0.00	1.00	1.00	84,583	0.28	0.45	0.00	1.00	1.00	37,353
	Undergraduate Degree	0.21	0.41	0.00	1.00	1.00	84,583	0.26	0.44	0.00	1.00	1.00	37,353
	Graduate Degree	0.11	0.31	0.00	1.00	1.00	84,583	0.15	0.35	0.00	1.00	1.00	37,353
	Grande Ecole	0.04	0.21	0.00	0.00	1.00	33,806	0.06	0.24	0.00	0.00	1.00	37,353
	Completed Required Training	0.21	0.41	0.00	1.00	1.00	84,583	0.22	0.41	0.00	1.00	1.00	37,353
Motivation and Expectations													
	Expectation: Growth	0.52	0.50	1.00	1.00	1.00	84,583	0.42	0.49	0.00	1.00	1.00	37,353
	Expectation: Sustain	0.27	0.45	0.00	1.00	1.00	84,583	0.39	0.49	0.00	1.00	1.00	37,353
	Expectation: Rebound	0.07	0.25	0.00	0.00	1.00	84,583	0.08	0.28	0.00	0.00	1.00	37,353
	Motivation: Peer Entrepreneurs	0.11	0.31	0.00	1.00	1.00	84,583	0.09	0.28	0.00	0.00	1.00	37,353
	Expect to Hire	0.24	0.43	0.00	1.00	1.00	84,583	0.26	0.44	0.00	1.00	1.00	37,353
	Motivation: New Idea	0.18	0.38	0.00	1.00	1.00	84,583	0.16	0.37	0.00	1.00	1.00	37,353
	Motivation: Opportunity	0.32	0.47	0.00	1.00	1.00	84,583	0.44	0.50	0.00	1.00	1.00	37,353
	Innovation	0.34	0.47	0.00	1.00	1.00	84,583	0.44	0.50	0.00	1.00	1.00	37,353

Continued on next page

		Training						Test					
	Variable	Mean	SD	p50	p90	p99	N	Mean	SD	p50	p90	p99	N
Venture Characteristics													
	Paris-based	0.10	0.30	0.00	1.00	1.00	84,583	0.08	0.28	0.00	0.00	1.00	37,353
	Marseille-based	0.02	0.14	0.00	0.00	1.00	84,583	0.03	0.18	0.00	0.00	1.00	37,353
	Lyon-based	0.02	0.13	0.00	0.00	1.00	84,583	0.02	0.13	0.00	0.00	1.00	37,353
	Bordeaux-based	0.02	0.14	0.00	0.00	1.00	84,583	0.02	0.13	0.00	0.00	1.00	37,353
	Business Services Industry	0.16	0.36	0.00	1.00	1.00	85,119	0.14	0.35	0.00	1.00	1.00	37,685
	Health and Social Work Industry	0.04	0.20	0.00	0.00	1.00	85,119	0.04	0.19	0.00	0.00	1.00	37,685
	Construction Industry	0.18	0.39	0.00	1.00	1.00	85,119	0.17	0.37	0.00	1.00	1.00	37,685
	High tech Industry	0.01	0.12	0.00	0.00	1.00	85,119	0.02	0.13	0.00	0.00	1.00	37,685
	Energy Industry	0.00	0.02	0.00	0.00	0.00	85,119	0.03	0.16	0.00	0.00	1.00	37,685
	B2B	0.33	0.47	0.00	1.00	1.00	84,583	0.32	0.47	0.00	1.00	1.00	37,353
	B2C	0.63	0.48	1.00	1.00	1.00	84,583	0.62	0.48	1.00	1.00	1.00	37,353
	International Customers	0.07	0.25	0.00	0.00	1.00	84,583	0.05	0.21	0.00	0.00	1.00	37,353
	Local Customers	0.53	0.50	1.00	1.00	1.00	84,583	0.58	0.49	1.00	1.00	1.00	37,353
	Domestic Customers	0.14	0.35	0.00	1.00	1.00	84,583	0.15	0.36	0.00	1.00	1.00	37,353
Venture Organization													
	Co-founders	0.12	0.32	0.00	1.00	1.00	84,583	0.14	0.35	0.00	1.00	1.00	37,353
	Outsourcing: Accounting	0.64	0.48	1.00	1.00	1.00	84,583	0.74	0.44	1.00	1.00	1.00	37,353
	Number of Employees	1.59	1.52	1.00	3.00	8.00	84,583	1.60	1.55	1.00	3.00	9.00	37,353
	10+ Clients	0.63	0.48	1.00	1.00	1.00	84,583	0.63	0.48	1.00	1.00	1.00	37,353
	Number of Paid Managers	0.15	0.46	0.00	1.00	2.00	84,583	0.17	0.42	0.00	1.00	2.00	37,353
	Customers from Prior Job	0.30	0.46	0.00	1.00	1.00	84,583	0.27	0.44	0.00	1.00	1.00	37,353
	Suppliers from Prior Job	0.23	0.42	0.00	1.00	1.00	84,583	0.21	0.41	0.00	1.00	1.00	37,353
	Help from Professionals	0.03	0.17	0.00	0.00	1.00	84,583	0.10	0.30	0.00	0.00	1.00	37,353
	Help from Family	0.27	0.44	0.00	1.00	1.00	84,583	0.17	0.38	0.00	1.00	1.00	37,353
	No External Help	0.44	0.50	0.00	1.00	1.00	84,583	0.27	0.44	0.00	1.00	1.00	37,353
Financial Characteristics													
(not included as input features)	Bank Loan	0.35	0.48	0.00	1.00	1.00	83,416	0.41	0.49	0.00	1.00	1.00	37,353
	Other Loan	0.08	0.27	0.00	0.00	1.00	83,416	0.09	0.29	0.00	0.00	1.00	37,353
	No Outside Financing	0.54	0.50	1.00	1.00	1.00	83,416	0.52	0.50	1.00	1.00	1.00	37,353
	Other Firm Financing	0.05	0.21	0.00	0.00	1.00	50,777	0.04	0.19	0.00	0.00	1.00	37,353
	Grant	0.21	0.41	0.00	1.00	1.00	83,416	0.08	0.27	0.00	0.00	1.00	37,353
Industries-Locations													
	Number of Industries						48						48
	Number of Regions						322						322

Algorithm trained on	Algorithm evaluated on						
	Revenue_5 (log)	Revenue_7 (log)	Top 5% Revenue_5	Top 5% Revenue_7	Imputed Valuation (log)	Revenue Growth	Exits
Revenue_5 (log)	6.05	5.58	0.60	0.53	1.48	0.15	4
Revenue_7 (log)	5.62	5.19	0.48	0.44	1.29	0.13	3
Top 5% Revenue_5	5.50	4.94	0.57	0.52	1.36	0.11	3
Top 5% Revenue_7	5.53	5.15	0.58	0.55	1.42	0.11	3
Imputed Valuation (log)	4.16	3.97	0.23	0.19	0.83	0.13	3
Revenue Growth	2.73	2.71	0.00	0.00	0.30	0.12	0
Exit (IPO/M&A)	4.09	3.50	0.32	0.28	0.87	0.10	10
	Comparison: Average performance measures						
	Revenue_5 (log)	Revenue_7 (log)	Top 5% Revenue_5	Top 5% Revenue_7	Imputed Valuation (log)	Revenue Growth	Exits
All firms in test set	2.43	2.02	0.05	0.05	0.28	-0.05	118
VC-backed firms	2.82	2.46	0.15	0.16	0.45	-0.01	10

Table 2: Performance of the Set of Best Predicted Performers Using Various Measures of Company Performance. This table presents the average observed outcomes for the 120 best predicted performers across various predictive models, each predicting different measures of company performance. The top panel rows indicate the outcome measures used to train the models: log of firm revenue at 5 and 7 years (in thousands of euros), the likelihood of being in the top 5% of cohort revenue at 5 and 7 years, log of imputed valuation (in million euros), 5-year revenue growth, and probability of exit via acquisition or IPO. For each trained model, the columns show the average performance of the top 120 identified companies across all outcome measures. The bottom panel provides comparative data: the first row shows mean performance measures for the entire 2010 cohort (test set), while the second row presents data for VC-backed firms only. Definitions of all outcome measures are detailed in Appendix B.

Panel A: Cost of Picking Bad Pred. Performers				
Dropping VC-backed w/ pred. perf:	# Port. Companies	Multiple (survivors only)	Multiple	% Increase
	(1)	(2)	(3)	(4)
bottom 10%	108	1.33	.76	9.2
bottom 25%	90	1.32	.79	13.3
bottom 50%	60	1.55	.98	39.9

Panel B: Cost of Passing On Best Pred. Performers				
	# Port. Companies	Multiple (survivors only)	Multiple	% Increase
	(1)	(2)	(3)	(4)
top 1%	373	2.52	2.05	193
top 0.5%	186	2.8	2.38	240
top 120	120	2.99	2.54	262.8

Table 3: What is the Cost of VCs' Errors? This table reports the imputed multiples (MOIC) on several portfolios designed to isolate the cost of selecting bad-predicted performers (Panel A) and the cost of passing on good predicted performers (Panel B). Rows 1 through 3 of Panel A contain the number, average multiple, and percentage increase in VCs' portfolio performance when dropping portfolio companies in the bottom 10%, 25%, and 50% of VC-backed companies' predicted performance, respectively. Rows 1 through 3 of Panel B contain the number, average multiple, and percentage increase in VCs' portfolio performance when selecting the top 1%, 0.05%, and the best 120 companies in terms of predicted performance in the 2010 cohort, respectively.

		Test Set (2010 cohort)						
		VC-backed			Best Pred. Performers			Difference
		Mean	SD	N	Mean	SD	N	T-Test
Predicted Performance								
	Pred. Revenue at Age 5 (log), k euros	2.87	1.07	120	5.81	0.45	120	-2.95***
Outcomes								
	Revenue at Age 5 (log), k euros	2.82	2.81	120	6.05	2.27	120	-3.24***
	Revenue at Age 5, k euros	283.21	686.47	120	1342.57	1998.41	120	-1059.35***
	alive_5	0.69	0.46	120	0.91	0.29	120	-0.22***
Founder Demographics								
	Entrepreneur's Age	41.26	10.58	120	43.23	9.45	120	-1.97
	Founder's Nationality (FR)	0.94	0.24	120	0.99	0.09	120	-0.05**
	Female	0.09	0.29	120	0.14	0.35	120	-0.05
Founder Professional Background								
	Same Prior Industry	0.52	0.50	120	0.91	0.29	120	-0.39***
	Serial Entrepreneur	0.10	0.30	120	0.03	0.16	120	0.07**
	Previously Employed in Small Firm	0.54	0.50	120	0.42	0.50	120	0.12*
	Graduate Degree	0.37	0.48	120	0.46	0.50	120	-0.09
	Grande Ecole	0.27	0.44	120	0.11	0.31	120	0.16***
Founder Motivation and Expectations								
	Expectation: Growth	0.57	0.50	120	0.57	0.50	120	0.01
	Motivation: Successful Peer Entrepreneurs	0.06	0.24	120	0.08	0.28	120	-0.03
	Expect to Hire	0.51	0.50	120	0.60	0.49	120	-0.09
	Motivation: New Idea	0.39	0.49	120	0.05	0.22	120	0.34***
	Motivation: Opportunity	0.38	0.49	120	0.58	0.50	120	-0.21***
	Innovation	0.68	0.47	120	0.39	0.49	120	0.29***
Venture Characteristics								
	Paris-based	0.21	0.41	120	0.07	0.25	120	0.14***
	High-Tech Industry	0.13	0.34	120	0.01	0.09	120	0.12***
Organization								
	Outsourcing: Accounting	0.91	0.29	120	0.82	0.39	120	0.09**
	Outsourcing: Management	0.09	0.29	120	0.26	0.44	120	-0.17***
	Outsourcing: Logistics	0.15	0.36	120	0.38	0.49	120	-0.23***
	Number of Employees	2.30	2.82	120	6.67	4.50	120	-4.38***
Industries-Locations								
	Number of Industries	.	.	37	.	.	25	
	Number of Regions	.	.	68	.	.	78	

Table 4: Differences Between VC-backed Companies and Best Predicted Performers. This table reports selected summary statistics for VC-backed and best predicted performers. We report t-tests for the difference in means. We assign a zero as the (log) revenue at age 5 of companies that do not survive. The data come from the entrepreneur survey (SINE) conducted by the French Statistical Office, tax files from the Ministry of Finance and the firm registry (SIRENE). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	VC-backed							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\widehat{h}(X)$	1.5***				1.6***	1.5***	1.6***	1.6***
	(.043)				(.044)	(.044)	(.044)	(.045)
$\widehat{m}(X)_{top5\_Revenue_5}$		.058***			-.0059			-.0067
		(.0066)			(.0067)			(.0087)
$\widehat{m}(X)_{exit}$			.49***			.042		.077
			(.055)			(.056)		(.062)
$\widehat{m}(X)_{Log\_Revenue_5}$				.0038***			-.0006	-.00047
				(.00053)			(.00053)	(.00064)
adj.R^2	.047	.0029	.0029	.0018	.047	.047	.047	.047
Observations	26,440	26,440	26,440	26,440	26,440	26,440	26,440	26,440

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 5: VCs' Decision Model is not Subsumed by Performance Predictions. We test whether our predictions of VCs' decisions are subsumed by predicted performance. This table reports the results of a regression of VC-backed status on the predictions from four estimators. $\widehat{h}(X)$ is a vector of predicted probabilities for whether a company is VC-backed; $\widehat{m}(X)_{top5Revenue_5}$ is a vector of predicted probabilities for whether a company will be in the top 5% of its cohort in terms of revenue at age 5; $\widehat{m}(X)_{exit}$ is a vector of predicted probabilities for whether a company will go through an IPO or M&A. $\widehat{m}(X)_{Log(Revenue_5)}$ is a vector of predicted values for the log of revenue at age 5 (in thousands of euros). All models are trained on a random split of 70% of the observations in the 1998, 2002, and 2010 cohorts, and tested out-of-sample on the remaining 30% of observations (see Section 6.2). These results pertain to the test set only.

Table 6: Full vs. Simple Models. We test whether VCs do not correctly weigh all entrepreneur characteristics in their decisions. This table reports the results of a regression of VC-backed status on predictions of *Exit* from our full model $\widehat{m}(X)_{Exit}$ and from the simple models $\widehat{m}_{simple}(.)$, which take as inputs only a subset X of features. All estimators in this table predict *Exit*, a dummy equal to one for firms that were acquired or became public. The algorithms are trained on the sample of all new companies in the 1998, 2002, and 2006 cohorts and tested on the 2010 cohort of entrepreneurs. Estimator $\widehat{m}_{simple}$(personal features) is trained by taking as inputs the founding entrepreneur's age, gender, education, nationality, and whether there are entrepreneurs among her relatives. Estimator $\widehat{m}_{simple}$(optimism) is trained taking as input a dummy equal to one if the entrepreneur expects to grow or hire. Estimator $\widehat{m}_{simple}$(startup traction) is trained taking as inputs the total number of workers, the number of clients, and the client's location.

Panel A: Entrepreneurs' features

	VC-backed									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$\widehat{m}(X)$(Exit)	.72***	.7***	.72***	.71***	.7***	.64***	.72***	.72***	.73***	.71***
	(.037)	(.038)	(.037)	(.037)	(.037)	(.034)	(.037)	(.037)	(.04)	(.037)
$\widehat{m}_{simple}$(Personal Characteristics)		.25**								
		(.11)								
$\widehat{m}_{simple}$(Age)			.068							
			(.16)							
$\widehat{m}_{simple}$(Gender)				.91***						
				(.33)						
$\widehat{m}_{simple}$(Graduate Degree)					.56***					
					(.16)					
$\widehat{m}_{simple}$(Elite School)						.85***				
						(.16)				
$\widehat{m}_{simple}$(French Nationality)							.13			
							(1.2)			
$\widehat{m}_{simple}$(Relatives)								-.6		
								(.66)		
$\widehat{m}_{simple}$(Optimism)									-.038	
									(.1)	
$\widehat{m}_{simple}$(Serial Entrepreneur)										.86***
										(.32)
Adj.R^2	.01	.01	.01	.01	.01	.012	.01	.01	.01	.01
Observations	37,353	37,353	37,353	37,353	37,353	37,353	37,353	37,353	37,353	37,353

Standard errors in parentheses

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Panel B: New ventures' features

	VC backed								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\widehat{m}(X)$(Exit)	.72***	.71***	.7***	.68***	.72***	.72***	.71***	.72***	.72***
	(.037)	(.038)	(.037)	(.04)	(.038)	(.037)	(.037)	(.037)	(.037)
$\widehat{m}_{simple}$(Innovative)		.36**							
		(.16)							
$\widehat{m}_{simple}$(New Idea)			.68***						
			(.18)						
$\widehat{m}_{simple}$(Startup Traction)				.24***					
				(.091)					
$\widehat{m}_{simple}$(High-tech Ind.)					.092				
					(.14)				
$\widehat{m}_{simple}$(Scientific R&D Ind.)						3.3***			
						(.67)			
$\widehat{m}_{simple}$(Paris)							.97***		
							(.33)		
$\widehat{m}_{simple}$(Marseille)								-.3	
								(.8)	
$\widehat{m}_{simple}$(Lyon)									-2.5
									(2.4)
Adj.R^2	.01	.01	.011	.01	.01	.011	.01	.01	.01
Observations	37,353	37,353	37,353	37,353	37,353	37,353	37,353	37,353	37,353

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Feature	Top 1%	Bottom 99%	Representativeness of best performers $\frac{Pr(X_i \mid \text{Top1})}{Pr(X_i \mid \text{Bottom99})}$
	(1)	(2)	(3)
Male	80.95	69.57	1.16
Graduate Degree	18.11	10.53	1.72
Elite School	3.91	1.76	2.22
Optimism	53.02	19.63	2.7
Serial Entrepreneur	13.02	3.68	3.53
Paris-based	16.92	10.14	1.67
High-tech Ind.	5.92	4.13	1.43

Table 7: Stereotypes of the Most Successful Entrepreneurs. This table reports the fraction of entrepreneurs with a given characteristic, as listed in the rows, among the best performing companies in column 1 (top 1% of revenue at age 5) and among all the other companies in column 2 (bottom 99% of revenue at age 5). A given characteristic is representative (or stereotypical) of the best performing companies if it scores high on the representativeness ratio (column 3) of the percentage in column 1 over that in column 2. The training data set (the 1998, 2002, and 2006 cohorts) is used in this table.

Bibliography

Azoulay, Pierre, Benjamin F Jones, J Daniel Kim, and Javier Miranda. 2020. "Age and high-growth entrepreneurship." *American Economic Review: Insights*, 2(1): 65–82.

Balachandra, Lakshmi, Tony Briggs, Kim Eddleston, and Candida Brush. 2019. "Don't pitch like a girl: How gender stereotypes influence investor decisions." *Entrepreneurship Theory and Practice*, 43(1): 116–137.

Bernstein, Shai, Arthur Korteweg, and Kevin Laws. 2017. "Attracting early-stage investors: Evidence from a randomized field experiment." *Journal of Finance*, 72(2): 509–538.

Bian, Bo, Yingxiang Li, and Casimiro Antonio Nigro. 2022. "Conflicting Fiduciary Duties and Fire Sales of VC-backed Start-ups." Working paper.

Bonelli, Maxime. 2023. "Data-drive Investors." *Available at SSRN 4362173*.

Bordalo, Pedro, John J Conlon, Nicola Gennaioli, Spencer Y Kwon, and Andrei Shleifer. 2023. "Memory and Probability*." *The Quarterly Journal of Economics*, 138(1): 265–311.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. "Stereotypes." *Quarterly Journal of Economics*, 131(4): 1753–1794.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2020. "Memory, Attention, and Choice*." *The Quarterly Journal of Economics*, 135(3): 1399–1442.

Bpifrance. 2021. "Les fonds partenaires et Bpifrance: Bilan 2021."

Brown, Gregory W, Robert S Harris, Wendy Hu, Tim Jenkinson, Steven N Kaplan, and David T Robinson. 2020. "Private equity portfolio companies: A first look at Burgiss holdings data." *Available at SSRN 3532444*.

Bryan, Kevin, and Jorge Guzman. 2021. "Entrepreneurial Migration." *Available at SSRN*.

Calder-Wang, Sophie, and Paul A Gompers. 2021. "And the children shall lead: Gender diversity and performance in venture capital." *Journal of Financial Economics*.

Catalini, Christian, Jorge Guzman, and Scott Stern. 2019. "Hidden in Plain Sight: Venture Growth with or without Venture Capital." National Bureau of Economic Research Working Paper 26521.

Chemmanur, Thomas J, Karthik Krishnan, and Debarshi K Nandy. 2011. "How does venture capital financing improve efficiency in private firms? A look beneath the surface." *Review of Financial Studies*, 24(12): 4037–4090.

Chen, Henry, Paul Gompers, Anna Kovner, and Josh Lerner. 2010. "Buy local? The geography of venture capital." *Journal of Urban Economics*, 67(1): 90–102. Special Issue: Cities and Entrepreneurship.

Chen, Hugh, Joseph D. Janizek, Scott Lundberg, and Su-In Lee. 2020. "True to the Model or True to the Data?"

Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost: A Scalable Tree Boosting System.” *CoRR*, abs/1603.02754.

Chung, Ji-Woong, Berk A. Sensoy, Léa H Stern, and Michael Weisbach. 2012. “Pay for Performance from Future Fund Flows: The Case of Private Equity.” *Review of Financial Studies*, 25(11): 3259–3304.

Cong, Lin William, and Yizhou Xiao. 2021. “Persistent Blessings of Luck: Theory and an Application to Venture Capital.” *Review of Financial Studies*, 35(3): 1183–1221.

Conlon, John, and Dev Patel. 2022. “What Jobs Come To Mind? Stereotypes about Fields of Study.” *Working Paper.*

Cook, Lisa D, Matt Marx, and Emmanuel Yimfor. 2023. “Funding Black High-Growth Startups.” Working paper.

Cour des comptes. 2023. “Rapport portant sur une entreprise publique: Les activités d’investissement de Bpifrance.” Exercices 2012-2021.

Davenport, Diag. 2022. “Predictably Bad Investments: Evidence from Venture Capitalists.” *Available at SSRN 4135861.*

Ekeland, Marie, Augustin Landier, and Jean Tirole. 2016. “Strengthening French venture capital.” *Notes du conseil danalyse economique*, 33(6): 1–12.

Erel, Isil, Léa H Stern, Chenhao Tan, and Michael S Weisbach. 2021. “Selecting Directors Using Machine Learning.” *Review of Financial Studies*, 34(7): 3226–3264.

Ewens, Michael. 2023. “Gender and race in entrepreneurial finance.”

Ewens, Michael, and Richard R Townsend. 2020. “Are early stage investors biased against women?” *Journal of Financial Economics*, 135(3): 653–677.

Fairlie, Robert, Alicia Robb, and David T. Robinson. 2022. “Black and White: Access to Capital Among Minority-Owned Start-ups.” *Management Science*, 68(4): 2377–2400.

Fazio, Catherine, Jorge Guzman, Fiona Murray, and Scott Stern. 2016. “A new view of the skew: Quantitative assessment of the quality of American entrepreneurship.” *Kauffman Foundation New Entrepreneurial Growth.*

Ferrati, Francesco, Moreno Muffatto, et al. 2021. “Entrepreneurial finance: emerging approaches using machine learning and big data.” *Foundations and Trends in Entrepreneurship*, 17(3): 232–329.

Frisch, Ragnar, and Frederick V Waugh. 1933. “Partial time regressions as compared with individual trends.” *Econometrica*, 387–401.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2022. “Predictably Unequal? The Effects of Machine Learning on Credit Markets.” *Journal of Finance*, 77(1): 5–47.

Giglio, Stefano, Matteo Maggiori, Johannes Stroebel, and Stephen Utkus. 2021. “Five Facts about Beliefs and Portfolios.” *American Economic Review*, 111(5): 1481–1522.

Gilles, Fabrice, Yannick L'Horty, and Ferhat Mihoubi. 2023. "Qu'avons-nous appris en évaluant les accélérateurs de Bpifrance?" *Revue d'économie financière*, , (2): 229–250.

Gompers, Paul A, and Steven N Kaplan. 2022. *Advanced introduction to private equity.* Edward Elgar Publishing.

Gompers, Paul, and Josh Lerner. 2001. "The venture capital revolution." *Journal of Economic Perspectives*, 15(2): 145–168.

Gompers, Paul A, Will Gornall, Steven N Kaplan, and Ilya A Strebulaev. 2020. "How do venture capitalists make decisions?" *Journal of Financial Economics*, 135(1): 169–190.

Gompers, Paul, Will Gornall, Steven N Kaplan, and Ilya A Strebulaev. 2021. "Venture capitalists and COVID-19." *Journal of Financial and Quantitative Analysis*, 56(7): 2474–2499.

Gornall, Will, and Ilya A Strebulaev. 2020. "Gender, race, and entrepreneurship: A randomized field experiment on venture capitalists and angels." *Available at SSRN 3301982.*

Guzman, Jorge, and Scott Stern. 2020. "The State of American Entrepreneurship: New Estimates of the Quantity and Quality of Entrepreneurship for 32 US States, 1988–2014." *American Economic Journal: Economic Policy*, 12(4): 212–43.

Harris, Robert S, Tim Jenkinson, and Steven N Kaplan. 2014. "Private equity performance: What do we know?" *Journal of Finance*, 69(5): 1851–1882.

Hebert, Camille. 2023. "Gender stereotypes and entrepreneur financing."

Hellmann, Thomas, and Manju Puri. 2000. "The interaction between product market and financing strategy: The role of venture capital." *Review of Financial Studies*, 13(4): 959–984.

Hellmann, Thomas, and Manju Puri. 2002. "Venture capital and the professionalization of start-up firms: Empirical evidence." *Journal of Finance*, 57(1): 169–197.

Hochberg, Yael V., Alexander Ljungqvist, and Annette Vissing-Jørgensen. 2013. "Informational Holdup and Performance Persistence in Venture Capital." *The Review of Financial Studies*, 27(1): 102–152.

Hochberg, Yael V, Alexander Ljungqvist, and Yang Lu. 2007. "Whom you know matters: Venture capital networks and investment performance." *Journal of Finance*, 62(1): 251–301.

Hombert, Johan, Antoinette Schoar, David Sraer, and David Thesmar. 2020. "Can unemployment insurance spur entrepreneurial activity? Evidence from France." *Journal of Finance*, 75(3): 1247–1285.

Howell, Sabrina T, and Ramana Nanda. 2019. "Networking frictions in venture capital, and the gender gap in entrepreneurship." National Bureau of Economic Research.

Hu, Allen, and Song Ma. 2021. "Persuading Investors: A Video-Based Study." National Bureau of Economic Research Working Paper 29048.

Jang, Young Soo, and Steven N Kaplan. 2023. "Venture Capital Start-up Selection." *Available at SSRN.*

Kanze, Dana, Laura Huang, Mark A. Conley, and E. Tory Higgins. 2018. "We Ask Men to Win and Women Not to Lose: Closing the Gender Gap in Startup Funding." *Academy of Management Journal*, 61(2): 586–614.

Kaplan, Steven N, and Josh Lerner. 2016. "Venture capital data: Opportunities and challenges." *Measuring entrepreneurial businesses: Current knowledge and challenges*, 413–431.

Kaplan, Steven N, and Per ER Strömberg. 2004. "Characteristics, contracts, and actions: Evidence from venture capitalist analyses." *Journal of Finance*, 59(5): 2177–2210.

Kaplan, Steven, Per Strömberg, and Berk Sensoy. 2002. "How Well Do Venture Capital Databases Reflect Actual Investments?" Working paper.

Kerr, William R., Ramana Nanda, and Matthew Rhodes-Kropf. 2014. "Entrepreneurship as Experimentation." *Journal of Economic Perspectives*, 28(3): 25–48.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human decisions and machine predictions." *Quarterly Journal of Economics*, 133(1): 237–293.

Lamont, Michele. 2002. "Money, Morals and Manners: The Culture of the French and American Upper-Middle Class." *Bibliovault OAI Repository, the University of Chicago Press*, 72.

Landier, Augustin, and David Thesmar. 2008. "Financial contracting with optimistic entrepreneurs." *Review of Financial Studies*, 22(1): 117–150.

Lerner, Josh, and Ramana Nanda. 2020. "Venture capital's role in financing innovation: What we know and how much we still need to learn." *Journal of Economic Perspectives*, 34(3): 237–61.

Liebersohn, Jack, and Victor Lyonnet. 2024. "Declining Business Formation and the Rise of Superstar Firms."

Ludwig, Jens, and Sendhil Mullainathan. 2024. "Machine Learning as a Tool for Hypothesis Generation*." *The Quarterly Journal of Economics*, 139(2): 751–827.

Lundberg, Scott, and Su-In Lee. 2017. "A unified approach to interpreting model predictions." *CoRR*, abs/1705.07874.

Malenko, Andrey, Ramana Nanda, Matthew Rhodes-Kropf, and Savitar Sundaresan. 2021. "Catching Outliers: Committee Voting and the Limits of Consensus when Financing Innovation." Harvard Business School Harvard Business School Working Papers 21-131.

Mallaby, Sebastian. 2022. *The power law: Venture capital and the making of the new future.* Penguin.

Masters, Blake, and Peter Thiel. 2014. *Zero to one: notes on start ups, or how to build the future.* Random House.

Mullainathan, Sendhil. 2002. "A Memory-Based Model of Bounded Rationality*." *The Quarterly Journal of Economics*, 117(3): 735–774.

Mullainathan, Sendhil, and Ziad Obermeyer. 2022. "Diagnosing physician error: A machine learning approach to low-value health care." *Quarterly Journal of Economics*, 137(2): 679–727.

Puri, Manju, and Rebecca Zarutskie. 2012. "On the life cycle dynamics of venture-capital-and non-venture-capital-financed firms." *Journal of Finance*, 67(6): 2247–2293.

Queiró, Francisco. 2021. "Entrepreneurial human capital and firm dynamics."

Raina, Sahil. 2019. "VCs, founders, and the performance gender gap."

Rambachan, Ashesh. 2024. "Identifying Prediction Mistakes in Observational Data*." *The Quarterly Journal of Economics*, 139(3): 1665–1711.

Röhm, Sarah, Markus Bick, and Martin Boeckle. 2022. "The Impact of Artificial Intelligence on the Investment Decision Process in Venture Capital Firms." 420–435. Berlin, Heidelberg:Springer-Verlag.

Schwienbacher, Armin. 2008. "Venture capital investment practices in Europe and the United States." *Financial Markets and Portfolio Management*, 22(3): 195–217.

Sebag, David-James, Rima Maitrehenry, and Gide Loyrette Nouel. 2020. "Venture capital investment in France: market and regulatory overview."

Strebulaev, Ilia, and Alex Dang. 2024. *The venture mindset: how to make smarter bets and achieve extraordinary growth.* Penguin.

Te, Yiea-Funk, Michèle Wieland, Martin Frey, Asya Pyatigorskaya, Penny Schiffer, and Helmut Grabner. 2022. "Predicting the Success of Startups Using Crunchbase and Linkedin Data." *Available at SSRN 4217648.*

Tversky, Amos, and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty." *Science*, 185(4157): 1124–1131.

Wachter, Jessica, and Michael Kahana. 2023. "A retrieved-context theory of financial decisions." *Working Paper.*

Żbikowski, Kamil, and Piotr Antosiuk. 2021. "A machine learning, bias-free approach for predicting business success using Crunchbase data." *Information Processing & Management*, 58(4): 102555.

Appendix A Description of a Subset of the Entrepreneur Survey Variables

Variables	Description
Entrepreneur demographics	
Entrepreneur's age	The entrepreneur's age in years.
Female	Dummy equal to one if the entrepreneur is female.
Entrepreneur's Nationality (FR)	Dummy equal to one if the entrepreneur is French.
Entrepreneurial family	Dummy equal to one if the entrepreneur has relatives who are entrepreneurs.
Entrepreneur professional background	
Self-employed	Dummy equal to one if the new company status is such that the entrepreneur is self-employed (*code juridique* starts with 1).
Previously employed	Dummy equal to one if the entrepreneur was employed prior to creating the new company.
Part-time Entrepreneur	Dummy equal to one if the entrepreneur is working for another firm while creating the new company.
Same Prior Industry	Dummy equal to one if the entrepreneur has worked in the same industry the new company is created in.
Serial entrepreneur	Dummy equal to one if the entrepreneur has created at least one firm before.
Previously employed in small firm	Dummy equal to one if the entrepreneur was employed in a firm with less than 10 employees prior to creating the new company.
Previously inactive	Dummy equal to one if the entrepreneur was either previously unemployed or not yet part of the workforce.
Below high school degree	Dummy equal to one if the entrepreneur's highest degree is below a high school degree.
Undergraduate degree	Dummy equal to one if the entrepreneur's highest degree is an undergraduate degree (2 or 3 years post high school).
Graduate degree	Dummy equal to one if the entrepreneur's highest degree is a graduate degree (5 or more years post high school).
Elite School	Dummy equal to one if the entrepreneur graduated from a Grande école or engineering school. This variable is not used in the algorithm training because it is not available for the 1998 and 2002 cohorts of the entrepreneur survey.
Completed required training	Dummy equal to one if the entrepreneur completed required training to create the new company.
Entrepreneur motivation and expectations	
Expectation: growth	Dummy equal to one if the entrepreneur expects the new company's business to grow over the next 12 months.
Expectation: sustain	Dummy equal to one if the entrepreneur expects to sustain the new company's business at its current level over the next 12 months.
Expectation: rebound	Dummy equal to one if the entrepreneur expects the new company's business to improve over the next 12 months.
Expectation: future hires	Dummy equal to one if the entrepreneur expects to hire over the next 12 months.
Expectation: no future hires	Dummy equal to one if the entrepreneur does not expect to hire over the next 12 months.
Motivation: successful peer entrepreneurs	Dummy equal to one if the entrepreneur was inspired by a successful entrepreneur they are related to.
Motivation: new idea	Dummy equal to one if the entrepreneur had a new idea for a product, service, or a new market.

Continued next page

Description of Variables (continued)

Variables	Description
Motivation: opportunity	Dummy equal to one if the entrepreneur had an opportunity to create a firm.
Innovation	Dummy equal to one if the entrepreneur is bringing a new innovation in terms of marketing, product, services, or organization.
Innovation: marketing, product, or services	Dummy equal to one if the entrepreneur's innovation is in terms of marketing, product, or services (i.e., not organization).
Venture characteristics	
Paris-based	Dummy equal to one if the new company is located in Paris.
Marseille-based	Dummy equal to one if the new company is located in Marseille.
Lyon-based	Dummy equal to one if the new company is located in Lyon.
Bordeaux-based	Dummy equal to one if the new company is located in Bordeaux.
Specialized construction industry	Dummy equal to one if the new company is in the specialized construction industry (naf2 code 43).
Retail trade industry	Dummy equal to one if the new company is in the retail trade industry (naf2 code 47).
High-tech industry	Dummy equal to one if the new company is in the high-tech industry, as defined by the OECD (naf2 codes 21, 26, 30, 32, 46, 58, 61, 62, 63, 95).
Scientific R&D industry	Dummy equal to one if the new company is in the scientific R&D industry (naf2 code 72).
B2B	Dummy equal to one if the new company is business-to-business.
B2C	Dummy equal to one if the new company is business-to-customer.
International customers	Dummy equal to one if the new company has international customers.
Local customers	Dummy equal to one if the new company has local customers.
Domestic customers	Dummy equal to one if the new company has domestic customers.
Co-founders	Dummy equal to one if the entrepreneur has co-founders.
Outsourcing: Accounting	Dummy equal to one if the new company outsources accounting services.
Number of employees	The number of employees in the new company.
10+ clients	Dummy equal to one if the new company has more than 10 customers.
Number of unpaid managers	The number of managers in the new company who are not employed.
Number of paid managers	The number of managers in the new company who are employed.
Customers from prior job	Dummy equal to one if the entrepreneur has customers they met in their previous job.
Suppliers from prior job	Dummy equal to one if the entrepreneur has suppliers they met in their previous job.
Help from professionals	Dummy equal to one if the entrepreneur sought help from professionals to create their firm.
Help from family	Dummy equal to one if the entrepreneur sought help from family members to create their firm.
No external help	Dummy equal to one if the entrepreneur did not seek for external help to create their firm.
Bank loan	Dummy equal to one if the entrepreneur obtained a bank loan to finance their firm.
Other loan	Dummy equal to one if the entrepreneur obtained another type of loan to finance their firm.
Personal resources	Dummy equal to one if the entrepreneur only used their personal resources to finance their firm.
Other firm financing	Dummy equal to one if the entrepreneur obtained capital from other companies to finance their firm.
Public grant	Dummy equal to one if the entrepreneur received a public grant to finance their firm.

Appendix B Outcome Measures

We provide the data source and details on the construction of the outcome variables used to train and test the models in Table 2, and additional summary statistics on other outcome measures in Table B.1.

Revenue (log). Ventures' revenue at age 5 and age 7 are in thousands of euros. These variables are extracted from the tax files used by the Ministry of Finance for corporate tax collection purposes.

Imputed valuation. To impute valuations, we use early-stage deals data in Pitchbook to calculate the median exit valuation multiple for each industry. We multiply each venture's revenue at age 5 (in millions) by their industry's median exit valuation multiple. The resulting valuations are in millions of euros. In Table 2, we take the natural log of this imputed valuation.

Revenue growth. Revenue growth is the average DHS growth rate between age 0 and 5, for firms with at least 2 DHS growth rates available by age 5. Ventures' revenue are extracted from the tax files used by the Ministry of Finance for corporate tax collection purposes.

Top 5% revenue. Top 5% revenue is an indicator variable equal to one for ventures whose revenue is in the top 5% of their respective cohort's revenue distribution. Ventures' revenue are extracted from the tax files used by the Ministry of Finance for corporate tax collection purposes.

Exit (IPO/M&A). We follow the literature and define exits as companies that were either acquired or went public (Fazio et al., 2016; Guzman and Stern, 2020). We create a dummy variable *Exit*, equal to one for such home runs, without imposing a horizon within which the exit must occur. To identify exits in our sample, we match the French administrative data with data from Pitchbook, CBInsights, Preqin, SDC, VentureXpert, CapitalIQ, Orbis, and Crunchbase. We identify 118 exits among the 2010 cohort of entrepreneurs, including 116 acquisitions and 2 IPOs. We consider these exits to be associated with positive returns for the VCs that have made seed or early stage investments.[46] We do not treat later-stage VC rounds as exits. While these are usually positive events for early-stage investors, they are also largely endogenous to VCs making seed or early-stage investments.

[46] Due to data limitations, we are unable to ensure that acquisitions are made at a premium or that the initial VC (if any) exits the deal. However, we recognize that VCs under liquidity pressure sometimes resort to fire sales, characterized by substantially lower sale prices and positive abnormal returns for the acquirer (see Bian, Li and Nigro, 2022). It is therefore possible that some of the acquisitions we identify were traded at a discount relative to VCs' early stage investment.

Table B.1: Summary Statistics: Outcomes. This table reports summary statistics for various outcome measures in our training (Panel A) and test (Panel B) sets. We assign a zero as the (log) revenue at age 5 of firms that do not survive. The data come from the entrepreneur survey (SINE) conducted by the French Statistical Office, tax files from the Ministry of Finance and the firm registry (SIRENE).

		Training						Test					
	Variable	Mean	SD	p50	p90	p99	N	Mean	SD	p50	p90	p99	N
Outcomes													
	Revenue at Age 5 (log), k euros	2.31	2.46	2.20	5.67	7.68	84,583	2.43	2.48	2.08	5.78	7.64	37,353
	Top 5% Revenue at Age 5	0.05	0.22	0.00	0.00	1.00	84,583	0.05	0.22	0.00	0.00	1.00	37,353
	Alive at Age 5	0.62	0.48	1.00	1.00	1.00	84,583	0.66	0.48	1.00	1.00	1.00	37,353
	Revenue at Age 7 (log)	1.95	2.43	0.00	5.55	7.67	84,583	2.02	2.51	0.00	5.73	7.76	37,353
	Top 5% Revenue at Age 7, k euros	0.05	0.22	0.00	0.00	1.00	84,583	0.05	0.22	0.00	0.00	1.00	37,353
	Alive at Age 7	0.54	0.50	1.00	1.00	1.00	84,583	0.58	0.49	1.00	1.00	1.00	37,353
	Ave. Revenue Growth	0.15	0.49	0.00	0.50	2.00	79,734	-0.05	0.48	0.00	0.29	0.67	37,066
	Imputed Valuation (log), mn euros	0.25	0.47	0.02	0.77	2.26	84,583	0.28	0.50	0.03	0.88	2.30	37,353
	Exit (IPO or M&A)	0.00	0.05	0.00	0.00	0.00	84,583	0.00	0.06	0.00	0.00	0.00	37,353

	Mean	SD	p1	p5	p10	p25	p50	p75	p90	p95	p99	N
Bpifrance: MOIC, deal-level	1.28	2.48	0.00	0.00	0.00	0.00	0.34	1.62	3.43	5.02	9.80	357
Pitchbook: TVPI (fund-level)	1.55	0.69	0.29	0.81	0.97	1.16	1.45	1.74	2.16	2.72	4.51	243
SINE: Imputed MOIC	0.77	2.02	0.00	0.00	0.00	0.00	0.03	0.55	2.08	4.26	8.18	120
SINE: Imputed MOIC (survivors only)	1.36	2.53	0.00	0.01	0.02	0.11	0.28	1.69	4.18	5.41	17.07	68

Table B.2: Comparison of French Deal Multiples Across Data Sources. This table presents the returns distribution for French deals in the Bpifrance and Pitchbook data, for comparison with the imputed returns derived from the SINE survey. The Bpifrance data are constructed from GP reports detailing deal-level returns (MOIC) in France for investments made between 2009 and 2014. The Pitchbook data are restricted to fund-level returns (TVPI) for funds located in France. We impute deal-level returns (MOIC) as in Davenport (2022) on French deals using SINE data (see Section 5.1). We report the distribution of returns for all VC-backed companies in our test set (the 2010 cohort) and for survivors only.

Appendix C Institutional Context: Venture Capital in France

The French VC market differs from the stereotyped VC market, such as that of Silicon Valley, in a number of ways. In order to gain insights into the French VC context, this Appendix compares the VC market in France and in the US and outlines differences that can shape the investment strategies and outcomes for VC firms in both countries. The main data source to contrast VC in the two countries is Pitchbook. We emphasize that one caveat is that unlike our SINE dataset, Pitchbook data is not representative. In addition, Pitchbook's coverage, particularly for France prior to the mid-2010s, is limited.

Market size and growth. The VC market in the US is significantly larger and more mature compared to the French market, with higher volumes of investments and a denser network of startups, VC firms and investors. Below, we present country-level VC investment statistics for France and the US, sourced from the OECD.[47] In 2010, total VC investments in the US were approximately forty times larger than in France. Moreover, while VC investments in France constituted around 0.03% of GDP, the corresponding figure for the US was 0.20% of GDP. Seed stage investments accounted for 5.4% of all VC investment amounts in France, while they accounted for only 1.5% in the US.

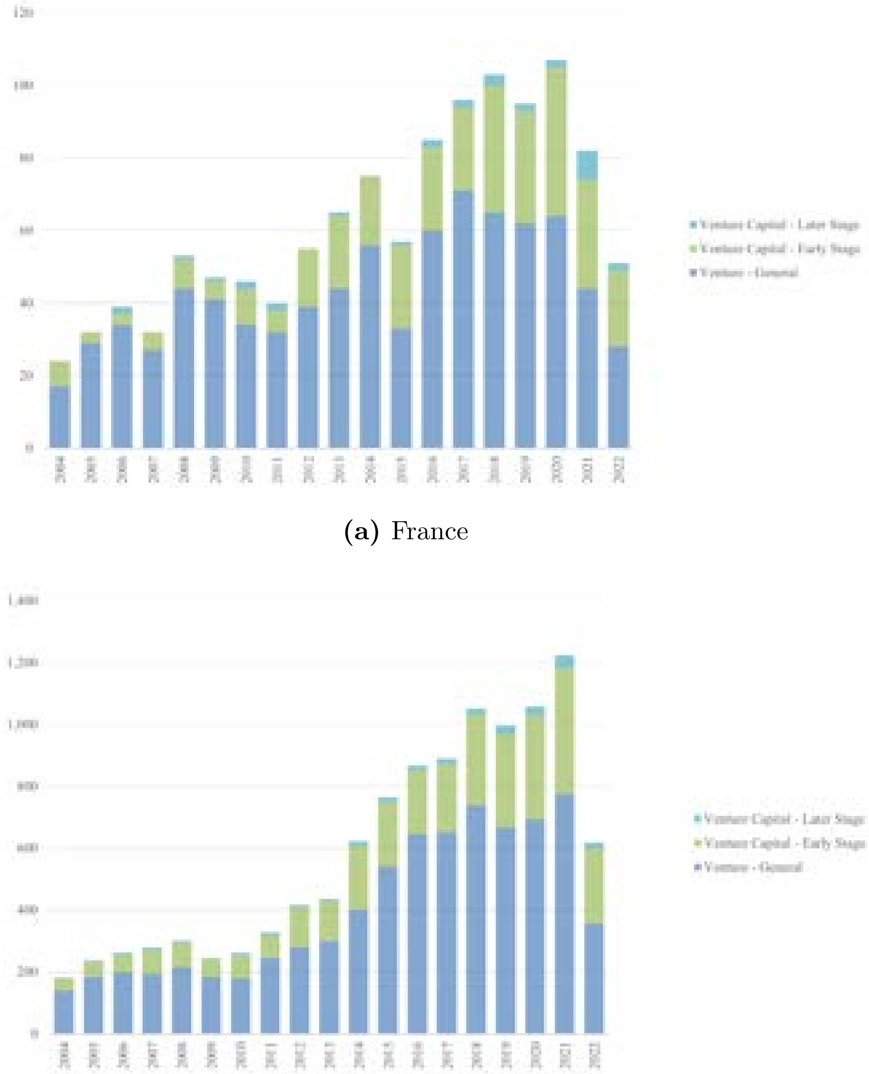
Starting in the early 2010s, the VC landscape in France has experienced a significant transformation. The funding raised by startups in France through VC tripled between 2013 and 2018, reaching 2.8b euros in 2018, and 4.9b euros in 2020.

As of 2010 (in million USD)	**France**	**US**
Total		
% GDP	.03%	.2%
Dollar amounts	736	30,481
Investment Stage		
Seed	40	463
Early stage	113	10,889
Late stage	583	19,129

Table C.1: Venture Capital Investment Activity. This table shows total VC investment amounts in 2010 in France and in the US, their size relative to the country's GDP, as well as investment amounts broken down by investment stage for both countries. Data source: OECD

[47] These statistics are as of 2010, to match the year of creation of companies in our test set.

Figure C.1 uses Pitchbook data to report fund counts, broken down by fund type, over the past two decades. Panel A displays the number of funds invested in France, while Panel B shows funds invested in the US for comparison. Despite approximately ten times more funds invested in the US (part of this gap is also due to the more limited coverage of French funds in Pitchbook), we note that both countries exhibit similar trends in terms of fund count growth, with an increased focus on early-stage VC over time.

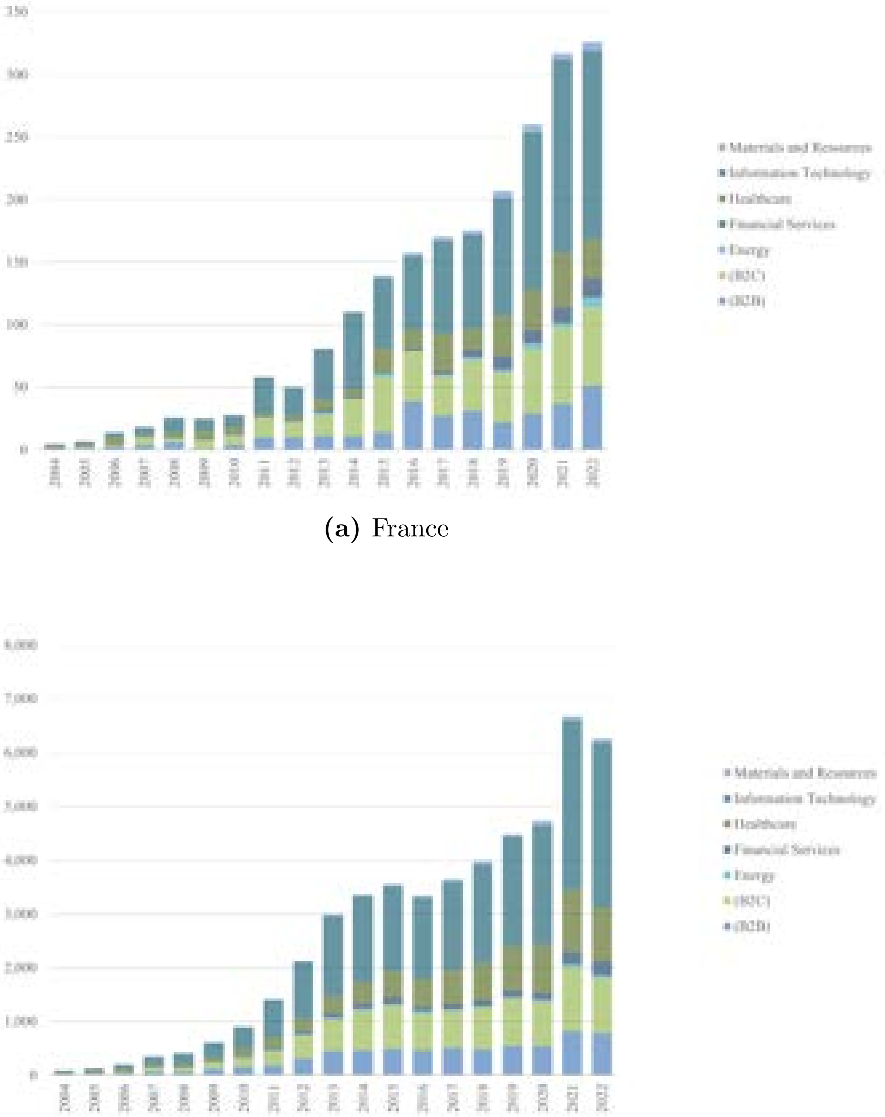


(a) France

(b) United States

Figure C.1: Venture Capital Fund Counts. This figure reports the number of VC funds invested in French companies (Panel A) and in US companies (Panel B) from 2004 to 2022, as well as their sector breakdown. Data source: Pitchbook.

Figure C.2 provides seed deal counts, categorized by industry, for both countries. Although there is again a considerable scale difference, the growth in seed funding in France mirrors that in the US, with the industry breakdown also showing substantial similarities.



(a) France

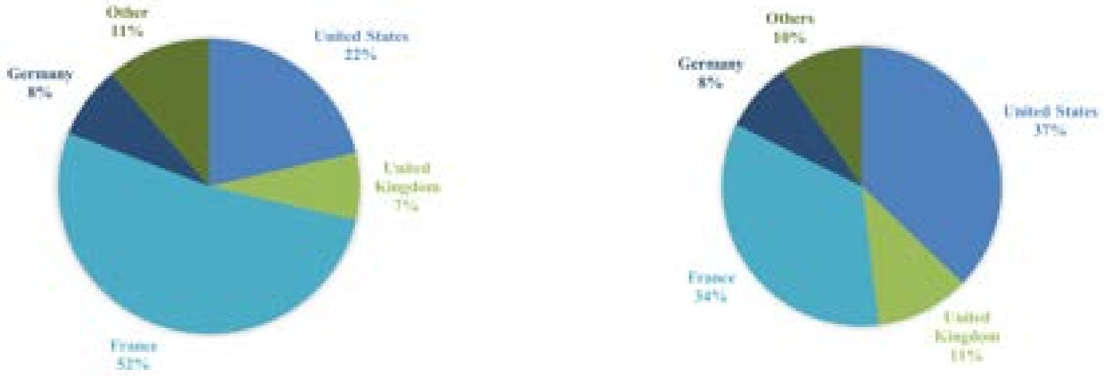
(b) United States

Figure C.2: Seed Deals Counts This figure reports the number of seed deals in France (Panel A) and the US (Panel B) from 2004 to 2022. Data source: Pitchbook.

VC Firms. The relatively localized nature of VC investments in France sets it apart from other European markets (Ekeland, Landier and Tirole, 2016). It is worth noting however that this trend is reverting, with increased contributions from international investors such as SoftBank and Sequoia Capital in the past few years.[48] Figure C.3 shows the country breakdown using fund counts (Panel

[48] https://vc-mapping.arkkapital.com/venture-capital-firms/france

A) and total amounts raised (Panel B) of VC funds that have made seed or early stage investments in French companies between 1998 and 2010. During our sample period, data from Pitchbook show that 52% of funds (counts) in French companies originate from France, while almost one quarter originate from the US. When looking at total amounts raised by French companies over this period, US VC funds have contributed slightly more than French funds, at 37% and 34%, respectively.



Panel A: VC Funds' Country of Origin, by Fund Counts

Panel B: VC Funds' Country of Origin, by Total Amounts Raised

Figure C.3: Country of Origin of VC Investors. This figure shows the countries of origin of the VC funds that have made seed or early-stage investments in French companies between 1998 and 2010. Panel A provides the country breakdown using fund counts, while Panel B reports the breakdown using dollar amounts. Data source: Pitchbook.

One notable difference with the US VC ecosystem is the absence of pension funds and university endowments as significant players in the French VC industry, which limits the available pool of funds for French VC investments. In addition, French universities' contributions to the innovation ecosystem are more limited than their US counterparts. Table C.2 reports the top ten partners in France and the US who invested at the seed or early stage in a French company between 1998 and 2010.

As a result of the limited pool of funds from pension funds and endowments, Bpifrance, a public investment bank created in 2012 through the merger of Oséo, CDC-Entreprises, the FSI, and FSI-Régions has been an influential player (see Table C.2). Owned 50% by the government and 50% by the French deposit and consignment office, Bpifrance has played a significant role in funding innovative ventures and supporting startup growth, shaping the overall VC landscape in France over the past decade. We are able to identify the investors for around 10% of the VC-backed companies in our 2010 SINE cohort of entrepreneurs using Pitchbook. We find that Bpifrance is one of the most frequently listed investors in our sample companies. In Table C.3, we report investment activities statistics for Bpifrance and the other three investors that appear most frequently.

In addition, we use data from Bpifrance and report deal-level returns (MOIC) for 357 deals

	Limited Partner Type	**Affiliated Funds**	**AUM** million USD	**Commitments in VC Funds**
	Panel A: French LP Investors			
Bpifrance	Sovereign wealth fund	61	35,251	47
ODDO BHF Group	Wealth management firm	11	160,555	11
Accexx Capital Partners	Fund of funds	11	14,948	9
Ardian	Fund of funds	8	150,000	7
Quartilium	Fund of funds	7		6
Caisse d'Epargne	Banking institution	4	2,490	3
Caisse des Depots Group	Sovereign wealth fund	3	716,783	3
CDC Enterprises	Fund of funds	5		3
CEA Investissement	Direct investment	3	78	3
Credit Agricole	Banking institution	3	2,748,265	3
	Panel B: US LP Investors			
Adams Street Partners	Fund of Funds	10	54,000	26
HarbourVest Partners	Fund of Funds	17	134,537	19
Calpers	Public pension fund	14	468,300	14
IBM Pers. Pension Plan	Corporate pension	14	52,130	9
Grovestreet	Fund of Funds	8	7,135	8
PA State Emp. Retir. Sys.	Public pension fund	10	34,700	8
HP Inc. Master Trust	Corporate pension	6	7,509	6
MN Life Insurance Co.	Insurance company	7	26,037	6
SBC Master Pension Trust	Corporate pension	7	58,120	6
IL Municipal Retir. Fund	Public pension fund	5	49,187	5

Table C.2: Limited Partners. This table reports the top ten French (Panel A) and US (Panel B) limited partners who made seed or early stage VC investments in French companies between 1998 and 2010. LPs are ordered by committed capital in French companies during the sample period. Data source: Pitchbook.

	Kima Ventures	**BPI France**	**Starquest Capital**	**Alven Capital Partners**
Location	Paris	Paris	Paris	Paris
Most backed sectors	IT, B2C	IT, B2B	IT, B2C	IT, B2C
VC investments	1,226	1,735	154	282
M&A exits	156	217	14	51
IPO exits	2	30	1	2
Median round amount	$1.59m	$2.6m	$1.08m	$6.26m
Median valuation	$7.23m	$10.64m	$4.51m	$28.47m
AUM	N/A	$35.2B	$332m	$2.00B
Number partners	4	121	8	12
Number of VC funds	1	26	1	8
Median fund size	N/A	$215m	$66.35m	$114m

Table C.3: French VC Investors. This table reports statistics on the most active VC investors in the sample of VC-backed companies in the 2010 cohort of the SINE survey that we identify in Pitchbook (around 10%). Data source: Pitchbook.

made between 2009 and 2014 in Table B.2. As is typical of VC investments, returns are heavily skewed. The median investment returns less than invested capital with a MOIC of 0.34, while the mean deal-level MOIC is 1.28, driven by the upper tail (top 1%) of the distribution, which generates ten times the invested capital.

French and European Government Initiatives. Over the past decade, the French government and the European Union (EU) have instituted a range of initiatives to bolster the growth of innovative startups, in the same vein as NSF's America Seed Fund.[49]

Some key initiatives and programs include the "Investments for the Future" program, a €57 billion launched by the French government in 2010 to promote economic growth, with a focus on renewable energy and adjacent domains. Sources of financial support include a combination of grants, refundable grants, and direct capital investment.

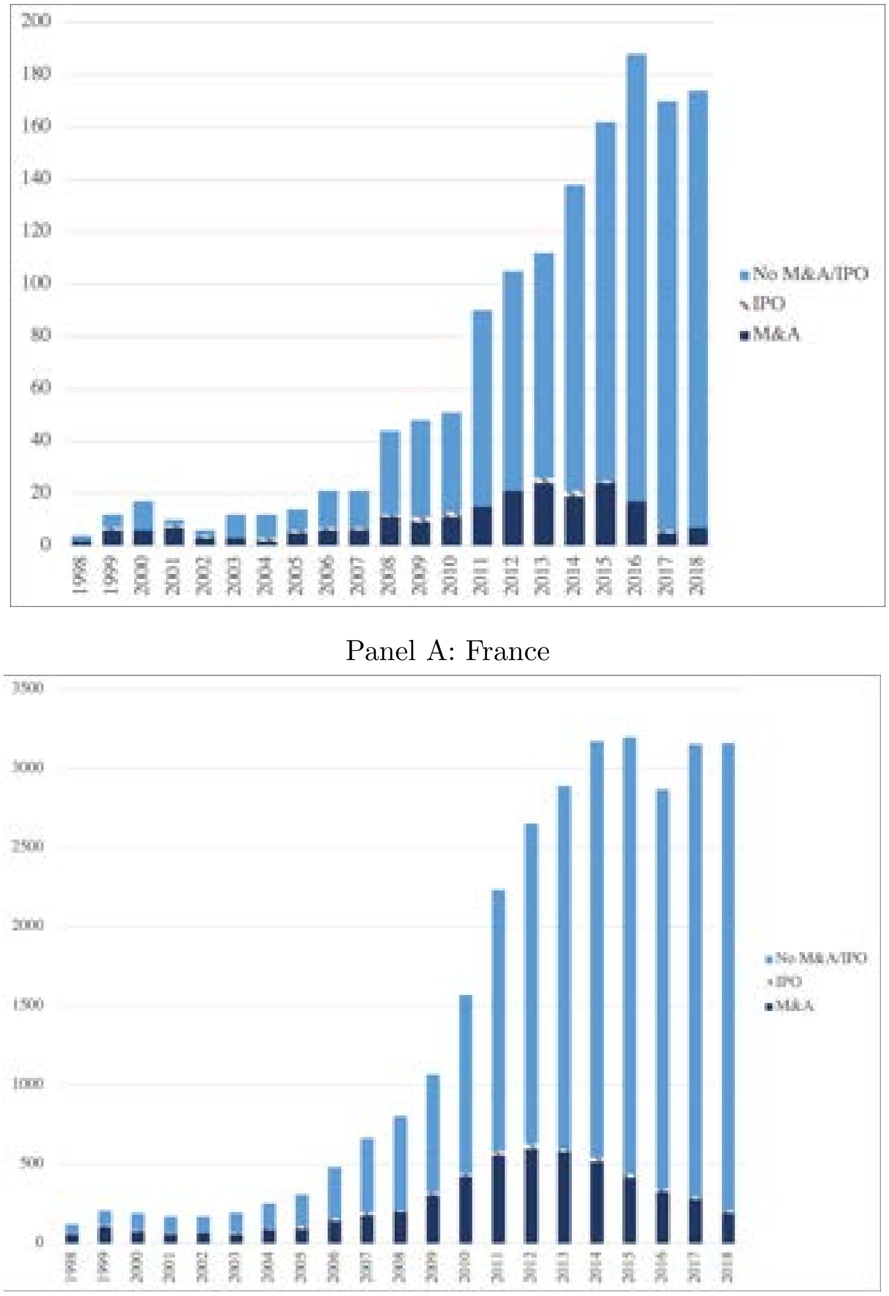
Another French government initiative is The French Tech, a government-driven accelerator launched in 2013 to nurture tech startups across multiple French cities by providing mentorship, accreditation and network services, and funding access. As mentioned above, Bpifrance has played a pivotal role since 2013 through equity investments, loans, guarantees, and advisory services. In addition, two programs launched in 2014. The Competitiveness of Enterprises and SMEs (COSME) program was created by the European Commission (EC) and partners with local financial institutions to provide financing (loans, guarantees as well as VC and equity investments) to entrepreneurs and businesses. Horizon 2020 was a €80 billion research and innovation program also launched by the EU to support research institutes and small and medium-sized businesses (SMEs). The program closed in 2020 and €2.8 billion was budgeted for private finance and venture capital.

VC Investments: Performance and Exits. French VCs are constrained by a limited range of exit strategies, primarily because France lacks a vibrant IPO market for young companies. As a result, VC funds in France (and Europe in general) often choose to exit successful companies through trade sales. This preference is driven by tax considerations, leading to trade sales frequently being structured as share deals (Sebag, Maitrehenry and Loyrette Nouel, 2020).

We use Pitchbook to compare exit rates of French and US companies that have received seed funding. In Figure C.4, Panel A (Panel B) illustrates the proportion of exits (M&A or IPO) for French (US) companies that have received seed investments by company founding year starting in 1998. Note that while these figures provide useful insights for comparison purposes, they are subject to reporting biases and limited coverage in the Pitchbook data, particularly for France in the first half of the sample period.

To better understand how VC investments and exits in France compare to those in the US, we provide statistics for companies founded after 1998 in Table C.4. In Pitchbook, 17% of US firms and 10% of French firms founded after 1998 have reported seed investments. Among companies with seed investments, the "success" rate (across all founding years) is comparable across the two

[49] https://seedfund.nsf.gov/



Panel A: France

Panel B: US

Figure C.4: Exits for Companies with Seed Investments. This figure reports the number and breakdown of exits (M&A or IPO) for companies with seed investments in France (Panel A) and the US (Panel B) categorized by company founding year. Data source: Pitchbook.

countries: 32% of US companies and 31% of French companies either received later stage VC, or went through an IPO or M&A. For these "successful" companies, the median capital invested in the US was $10 million, almost double that of French "successful" companies. Additionally, the median post-valuation of US companies was almost four times that of French companies.

	US	France
Companies founded 1998 - present	228,666	19,489
with seed investment	39,743	1,978
with later stage VC or IPO or M&A	12,590	611
"Success" rate of seed investments	32%	31%
Median capital invested (in million USD)	10	5.51
Median revenue (in million USD)	7	2.1
Median post-valuation (in million USD)	66.55	17.52
with later stage VC	8,396	427
with IPO	330	18
with M&A	5,283	217

Table C.4: Company Counts by Country. This table reports the number of French and US companies that were founded after 1998 in Pitchbook, how many of these have received seed investments. The "success" rate of seed investments in a given country is calculated as the proportion of companies that have received later-stage VC funding, were acquired, or underwent an IPO, out of the total number of companies that received seed investments. Data source: Pitchbook

Table C.5 presents fund statistics in France and the US. In Panel A, we present fund statistics for funds that invested in a French company at the seed or early stage between 1998 and 2010. Funds are categorized by their country of origin, and the statistics are provided for the top four countries with the largest number of funds. The median US VC fund that invested in France is over four times larger than the median French VC fund that invested in a French company. Funds that invest abroad, perhaps not surprisingly, tend to be much larger. In Panel B, where fund statistics are shown for funds that have invested in a US company, the median French fund is larger than the median US fund. For the subset of funds with performance metrics available in Pitchbook, French VCs do not appear to have underperformed relative to their US peers.

Figure C.5 presents the median fund TVPI ratio for funds (all countries of origin included) that have made seed or early-stage VC investments in French and in US companies (shown separately), between 1998 and 2010. The performance of early stage VCs that have made investments in France appears to be at par with those that have made investments in the US.

Deal structures, Incentives, and VC Firms Operations. VC firms in Europe generally operate similarly to their US counterparts, with a few notable distinctions. Compared to American VCs, European VCs tend to adopt a more hands-off approach. Hellmann and Puri (2002) show that VCs in the US actively participate in guiding strategy, hiring decisions, and accelerating growth.

In a survey conducted by Schwienbacher (2008) in 2001, which included 104 VC firms in Europe

VC Fund Country	Fund Count	IRR Count	IRR Mean	IRR Median	Fund Size Mean	Fund Size Median	TVPI Mean	TVPI Median	TVPI Count
				Panel A: VC Investors in French Companies					
US	50	19	18.03%	1.70%	250.40	177.50	1.70x	1.17x	17
UK	27	10	17.32%	15.31%	377.21	194.77	2.02x	1.64x	9
France	244	11	7.96%	9.00%	106.17	39.56	1.59x	1.46x	10
Germany	15	1	-4.20%	-4.20%	174.99	135.18	0.69x	0.69x	1
				Panel B: VC Investors in US Companies					
US	3,244	1,075	10.41%	7.00%	251.37	85.00	1.86x	1.44x	965
UK	137	41	11.09%	9.71%	449.71	147.50	1.71x	1.51x	35
France	66	5	7.24%	10.70%	132.39	104.52	1.57x	1.33x	4
Germany	41	3	-2.50%	-4.20%	125.96	107.06	0.92x	0.75x	3

Table C.5: VC Funds Statistics. This table reports fund size, fund count, and fund performance (as of the most recent reporting quarter with IRR) for VC funds that have made seed or early stage VC investments in French (Panel A) and US (Panel B) companies between 1998 and 2010, by the funds' country of origin. The four countries with the highest number of funds are shown. Data source: Pitchbook.

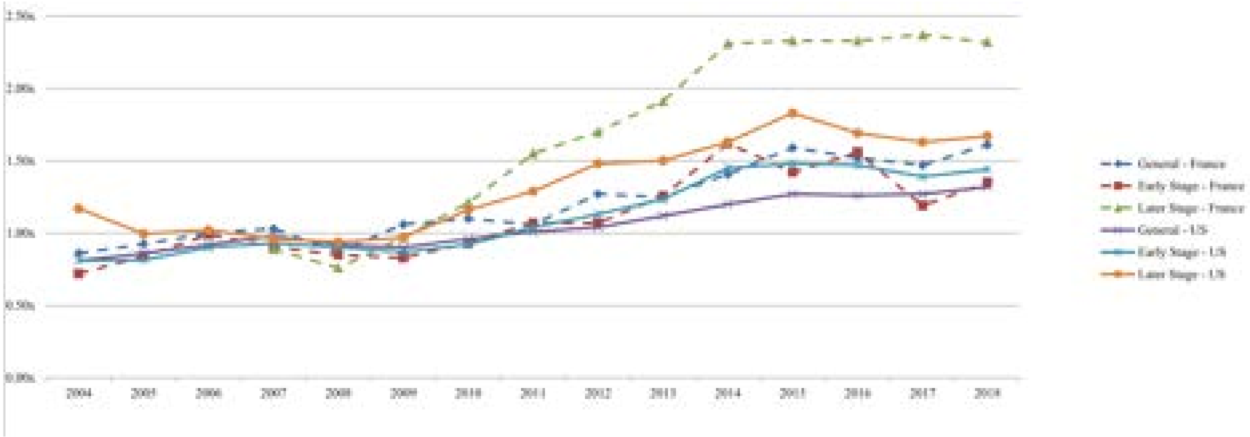


Figure C.5: VC Funds TVPI. This figure reports the median TVPI for funds that were invested between 1998 and 2010 in seed or early-stage VC in France and the US. Data source: Pitchbook.

(including 13 French VCs) and 67 American VC firms, several differences emerged. On average, European VCs held their investments for a longer duration (3.6 years vs. 2.9 years) and utilized convertible securities, convertible debt, and convertible preferred stock less frequently (17% vs. 59%) compared to their American counterparts. Additionally, European VCs replaced management less often (19% vs. 34%) and engaged in co-investments with other VCs less frequently (56% vs. 81%). These variations, especially the hands-off approach adopted by VCs in Europe, may contribute to the relatively lower performance of their portfolio companies.

Profiles of VC-backed Entrepreneurs. Are French VC-backed entrepreneurs similar to US VC-backed entrepreneurs? In Table 4, we report demographics and other statistics for entrepreneurs who were part of the SINE survey and declared having received VC as part of the survey. To compare the typical VC-backed French entrepreneurs to their American counterpart, we leverage data in Pitchbook. Using data on all VC-backed and formerly VC-backed companies founded between 1998 and 2010 in Pitchbook, in France and in the US, we find that 9% of VC-backed French founders are female, compared to 11.5% of founders in the US. Moreover, 26% of French VC-backed founders with an institution reported in Pitchbook have a degree from one of the top 3 French elite schools. This is similar to 24% of US founders coming from an Ivy League school.[50]

[50] The three French schools are: HEC Paris, Ecole Polytechnique and Ecole Centrale de Paris. The US schools are: Harvard, Stanford, Princeton, Dartmouth, University of Pennsylvania, MIT, Cornell, Yale, Columbia and Brown.

Appendix D Model Interpretability

Lundberg and Lee (2017) introduce a model interpretability approach based on Shapley values, which are rooted in coalitional game theory. The input feature values for an observation act as players in a coalition. An input feature's SHAP value for a given observation represents its contribution to shifting the model's output from its unconditional expectation. This value is calculated as the average change in expected model output across all possible orderings of other features. SHAP values can be aggregated across observations to facilitate the model's global interpretability, providing a ranking of features based on their predictive importance. We emphasize that while SHAP values offer insights into the model's prediction process, they do not imply causality and have interpretational limitations, as discussed by (Chen et al., 2020).

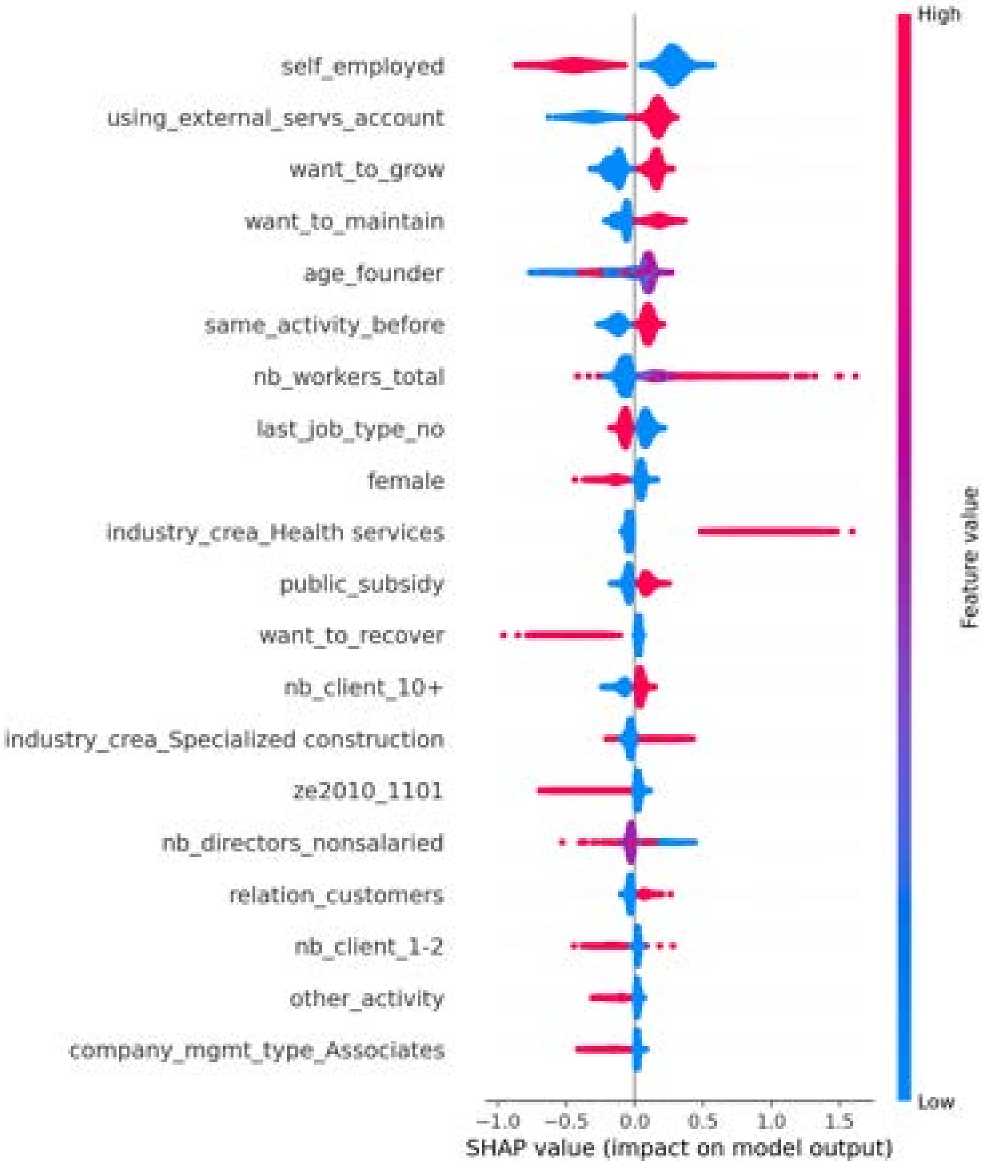


Figure D.1: SHAP Values of Top Predictors for Operating Performance. This figure displays the SHAP values for the twenty most influential features in predicting operating performance, measured as log revenue at age 5. Features are ranked in decreasing order of importance. Each point represents an individual observation, with its position on the x-axis indicating its SHAP value. Positive SHAP values suggest the feature increased the predicted operating performance for that observation, while negative values indicate a decrease. The color of each point reflects the feature's value for that observation. The predictive model is trained on all new companies in the 1998, 2002, and 2006 cohorts using ten-fold cross validation.

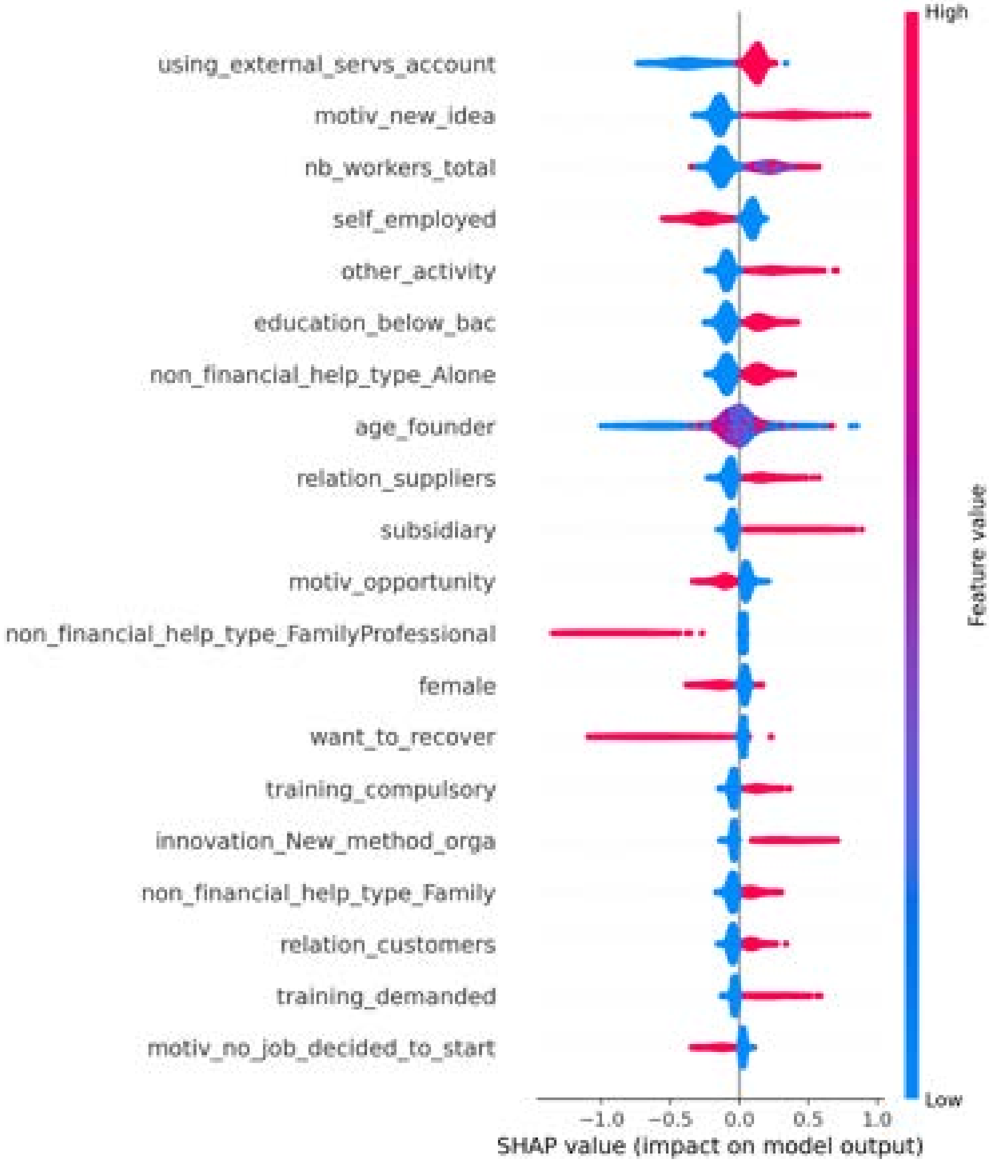


Figure D.2: SHAP Values of Top Predictors for VC backing. This figure displays the SHAP values for the twenty most influential features in predicting whether a firm is VC-backed. Features are ranked in decreasing order of importance. Each point represents an individual observation, with its position on the x-axis indicating its SHAP value. Positive SHAP values suggest the feature increased the predicted operating performance for that observation, while negative values indicate a decrease. The color of each point reflects the feature's value for that observation. The predictive model is trained on a random sample of all new companies in the 1998, 2002, and 2010 cohorts using five-fold cross validation.

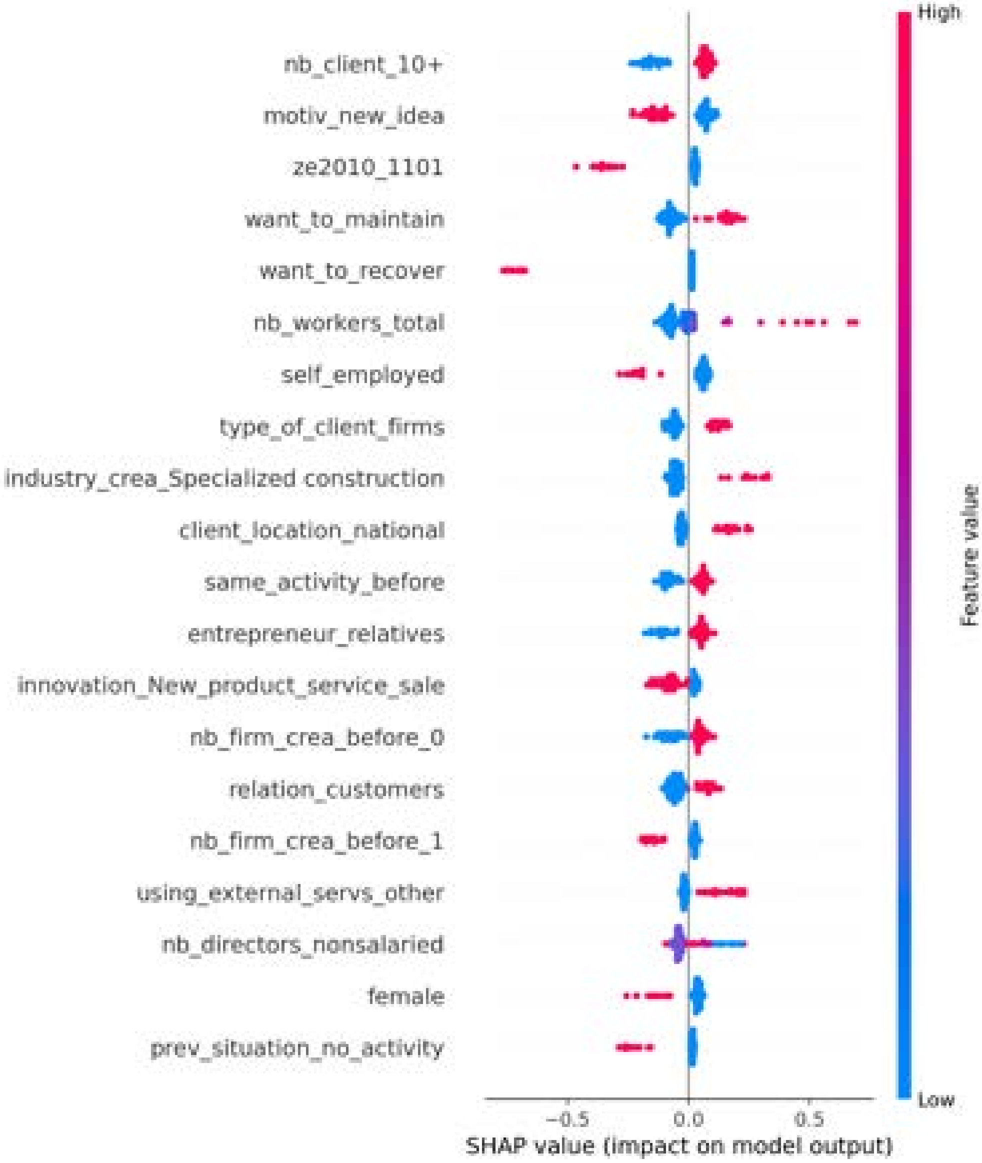


Figure D.3: SHAP Values of Top Predictors for Operating Performance When the Model is Trained on VC-backed companies Only. This figure displays the SHAP values for the twenty most influential features in predicting operating performance (log of revenue at age 5). Features are ranked in decreasing order of importance. Each point represents an individual observation, with its position on the x-axis indicating its SHAP value. Positive SHAP values suggest the feature increased the predicted operating performance for that observation, while negative values indicate a decrease. The color of each point reflects the feature's value for that observation. The predictive model is trained on new VC-backed companies in the 1998 and 2002 cohorts using ten-fold cross validation.

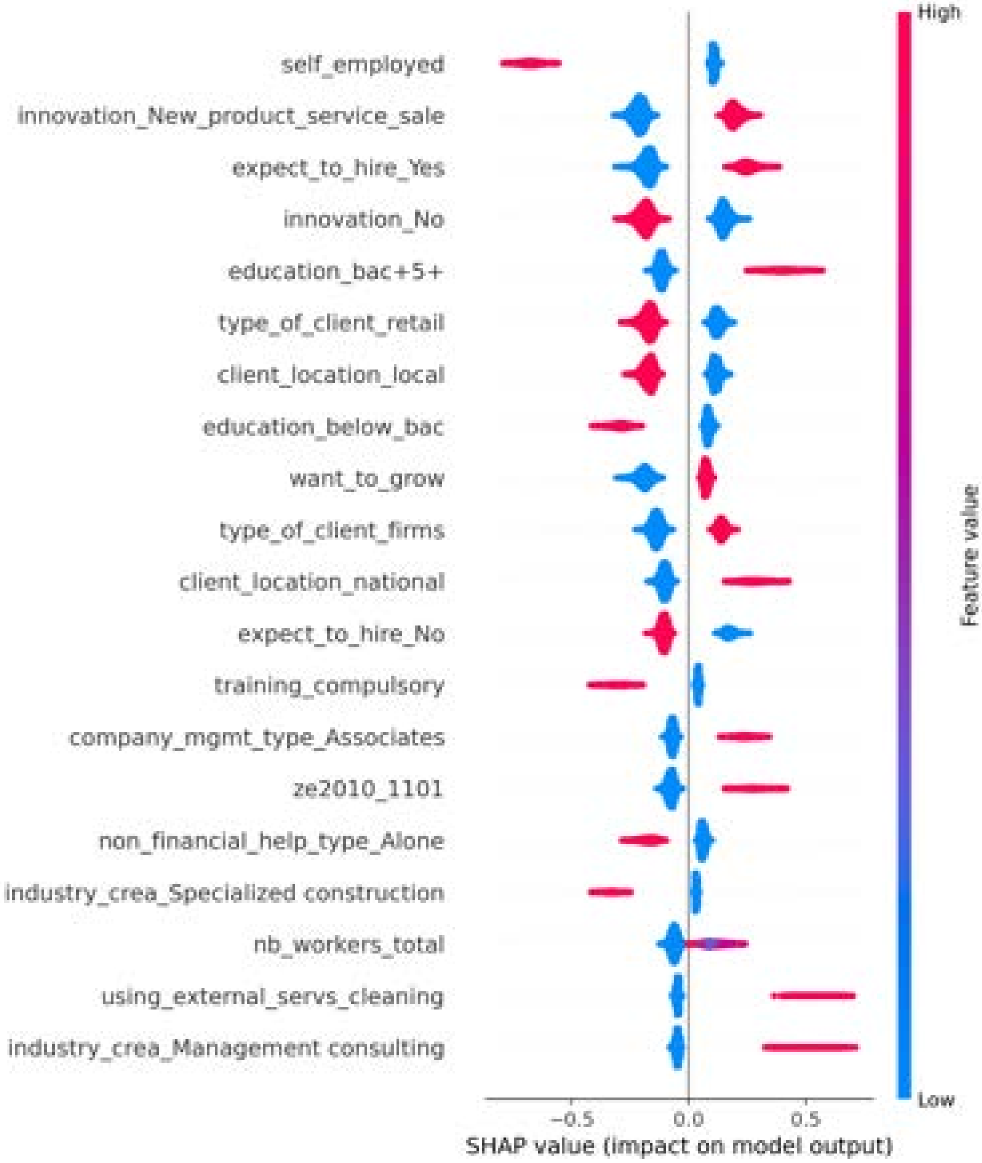


Figure D.4: SHAP Values of Top Predictors for Exits. This figure displays the SHAP values for the twenty most influential features in predicting *Exit*, which takes value one if the firm is acquired or if it becomes public. We use data from Pitchbook, CBInsights, Preqin, SDC, VentureXpert, CapitalIQ, Orbis, and Crunchbase to construct this measure. Features are ranked in decreasing order of importance. Each point represents an individual observation, with its position on the x-axis indicating its SHAP value. Positive SHAP values suggest the feature increased the predicted operating performance for that observation, while negative values indicate a decrease. The color of each point reflects the feature's value for that observation. The predictive model is trained on the sample of all new companies in the 1998, 2002 and 2006 cohorts using ten-fold cross validation.

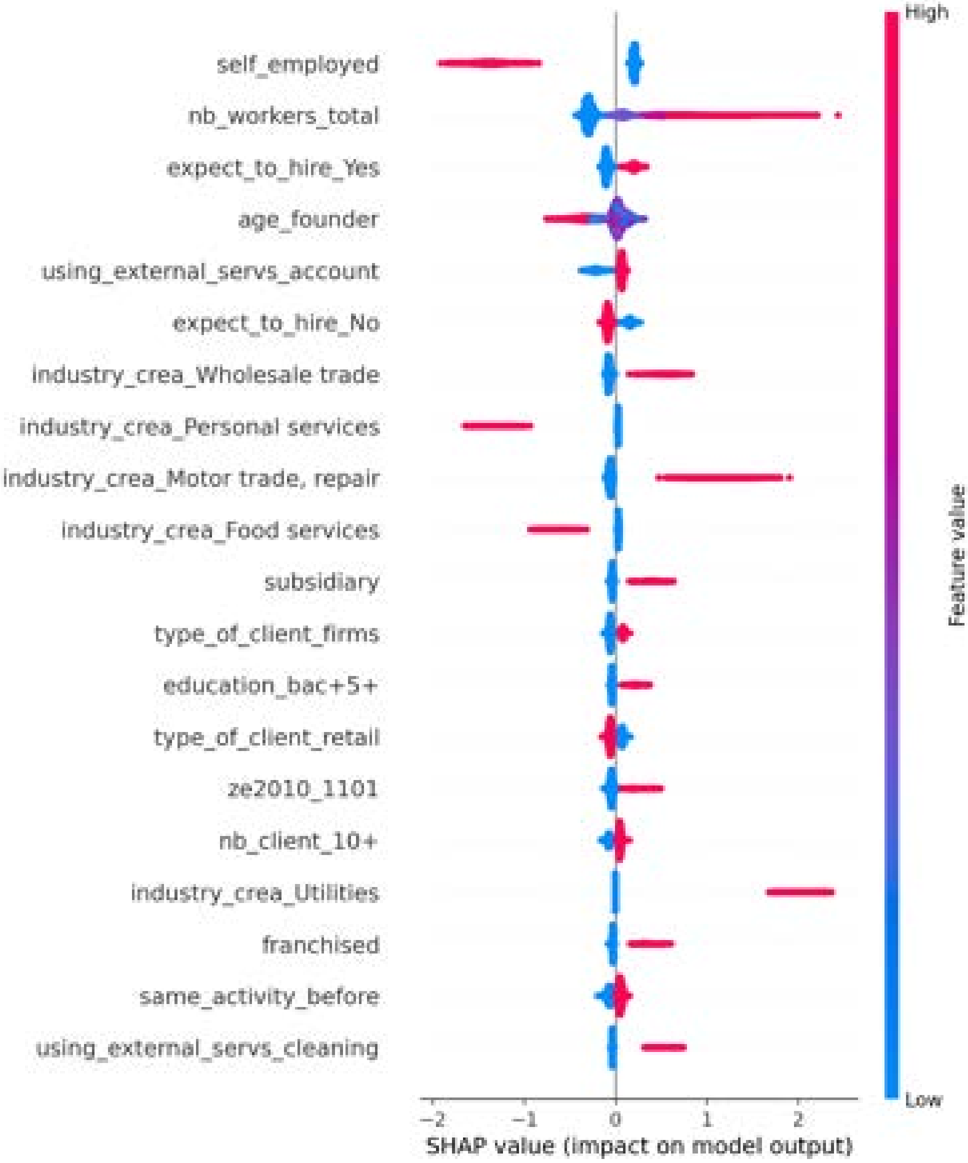


Figure D.5: SHAP Values of Top Predictors for Best Predicted Performers. This figure displays the SHAP values for the twenty most influential features in predicting whether a firm will be in the top 5% of its cohort in terms of operating performance (revenue at age 5). Features are ranked in decreasing order of importance. Each point represents an individual observation, with its position on the x-axis indicating its SHAP value. Positive SHAP values suggest the feature increased the predicted operating performance for that observation, while negative values indicate a decrease. The color of each point reflects the feature's value for that observation. The predictive model is trained on the sample of all new companies in the 1998, 2002 and 2006 cohorts using ten-fold cross validation.

Appendix E Additional Tables and Figures

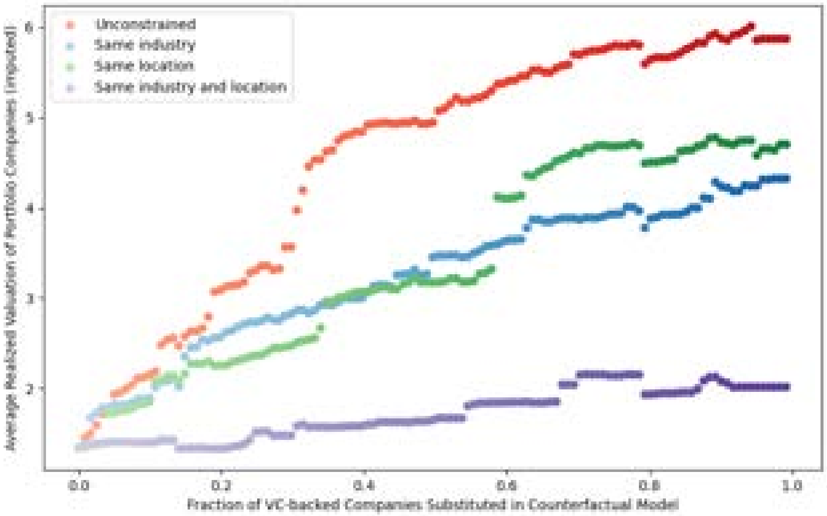


Figure E.1: Counterfactual Models Evaluated with Imputed Valuations. This figure shows the average performance (imputed valuations in millions of euros) of companies selected by several counterfactual models. The counterfactual models sequentially drop VC-backed companies with the lowest $\hat{m}(x_i)$ and replace them with the best predicted performers (i.e., companies in $\mathcal{A}_s$ with the highest $\hat{m}(x_i)$) selected from various investable pools $\mathcal{D}$, ensuring the total number of portfolio companies stays constant at $|\mathcal{V}_s| = 120$. The x-axis shows the fraction of VC-backed companies replaced. The y-axis reports the average performance of the companies in the portfolio (in millions of euros). The red line shows the performance of the unconstrained counterfactual model, that is, the best predicted performers are not constrained within a specific set of companies. Other lines represent the performance of a counterfactual model constrained to replace each VC-backed company it excludes with a company that is in the same industry (in blue), the same location (in green), or both the same industry *and* location (in purple). We calculate industry-level median exit valuation multiples for early deals in Pitchbook data starting in 2000. Imputed valuations (in million of euros) are constructed by multiplying the companies' observed revenue at age 5 by their respective industry's median exit valuation multiple.

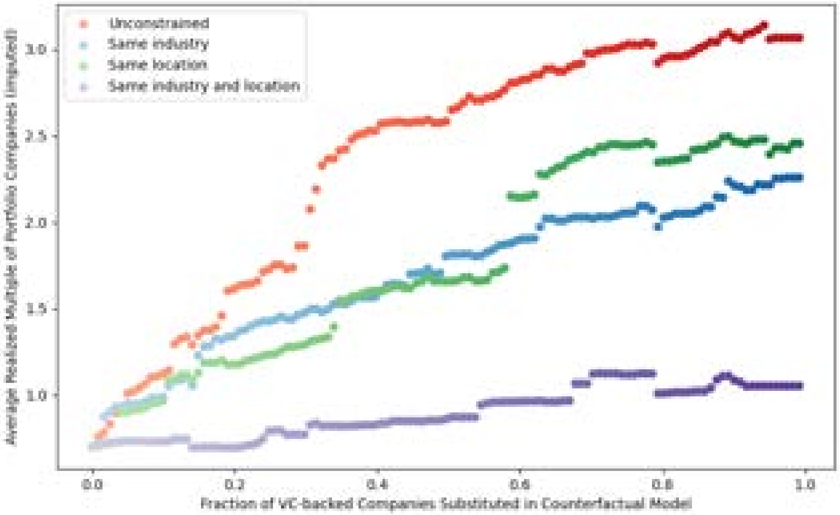


Figure E.2: Counterfactual Models Evaluated with Imputed Investment Multiples. This figure shows the average performance (imputed investment multiple) of companies selected by several counterfactual models. The counterfactual models sequentially drop VC-backed companies with the lowest $\hat{m}(x_i)$ and replace them with the best predicted performers (i.e., companies in $\mathcal{A}_s$ with the highest $\hat{m}(x_i)$) selected from various investable pools $\mathcal{D}$, ensuring the total number of portfolio companies stays constant at $|\mathcal{V}_s| = 120$. The x-axis shows the fraction of VC-backed companies replaced. The y-axis reports the average performance of the companies in the portfolio (imputed MOIC). The red line shows the performance of the unconstrained counterfactual model, that is, the best predicted performers are not constrained within a specific set of companies. Other lines represent the performance of a counterfactual model constrained to replace each VC-backed company it excludes with a company that is in the same industry (in blue), the same location (in green), or both the same industry *and* location (in purple). Imputed valuations (in million of euros) are constructed by multiplying the companies' observed revenue at age 5 by their respective industry's median exit valuation multiple. Imputed multiples are constructed by multiplying imputed valuations by the industry's median deal terms for early stage deals accounting for dilution (see Equation (7)).

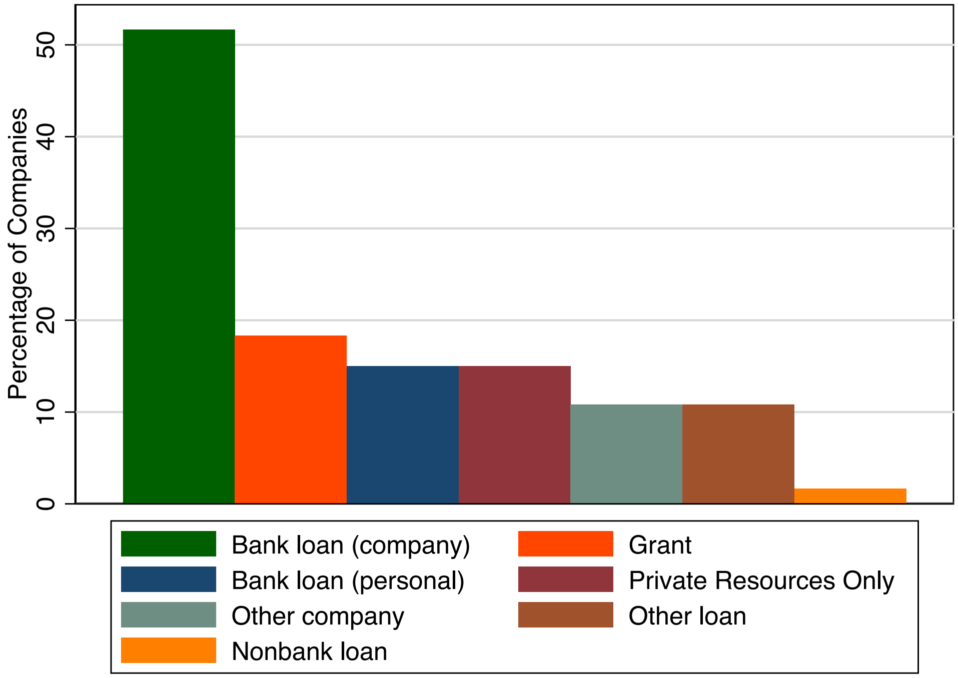


Figure E.3: Best Predicted Performers: Sources of Outside Funding at Firm Creation. This figure reports the distribution of responses to the survey question that asks founders about sources of outside funding at firm creation. Founders in this set include the best predicted performers in $\mathcal{A}_{s=120}$ when the investable pool $\mathcal{D}$ is restricted to founders who match VC-backed firms on financial constraints, industry and growth prospects (yellow line in Figure 6).

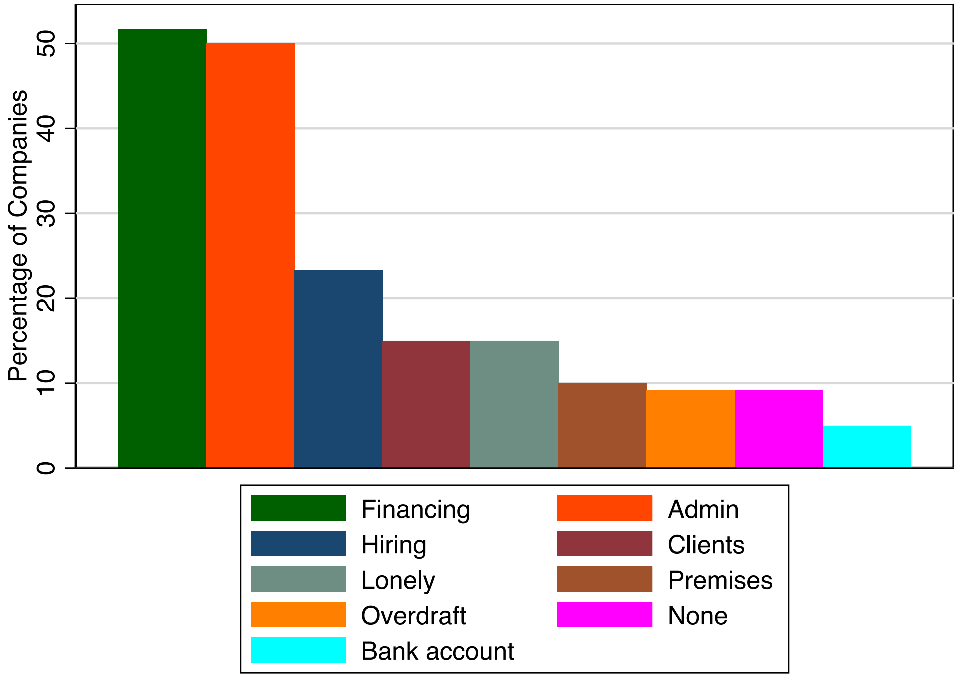


Figure E.4: Best Predicted Performers: Main Obstacles at Firm Creation. This figure reports the distribution of responses to the survey question that asks founders about the main obstacles they faced at firm creation. Founders in this set include the best predicted performers in $\mathcal{A}_{s=120}$ when the investable pool $\mathcal{D}$ is restricted to founders who match VC-backed firms on financial constraints, industry and growth prospects (yellow line in Figure 6).

	Percentile rank in exit valuation distribution (averaged across sectors)	
	Average venture	Median venture
Firms in top 1% of revenue	90	96
Firms in top 5% of revenue	88	95
Firms in top 10% of revenue	84	91

Table E.1: Correspondence Between Revenue and Valuation at Exit. This table reports the correspondence between VC-backed companies' revenue and valuation, both at exit. Rows 1, 2, and 3 focus on companies in the top 1%, 5%, and 10% of the revenue distribution, respectively. For each set of companies, Column 1 shows the percentile rank of the average firm in the distribution of exit valuations, and Column 2 shows the percentile rank of the median firm in the distribution of exit valuations. All percentile ranks are calculated at the sector level and then averaged across sectors. The data come from Pitchbook and comprise French and US companies with recorded exits for which post valuation and revenue are available.

		Percentiles in Deal Terms Distribution
Quintiles of Revenue Multiple for Best Predicted Performers	Q1	34;71
	Q2	14;87
	Q3	4;97
	Q4	1;99
	Q5	1;99
	Median	8;93

Table E.2: Sensitivity Analysis: Varying Revenue Multiple Assumptions for Best Predicted Performers This table reports the results of a sensitivity analysis for the results presented in Figure 4 under different revenue multiple assumptions for the portfolio of best predicted performers. VC-backed companies are assumed to secure their respective industry's median revenue multiple, estimated using US Pitchbook data. The last column represents two percentiles from the distribution of deal terms in early-stage French deals, following the interpretation of the results in Figure 4. The first number indicates how unfavorable deal terms would have to be for hypothetical investors backing the best-predicted performers for $MOIC_\alpha - MOIC_h$ to turn negative. The second number captures how favorable the deal terms would have to be for investors in the VC-backed companies for the MOIC difference to turn negative, assuming hypothetical investors in the best predicted performers secured median deal terms. The distribution of deal terms reflects the distribution of early-stage French deal terms. Data source: Pitchbook

Investable Pool ($\mathcal{D}$)	Revenue at Age 5 (log)	
	Mean	S.D.
Unconstrained	6.05	2.27
Location	5.64	2.3
Industry	5.25	2.88
Growth, innovation and hiring	5.38	2.83
Financially constrained	5.13	2.69
Location and industry	3.9	2.88
Loc., ind., and fin. constrained	3.35	2.96
Growth, fin. cons. and industry	3.91	3
Same revenue at birth	4.89	2.73
Comparison:	Revenue at Age 5 (log)	
	Mean	S.D.
All firms in test set	2.43	2.48
VC-backed firms	2.82	2.81

Table E.3: Performance of the Set of Best Predicted Performers When Varying the Investable Pool $\mathcal{D}$. To quantify the performance of the best predicted performers and the importance of supply and demand factors, we create counterfactual models that sequentially drop VC-backed companies with the lowest $\hat{m}(x_i)$ and replace them with best predicted performers (i.e., companies in $\mathcal{A}_s$ with the highest $\hat{m}(x_i)$) such that the total number of portfolio companies stays constant at $|\mathcal{V}_s| = 120$. Realized performance is measured in terms of revenue at age 5 in thousands of euros. Companies that fail by age 5 are included and assigned zero revenue. The first row shows the realized performance of the best predicted performers in the entire set of new companies (i.e., without any constraints on the pool $\mathcal{D}$ from which the algorithm selects). The next rows show the realized performance of the best predicted performers in pools $\mathcal{D}$ restricted to simulate VCs' constraints or preferences. To build these counterfactuals, our algorithm ranks companies in the test set by predicted performance using the function $M(x_i)$, as before. However, unlike the previous approach where companies in $\mathcal{A}_s$ were selected based on $M(x_i) > 1 - s$, we now require that the best predicted performers match VC-backed companies on one or more criteria. The last two rows of the table repeat the performance of all companies in the 2010 cohort and VC-backed companies only in that cohort.

		Test Set (2010 cohort)										
		VC-backed			$\mathcal{A}_{s=0.5\%}$			$\mathcal{A}_{s=1\%}$			Diff. VC-$\mathcal{A}_{s=0.5\%}$	Diff. VC-$\mathcal{A}_{s=1\%}$
		Mean	SD	N	Mean	SD	N	Mean	SD	N	T-Test	T-Test
Predicted Performance												
	Pred. Revenue at Age 5 (log), k euros	2.87	1.07	120	5.61	0.45	187	5.29	0.46	374	-2.75***	-2.42***
Outcomes												
	Revenue at Age 5 (log), k euros	2.82	2.81	120	5.80	2.37	187	5.52	2.40	374	-2.98***	-2.70***
	Revenue at Age 5, k euros	283.21	686.47	120	1247.77	2238.84	187	925.61	1681.64	374	-964.56***	-642.40***
	Alive at Age 5	0.69	0.46	120	0.89	0.31	187	0.91	0.29	374	-0.20***	-0.21***
Founder Demographics												
	Entrepreneur's Age	41.26	10.58	120	42.53	9.30	187	42.14	9.22	374	-1.27	-0.89
	Founder's Nationality (FR)	0.94	0.24	120	0.98	0.13	187	0.98	0.13	374	-0.04**	-0.04**
	Female	0.09	0.29	120	0.15	0.36	187	0.14	0.35	374	-0.06	-0.05
Founder Professional Background												
	Same Prior Industry	0.52	0.50	120	0.92	0.27	187	0.92	0.27	374	-0.40***	-0.41***
	Serial Entrepreneur	0.10	0.30	120	0.02	0.15	187	0.06	0.24	374	0.08***	0.04
	Previously Employed in Small Firm	0.54	0.50	120	0.44	0.50	187	0.44	0.50	374	0.10*	0.10**
	Graduate Degree	0.37	0.48	120	0.49	0.50	187	0.40	0.49	374	-0.12**	-0.04
	Elite School	0.27	0.44	120	0.10	0.30	187	0.09	0.29	374	0.17***	0.17***
Founder Motivation and Expectations												
	Expectation: Growth	0.57	0.50	120	0.59	0.49	187	0.60	0.49	374	-0.01	-0.02
	Motivation: Successful Peer Entrepreneurs	0.06	0.24	120	0.07	0.26	187	0.08	0.27	374	-0.01	-0.02
	Expect to Hire	0.51	0.50	120	0.60	0.49	187	0.58	0.49	374	-0.09	-0.07
	Motivation: New Idea	0.39	0.49	120	0.07	0.26	187	0.08	0.27	374	0.32***	0.31***
	Motivation: Opportunity	0.38	0.49	120	0.60	0.49	187	0.54	0.50	374	-0.22***	-0.17***
	Innovation	0.73	0.44	120	0.43	0.50	187	0.43	0.50	374	0.31***	0.30***
Venture Characteristics												
	Paris-based	0.21	0.41	120	0.07	0.26	187	0.07	0.25	374	0.13***	0.14***
	High-Tech Industry	0.13	0.34	120	0.02	0.13	187	0.02	0.14	374	0.12***	0.11***
Organization												
	Outsourcing: Accounting	0.90	0.30	114	0.83	0.38	187	0.87	0.33	374	0.07*	0.03
	Outsourcing: Management	0.10	0.30	114	0.26	0.44	187	0.21	0.41	374	-0.17***	-0.11***
	Outsourcing: Logistics	0.16	0.37	114	0.34	0.47	187	0.29	0.46	374	-0.18***	-0.14***
	Number of Employees	2.37	2.87	114	5.96	4.32	187	5.30	4.15	374	-3.59***	-2.93***
Industries-Locations												
	Number of Industries	.	.	37	.	.	29	.	.	36		
	Number of Regions	.	.	68	.	.	101	.	.	143		

Table E.4: Differences Between VC-backed and Best Predicted Performers When Varying the Selectivity Threshold s. We verify that the results in Table 4, where the number of companies in $\mathcal{A}_s$ matches the number of VC-backed companies in the test set (120 companies), do not depend on the number of best predicted performers chosen by the model. We report the same statistics as Table 4 for VC-backed companies and two sets of best predicted performers corresponding to two selectivity thresholds, $s = 0.5\%$ and $s = 1\%$ (187 and 374 best predicted performers, respectively). We report t-tests for the difference in means. We assign zero as the (log) revenue at age 5 for companies that do not survive. The data come from the entrepreneur survey (SINE) conducted by the French Statistical Office, tax files from the Ministry of Finance, and the firm registry (SIRENE). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Feature	Top 5%	Bottom 95%	Representativeness of best performers $\frac{Pr(X_i \mid \text{Top5})}{Pr(X_i \mid \text{Bottom95})}$	Top 1%	Bottom 99%	Representativeness of best performers $\frac{Pr(X_i \mid \text{Top5})}{Pr(X_i \mid \text{Bottom95})}$	Fraction among VC-backed companies $Pr(X_i \mid \text{VC-backed})$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VC Hub	65.72	61.65	1.07	78.83	61.68	1.28	61.88
California	44.13	38.57	1.14	60.22	38.64	1.56	38.85
Massachusetts	8.46	8.98	.94	5.84	8.98	.65	8.96
New York	8.02	7.28	1.1	8.03	7.31	1.1	7.34
Texas	5.11	6.82	.75	4.74	6.76	.7	6.73
Most VC-backed Industries	77.61	75.11	1.03	79.2	75.2	1.05	75.21
Information Technology	48.8	44.14	1.11	61.31	44.21	1.39	44.36
Health Care	19.77	21.72	.91	9.12	21.75	.42	21.6
Consumer Discretionary	9.04	9.25	.98	8.76	9.25	.95	9.25
Industrials	6.13	7.84	.78	2.55	7.81	.33	7.75
Communication	7.22	6.63	1.09	9.85	6.63	1.49	6.65

Table E.5: Success Representativeness in U.S. MSCI-Burgiss Data (Using TVPI). This table reports the fraction of entrepreneurs with a given characteristic among the best performing companies and among the other companies. The two deal characteristics available in the Burgiss data are the company location and industry. The sample is restricted to U.S. realized deals with available industry, location, and TVPI. We focus on the four largest U.S. states, and the four largest industries, in terms of deals number. "VC Hub" and "Largest Industries" are defined as the four largest U.S. states and industries, respectively. We use TVPI as a measure of performance. In columns 1 and 2, the best performing companies are in the top 5%, and the other companies in the bottom 95%, in terms of TVPI. In columns 4 and 5, the best performing companies are in the top 1%, and the other companies in the bottom 99%, in terms of TVPI. A given characteristic is representative (or stereotypical) of the best performing companies if it scores high on the representativeness ratio (columns 3 and 6) of the percentage in columns 1 or 4 over that in column 2 or 5.

Feature	Top 5%	Bottom 95%	Representativeness of best performers $\frac{Pr(X_i \mid \text{Top5})}{Pr(X_i \mid \text{Bottom95})}$	Top 1%	Bottom 99%	Representativeness of best performers $\frac{Pr(X_i \mid \text{Top5})}{Pr(X_i \mid \text{Bottom95})}$	Fraction among VC-backed companies $Pr(X_i \mid \text{VC-backed})$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VC Hub	64.64	61.67	1.05	67	61.76	1.08	61.88
California	40.77	38.38	1.06	44.83	38.42	1.17	38.85
Massachusetts	10.12	8.74	1.16	11.82	8.79	1.35	8.96
New York	7.66	7.58	1.01	6.4	7.59	.84	7.34
Texas	6.09	6.98	.87	3.94	6.96	.57	6.73
Most VC-backed Industries	76.92	74.85	1.03	82.76	74.88	1.11	75.21
Information Technology	44.01	45.3	.97	43.35	45.24	.96	44.36
Health Care	26.13	20.32	1.29	34.48	20.48	1.68	21.6
Consumer Discretionary	6.78	9.23	.73	4.93	9.16	.54	9.25
Industrials	6.58	8.21	.8	4.93	8.16	.6	7.75
Communication	5.89	6.3	.94	5.91	6.28	.94	6.65

Table E.6: Success Representativeness in U.S. MSCI-Burgiss Data (Using IRR). This table reports the fraction of entrepreneurs with a given characteristic among the best performing companies and among the other companies. The two deal characteristics available in the Burgiss data are the company location and industry. The sample is restricted to U.S. realized deals with available industry, location, and IRR. We focus on the four largest U.S. states, and the four largest industries, in terms of deals number. "VC Hub" and "Largest Industries" are defined as the four largest U.S. states and industries, respectively. We use IRR as a measure of performance. In columns 1 and 2, the best performing companies are in the top 5%, and the other companies in the bottom 95%, in terms of IRR. In columns 4 and 5, the best performing companies are in the top 1%, and the other companies in the bottom 99%, in terms of IRR. A given characteristic is representative (or stereotypical) of the best performing companies if it scores high on the representativeness ratio (columns 3 and 6) of the percentage in columns 1 or 4 over that in column 2 or 5.

Table E.7: Summary Statistics: Entrepreneur and Venture Characteristics in the 2014 and 2018 cohorts. This table reports summary statistics for a subset of features in the 2014 (Panel A) and 2018 (Panel B) cohorts of entrepreneurs. The data come from the entrepreneur survey (SINE) conducted by the French Statistical Office, tax files from the Ministry of Finance, and the firm registry (SIRENE). Outcome measures are not available for these cohorts. Appendix E describes the variables in the entrepreneur survey.

		2014 cohort						2018 cohort					
	Variable	Mean	SD	p50	p90	p99	N	Mean	SD	p50	p90	p99	N
Demographics													
	Entrepreneur's Age	40.45	10.49	37.50	57.50	60.50	32,023	40.86	10.29	37.50	57.50	60.50	16,675
	Female	0.28	0.45	0.00	1.00	1.00	32,023	0.28	0.45	0.00	1.00	1.00	16,675
	Entrepreneur's Nationality (FR)	0.91	0.28	1.00	1.00	1.00	32,023	0.92	0.28	1.00	1.00	1.00	16,675
	Entrepreneurial Family	0.69	0.46	1.00	1.00	1.00	32,023	0.73	0.44	1.00	1.00	1.00	16,675
Professional Background													
	Self-employed	0.32	0.47	0.00	1.00	1.00	32,023	0.19	0.39	0.00	1.00	1.00	16,675
	Previously Employed	0.61	0.49	1.00	1.00	1.00	32,023	0.69	0.46	1.00	1.00	1.00	16,675
	Part-time Entrepreneur	0.22	0.42	0.00	1.00	1.00	32,023	0.24	0.43	0.00	1.00	1.00	16,675
	Same Prior Industry	0.59	0.49	1.00	1.00	1.00	32,023	0.59	0.49	1.00	1.00	1.00	16,675
	Serial Entrepreneur	0.03	0.18	0.00	0.00	1.00	32,023	0.05	0.21	0.00	0.00	1.00	16,675
	Previously Employed in Small Firm	0.60	0.49	1.00	1.00	1.00	32,023	0.57	0.50	1.00	1.00	1.00	16,675
	Previously Inactive	0.00	0.00	0.00	0.00	0.00	32,023	0.00	0.00	0.00	0.00	0.00	16,675
	Below High School Degree	0.24	0.43	0.00	1.00	1.00	32,023	0.18	0.38	0.00	1.00	1.00	16,675
	Undergraduate Degree	0.28	0.45	0.00	1.00	1.00	32,023	0.29	0.45	0.00	1.00	1.00	16,675
	Graduate Degree	0.18	0.38	0.00	1.00	1.00	32,023	0.27	0.44	0.00	1.00	1.00	16,675
	Grande Ecole	0.06	0.24	0.00	0.00	1.00	32,023	0.09	0.29	0.00	0.00	1.00	16,675
	Completed Required Training	0.20	0.40	0.00	1.00	1.00	32,023	0.20	0.40	0.00	1.00	1.00	16,675
Motivation and Expectations													
	Expectation: Growth	0.36	0.48	0.00	1.00	1.00	32,023	0.47	0.50	0.00	1.00	1.00	16,675
	Expectation: Sustain	0.34	0.47	0.00	1.00	1.00	32,023	0.31	0.46	0.00	1.00	1.00	16,675
	Expectation: Rebound	0.06	0.24	0.00	0.00	1.00	32,023	0.04	0.19	0.00	0.00	1.00	16,675
	Motivation: Peer Entrepreneurs	0.10	0.30	0.00	1.00	1.00	32,023	0.12	0.33	0.00	1.00	1.00	16,675
	Expect to Hire	0.22	0.41	0.00	1.00	1.00	32,023	0.29	0.46	0.00	1.00	1.00	16,675
	Motivation: New Idea	0.15	0.36	0.00	1.00	1.00	32,023	0.18	0.38	0.00	1.00	1.00	16,675
	Motivation: Opportunity	0.35	0.48	0.00	1.00	1.00	32,023	0.38	0.48	0.00	1.00	1.00	16,675
	Innovation	0.49	0.50	0.00	1.00	1.00	32,023	0.50	0.50	1.00	1.00	1.00	16,675
Venture Characteristics													
	Paris-based	0.11	0.31	0.00	1.00	1.00	32,023	0.15	0.36	0.00	1.00	1.00	16,675
	Marseille-based	0.01	0.11	0.00	0.00	1.00	32,023	0.02	0.15	0.00	0.00	1.00	16,675
	Lyon-based	0.02	0.13	0.00	0.00	1.00	32,023	0.02	0.14	0.00	0.00	1.00	16,675
	Bordeaux-based	0.02	0.14	0.00	0.00	1.00	32,023	0.02	0.14	0.00	0.00	1.00	16,675
	Specialized Construction Industry	0.14	0.35	0.00	1.00	1.00	32,023	0.10	0.30	0.00	1.00	1.00	16,675
	Retail Trade Industry	0.08	0.26	0.00	0.00	1.00	32,023	0.05	0.22	0.00	0.00	1.00	16,675
	Wholesale Trade Industry	0.04	0.20	0.00	0.00	1.00	32,023	0.03	0.16	0.00	0.00	1.00	16,675
	High tech Industry	0.05	0.21	0.00	0.00	1.00	32,023	0.06	0.25	0.00	0.00	1.00	16,675
	B2B	0.32	0.47	0.00	1.00	1.00	32,023	0.37	0.48	0.00	1.00	1.00	16,675
	B2C	0.63	0.48	1.00	1.00	1.00	32,023	0.57	0.50	1.00	1.00	1.00	16,675

Continued on next page

		2014 cohort						2018 cohort					
	Variable	Mean	SD	p50	p90	p99	N	Mean	SD	p50	p90	p99	N
	International Customers	0.05	0.22	0.00	0.00	1.00	32,023	0.06	0.25	0.00	0.00	1.00	16,675
	Local Customers	0.61	0.49	1.00	1.00	1.00	32,023	0.58	0.49	1.00	1.00	1.00	16,675
	Domestic Customers	0.14	0.35	0.00	1.00	1.00	32,023	0.16	0.37	0.00	1.00	1.00	16,675
Venture Organization													
	Co-founders	0.14	0.35	0.00	1.00	1.00	32,023	0.16	0.37	0.00	1.00	1.00	16,675
	Outsourcing: Accounting	0.70	0.46	1.00	1.00	1.00	32,023	0.80	0.40	1.00	1.00	1.00	16,675
	Number of Employees	1.69	1.73	1.00	3.00	12.00	32,023	1.43	1.23	1.00	2.00	7.00	16,675
	10+ Clients	0.61	0.49	1.00	1.00	1.00	32,023	0.59	0.49	1.00	1.00	1.00	16,675
	Number of Paid Managers	0.15	0.46	0.00	1.00	2.00	32,023	0.12	0.38	0.00	1.00	2.00	16,675
	Customers from Prior Job	0.13	0.33	0.00	1.00	1.00	32,023	0.67	0.47	1.00	1.00	1.00	16,675
	Suppliers from Prior Job	0.68	0.47	1.00	1.00	1.00	32,023	0.27	0.44	0.00	1.00	1.00	16,675
	Help from Professionals	0.06	0.24	0.00	0.00	1.00	32,023	0.13	0.34	0.00	1.00	1.00	16,675
	Help from Family	0.16	0.36	0.00	1.00	1.00	32,023	0.00	0.00	0.00	0.00	0.00	16,675
	No External Help	0.39	0.49	0.00	1.00	1.00	32,023	0.42	0.49	0.00	1.00	1.00	16,675
Financial Characteristics													
(not included as input features)	Bank Loan	0.30	0.46	0.00	1.00	1.00	32,023	0.31	0.46	0.00	1.00	1.00	16,675
	Other Loan	0.09	0.28	0.00	0.00	1.00	32,023	0.08	0.26	0.00	0.00	1.00	16,675
	No Outside Financing	0.63	0.48	1.00	1.00	1.00	32,023	0.64	0.48	1.00	1.00	1.00	16,675
	Other Firm Financing	0.04	0.19	0.00	0.00	1.00	32,023	0.05	0.22	0.00	0.00	1.00	16,675
	Grant	0.05	0.22	0.00	0.00	1.00	32,023	0.04	0.18	0.00	0.00	1.00	16,675
Industries-Locations													
	Number of Industries						48						48
	Number of Regions						322						323

Table E.8 Description. We test the robustness of the algorithm's predictive accuracy in Table E.8. We report the predicted and realized performance of the best predicted performers (i.e., companies in $\mathcal{A}_s$ with the highest $\hat{m}(x_i)$) for various designs of our algorithm trained and tested on different sets, and for different selectivity thresholds s. Panel A tests various cohort-based training and test sets, and Panel B tests various training and test random splits. Columns 1 and 2 show the predicted and realized performance of the best predicted performers such that the total number of portfolio companies in $|\mathcal{A}_s| = 120$ equals that in $|\mathcal{V}_s|$. Columns 3 and 4 increase the number of selected best predicted performers by decreasing the selectivity threshold to $s = 0.5\%$ ($|\mathcal{A}_s| = 187$), and columns 5 and 6 decrease it further to $s = 1\%$ ($|\mathcal{A}_s| = 374$). See Table E.4 for more details on these sets of best predicted performers. Columns 7, 9, and 11 report the median percentile of these three sets of best predicted performers in the distribution of realized performance, and columns 8, 10, and 12 report the percentage overlap between the best predicted performers and the best realized performers for each respective definition of best performers. This exercise is conducted without any constraints on the pool $\mathcal{D}$ from which the algorithm selects.

Panel A: Cohort-based Splits

		Ave. Revenue at Age 5 (log)						Realized Distribution Rank					
		Best		$\mathcal{A}_{s=0.5\%}$		$\mathcal{A}_{s=1\%}$		Best		$\mathcal{A}_{s=0.5\%}$		$\mathcal{A}_{s=1\%}$	
Training	Test	Pred.	Real.	Pred.	Real.	Pred.	Real.	Pctl. (Med.)	Overlap (%)	Pctl. (Med.)	Overlap (%)	Pctl. (Med.)	Overlap (%)
1998	2002	5.62	4.93	5.46	5.01	5.19	4.81	93	7	92	8	90	13
1998	2006	5.72	4.94	5.57	4.57	5.31	4.36	94	4	92	4	88	8
1998	2010	5.98	5.29	5.8	5.07	5.5	4.86	95	6	94	7	91	10
1998, 2002	2006	5.69	5.35	5.51	5.36	5.21	4.79	97	8	96	4	93	10
1998, 2002	2010	5.89	5.62	5.69	5.56	5.38	5.22	95	6	95	7	94	11
1998, 2002, 2006	2010	5.81	6.05	5.61	5.8	5.29	5.52	96	7	95	7	94	9

Panel B: Random Splits

	Ave. Revenue at Age 5 (log)						Realized Distribution Rank					
	Best		$\mathcal{A}_{s=0.5\%}$		$\mathcal{A}_{s=1\%}$		Best		$\mathcal{A}_{s=0.5\%}$		$\mathcal{A}_{s=1\%}$	
Training & Test	Pred.	Real.	Pred.	Real.	Pred.	Real.	Pctl. (Med.)	Overlap (%)	Pctl. (Med.)	Overlap (%)	Pctl. (Med.)	Overlap (%)
1998	5.2	5	5.02	4.77	4.67	4.53	89	11	87	16	86	23
1998, 2002	5.27	5.47	5.08	5.2	4.79	4.87	95	12	93	14	90	20
1998, 2002, 2006	5.5	5.3	5.33	5.32	5.03	5.21	93	5	93	9	92	14
1998, 2002, 2010	5.83	5.77	5.6	5.57	5.23	5.17	97	8	96	14	93	16
1998, 2002, 2006, 2010	5.9	5.86	5.67	5.65	5.31	5.38	97	8	96	8	94	12

Table E.8: Algorithm Design Robustness: Best Predicted Performers' Realized Performance for Various Train and Test Sets and Selectivity Thresholds. This table tests the robustness of the algorithm's predictive accuracy by varying the designs of our algorithm, trained and tested on different sets and for different selectivity thresholds s. Realized and predicted revenue are in thousands of euros. ompanies that fail by age 5 are included and assigned zero revenue. Please refer to page 94 for a complete table description.