Venture Capital (Mis)Allocation in the Age of AI

Victor Lyonnet Léa H. Stern[*]

31st December, 2022

Abstract

How do venture capitalists (VCs) make investment decisions? Using a large administrative data set on French entrepreneurs that contains VC-backed as well as non-VC-backed firms, we use algorithmic predictions of new ventures' performance to identify the most promising ventures. We find that VCs invest in some firms that perform predictably poorly and pass on others that perform predictably well. Consistent with models of stereotypical thinking, we show that VCs select entrepreneurs whose characteristics are representative of the most successful entrepreneurs (i.e., characteristics that occur more frequently among the best performing entrepreneurs relative to the other ones). Although VCs rely on accurate stereotypes, they make prediction errors as they exaggerate some representative features of success in their selection of entrepreneurs (e.g., male, highly educated, Paris-based, and high-tech entrepreneurs). Overall, algorithmic decision aids show promise to broaden the scope of VCs' investments and founder diversity.

Keywords: venture capital, deal selection, stereotypes.

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

[*] For helpful comments, we thank Ajay Agrawal (discussant), Will Cong, Murray Frank, Nicola Gennaioli, Jorge Guzman (discussant), Camille Hebert (discussant), Johan Hombert, Chris Hrdlicka, Hyeik Kim, Arthur Korteweg (discussant), Josh Lerner, Clément Mazet-Sonilhac, David Robinson, Daniel Rock, Francesco Saita, Amin Shams, Andrei Shleifer, Michael Weisbach, Karen Wruck, and conference and seminar participants at the University of Colorado Boulder, Ohio State, University of Washington, the 2022 RFS-GSU Fintech conference, 2022 Frontiers in Finance Conference, University of Calgary, 2022 Private Equity Research Oxford Symposium, AI & Big Data in Finance Research Forum, 2022 NBER SI Entrepreneurship, HEC Paris, University of Mannheim, Bocconi University, the 2022 Carey Finance Conference, CKGSB, Virtual Corporate Finance Wednesdays, and Chicago Booth. Samin Jalali, George Nurisso, Boyu Wang, Daisy Wang, and especially Danial Salman provided excellent research assistance. We thank Wendy Hu, Greg Brown, and IPC/PERC for letting us use Burgiss data for parts of our analysis. Victor Lyonnet is with the Ohio State University (Email: lyonnet.1@osu.edu). Léa H. Stern is with the University of Washington (Email: leastern@uw.edu). All errors are our own.

1 Introduction

Each year, hundreds of thousands of entrepreneurs start new ventures. While not all aspire to create high-growth startups, for those who do, venture capital (VC) is the dominant source of financing (Lerner and Nanda, 2020). How do venture capitalists decide which ones to back? A typical VC considers approximately two hundred new ventures each year and ends up backing around four (Gompers et al., 2020). VCs view this deal selection as a crucial determinant of their returns, and successful deal selection hinges on their ability to assess new ventures' potential and predict their future performance. Without historical data on new firms and a complex set of entrepreneur and new firm characteristics however, selecting the most promising firms is an extremely difficult task. Which firms have the highest chance of success? Do VCs back the best firms? How do they make their investment decisions?

A key obstacle to answering these questions is the difficulty of observing VCs' full choice set, and of evaluating VCs' investment decisions within this choice set. To explore these questions, we make use of a unique dataset that includes VCs' full choice set to contrast VCs' selections of firms to those of an algorithmic policy that selects firms with the highest chance of success. Because there is no one-size-fits-all recipe for venture success, theory cannot guide the choice of model to make out-of-sample predictions of new ventures' performance. Some covariates likely matter in some cases but not others, and may interact in nonlinear ways. We therefore use machine learning (ML) prediction methods to form predictions of new firms' future performance. ML methods avoid relying on specific parametric assumptions and instead rely on a rigorous data-driven model selection that produces accurate predictions of the distribution of outcomes. These methods are designed to maximize out-of-sample predictive accuracy and are well adapted for structured data (Chen and Guestrin, 2016; Erel et al., 2021).

We use French administrative data on 123,511 new firms from four cohorts of entrepreneurs who created a firm between 1998 and 2010. Our sample is representative of the entire population of entrepreneurs and free from survivorship and selection bias.[1] We observe detailed information on over one hundred features of entrepreneurs and new firm characteristics. We use these input features to train a Gradient Boosting Trees algorithm (*XGBoost*) on the first three cohorts of entrepreneurs to predict new firms' performance, and evaluate the algorithm's out-of-sample predictions in the

[1] While VC investment involves a complex two-sided matching process subject to negotiations (Cong and Xiao, 2021), we observe a firm's VC-backed status in the aggregate, that is, whether it is backed by *any* VC. We are not limited to observing whether a firm matched with one particular VC, which mitigates concerns related to negotiations, the two-sided matching process of VC investment, and portfolio considerations.

last cohort of entrepreneurs. All results pertain exclusively to the algorithm's performance in the test set, which is left untouched during training.

Entirely circumventing the selective labels problem (in which non-selected ventures observably similar to selected ones may in fact have different outcomes due to unobservables), would require observing all firms' performance if they were VC-backed. While this is not possible due to the fact that only a small subset of firms receive VC-backing, we observe a close substitute: the operational performance for all firms, regardless of their VC-backed status. These "quasi-labels" (Erel et al., 2021) offer several key advantages. First, operational performance is available for *all* new firms, not only VC-backed firms. Second, to understand and evaluate VCs' decisions, our performance measure should be one that maps well with what VCs aim to predict when they make investment decisions. Firms' operational performance satisfies this criteria. Third, because of VCs' preference for skewness, firms in the right tail of our performance distribution should also populate the right tail of exit valuations distribution. We use Pitchbook data to show that ventures' revenue at age five (a typical VC investment horizon, see Acharya et al. (2013)) satisfies these requirements. Finally, any entrepreneur and/or venture characteristics affecting venture performance that are observable to VCs but not the econometrician (e.g., the entrepreneur's passion and drive) are accounted for by these quasi-labels. Therefore, the extent of the selective labels problem is limited, since "unobservables" that may lead to differences in outcomes for observably similar ventures are restricted to the impact of VCs' resources and support on firm performance (Chemmanur, Krishnan and Nandy, 2011; Puri and Zarutskie, 2012; Bernstein, Giroud and Townsend, 2016). We discuss the role of this VC treatment effect in our simple framework below.

We first show that our algorithm makes accurate out of sample predictions of the distribution of firm outcomes, and we verify that this predictive accuracy is not dependent on our characterization of venture performance.

Next, we design an investment policy that selects the 115 most promising ventures based on these algorithmic predictions. We study patterns of discrepancy between these firms and the 115 VC-backed firms in our test set. We do not assume that the algorithm's predictions are correct. The algorithm's predictions help us isolate potential errors, but we always rely on realized outcomes to identify actual errors.

The only restrictions the algorithmic investment policy is subject to in this exercise is to limit the investable pool to firms in industries that receive VC in our data. This analysis reveals two potential types of errors in VCs' investment decisions. First, VCs invest in some firms that perform *predictably* poorly. Second, VCs pass on some new firms that perform *predictably* well. Although the performance of this unconstrained policy is interesting as an assessment of the upper-bound performance gains from introducing a VC algorithmic decision aid, there are many reasons why this policy is not realistic. First, VCs are typically specialized in specific industries in which their expertise allows them to best support their portfolio firms. Second, VCs cannot always invest in companies that are geographically distant, perhaps due to their on-site involvement with their portfolio companies (e.g., Bernstein, Giroud and Townsend, 2016). To account for VCs' constraints as well as demand-side considerations, we design a menu of *constrained* algorithmic investment policies as realistic algorithmic counterfactuals to VCs' selection. In our preferred counterfactual, the constrained algorithmic policy is restricted to investing in the most promising firms that must match both the industry and the location of the VC-backed firms it chooses to replace.

We face a trade-off in defining the algorithmic investment policy's pool. On the one hand, we would like to zoom in on firms that VCs could have invested in and that would have accepted VCs. On the other hand, we do not want to exactly reproduce VCs' portfolio. We show results for policies at different points along this trade-off. Our preferred constrained policy's investable pool comprises new ventures in the same industries and locations as the VC-backed firms in our test set. We find that the average revenue at age 5 of portfolio firms selected by this algorithmic policy are higher than the average of VC-backed portfolio firms.

We obtain similar results with alternative measures of venture success, including measures that account for VCs' preference for skewness (i.e., "home run" deals). We use two definitions of home runs: firms that experience a successful exit (M&A, IPO, or additional funding round), and firms that end up in the top 5% of operating performance in their cohort. We show that the firms selected by our main algorithm (which predicts operating performance) perform better not only on average, but they are also more likely to become home runs. Regardless of the home run measure we use, any algorithmic policy from a model predicting home runs attains a higher average operating performance than VC-backed firms. These findings suggest that our main algorithm captures VCs' preference for skewness.

Why do firms selected by the algorithmic policy perform better? To understand why VCs pass on some predictably good performers and select some predictably bad ones, we train a model that predicts for each new firm whether it is VC-backed. This model predicts VCs' decisions well, which confirms our prior that these decisions are not random. One striking result is that almost half of the predictable component of VCs' decisions can be attributed to three entrepreneur demographics (gender, age, education). In comparison, our predictive model of firm performance does not perform well when it is restricted to these three features. This confirms the idea that entrepreneurial success is driven by a complex combination of multiple factors. That much of the predictability in VCs' decisions comes from three demographics features can be viewed as suggestive evidence that VCs rely on a sparse model to make investment decisions.

Comparing VC-backed firms to algorithm-selected firms, we find that VCs select entrepreneurs whose features are representative of the most successful entrepreneurs, that is, characteristics that occur more frequently among the best performing entrepreneurs relative to the other ones (Tversky and Kahneman, 1974; Bordalo et al., 2016). Given their preference for skewness, it is not surprising that VCs select firms whose characteristics fit the stereotype of the best performing firms. We complement our analysis of stereotypes using U.S. data from Burgiss on deal-level returns, and find that in the U.S. also, VCs tend to invest more into geographical areas (e.g., California) and industries (e.g., Information Technologies) that are representative of the most successful U.S. entrepreneurs.

We then ask whether VCs' reliance on a restricted set of entrepreneur features is efficient (Lerner and Nanda, 2020). We follow the approach in Mullainathan and Obermeyer (2022) and create simple models that predict VCs' decisions based on a restricted set of features that are representative of the best performing firms. This analysis shows that VCs exaggerate some representative features of success in their selection of entrepreneurs. In particular, VCs tend to overweight gender (Howell and Nanda, 2019; Hebert, 2020), education (Queiró, 2021), optimism (Landier and Thesmar, 2008), startup experience, and the venture's industry and location (Chen et al., 2010).

This paper contributes to the literature on VCs' decision making (Kaplan and Strömberg, 2004; Gompers et al., 2020). Recent work has documented the importance of founding teams in attracting VC and the presence of frictions in VCs' decision making (Hellmann and Puri, 2002; Kaplan, Sensoy and Strömberg, 2009; Bernstein, Korteweg and Laws, 2017). We contribute to this literature by leveraging algorithmic predictions to quantify the shadow cost of constraints faced by VCs, and to show that VCs do not always select the most promising entrepreneurs. Our results show that algorithmic decision aids hold promise to broaden the range of businesses that receive private capital, addressing key concerns about the narrowness of the VC industry raised in Lerner and Nanda (2020).

Our article is related to recent work that uses machine-learning tools to predict new firms' potential (e.g., Arroyo et al., 2019; Ferrati, Muffatto et al., 2021; Te et al., 2022; Davenport, 2022). These papers use commercial data on VC-backed firms such as Pitchbook, Crunchbase or Linkedin, hence they are subject to the selective label issue and to selection and look-ahead biases (Kleinberg et al., 2018; Żbikowski and Antosiuk, 2021). Instead, our analysis relies on administrative data that is representative of all new firms in France. These data allow us to not only identify predictably poor performers among VCs' portfolio firms, but also predictably good performers among the firms VCs did not select. Thanks to detailed information on entrepreneurs and their startups, all at the time of firm creation, we document the dimensions along which VC-backed firms differ from algorithm-selected ones and use the method developed in Mullainathan and Obermeyer (2022) to study the reasons why VCs do not select the best entrepreneurs.

Our approach helps reconcile a series of existing evidence on VCs' investment decisions and identify their root cause. Consistent with Azoulay et al. (2020), we find that VCs select more young entrepreneurs than the algorithmic policy. We also find that VCs select fewer female entrepreneurs compared to the algorithmic policy, in line with evidence that investors appear to be biased towards male entrepreneurs (e.g., Raina, 2019; Balachandra et al., 2019; Ewens and Townsend, 2020; Gornall and Strebulaev, 2020; Hebert, 2020; Hu and Ma, 2020; Calder-Wang and Gompers, 2021). Overall, our findings are consistent with homophily and network effects (e.g., Hochberg, Ljungqvist and Lu, 2007; Howell and Nanda, 2019; Gompers et al., 2020). They are also consistent with VCs' performance being driven by their non-local investments (Chen et al., 2010), and could in part explain why entrepreneurs migrate after having founded their firm in a VC hub (Bryan and Guzman, 2021).

Our results suggest that VCs' decisions are consistent with models of stereotypical thinking (Gennaioli and Shleifer, 2010*a*; Bordalo et al., 2016). Although representativeness-based stereotypes are accurate, in the sense that they arise from true differences between groups, they induce VCs to exaggerate the true differences between entrepreneurs. Our findings provide an account of how VCs make decisions, explain the observed standard casting of the stereotypical entrepreneur, and rationalize why a machine learning algorithm outperforms VCs' selection of entrepreneurs.

2 Framework

We propose a simple framework of VCs' investment decisions that builds on Kleinberg et al. (2018) and Erel et al. (2021). In our two-period model, each new venture i is characterized by features (x_i, z_i). Both x_i and z_i are observed by VCs, but only x_i is recorded in the data and observable to the econometrician. Features z_i are unobservables that affect firm performance. At $t = 0$, VCs observe (x_i, z_i) for the new ventures in their investable pool, $\mathcal{D}$, and form rational predictions of performance if VC-backed, y_i, which map into a percentile ranking of new ventures $R(x_i, z_i)$.

VCs choose an investment policy h:

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

such that they maximize their expected payoff:

$$\pi_{VC}(h) = \sum_{i \in \mathcal{D}} h_i E[y_i|h]$$

VCs' *optimal* policy is therefore to invest only in the most promising new firms:

$$h_i = 1 \text{ iff } R(x_i, z_i) > t, \tag{1}$$

with $P(y \leq P_t) = t$ the percentile threshold under which the VCs do not invest and t is such that $\|h\|_0 = N$.[2]

At $t = 1$, performance y_i is realized.

We ask whether the VCs' actual policy deviates from the optimal policy:

$$h_i = 1 \text{ iff } R(x_i, z_i) > t + \Delta(x_i, z_i) \tag{2}$$

where $\Delta(x_i, z_i)$ captures shifts in the investment threshold for entrepreneurs with characteristics (x_i, z_i), and makes VCs depart from the optimal investment policy in (1). We therefore search for an alternative investment policy α:

$$\alpha \in \{0,1\}^{|\mathcal{D}|} \ and \ \|\alpha\|_0 = N$$

with expected payoff

$$\pi_{VC}(\alpha) = \sum_{i \in \mathcal{D}} \alpha_i E[y_i|\alpha_i]$$

such that

$$\pi_{VC}(\alpha) > \pi_{VC}(h)$$

We do not observe $E[y_i|\alpha]$, the venture's performance if it were VC-backed, when the firm is selected

[2] The threshold t is determined outside our model and depends on VCs' financing and operational constraints. Empirically, $t \simeq 0.995$, so that VCs invest in only about 0.5% of all new firms (Lerner and Nanda, 2020). For the 2010 cohort of entrepreneurs, which is our test test, $N = 115$.

by the algorithmic policy but not by VCs (the selective labels problem). However, we observe *quasi-labels* $E[y_{-i}|\alpha]$, the operational performance for all ventures, which account for z_i. Therefore, the effect of unobservables on firm performance is limited to the VC treatment effect: $E[y_i|\alpha]-E[y_{-i}|\alpha]$. Therefore, VCs do not follow optimal policy if there exists α such that

$$\underbrace{E[y_{-i}|\alpha]}_{performance\ for\ alternative\ firms} > \underbrace{E[y_i|h]}_{performance\ for\ VC-backed\ firms}$$

where the underlying assumption on the VC treatment effect is that

$$\underbrace{E[y_i|\alpha]-E[y_{-i}|\alpha]}_{treatment\ effect} > -\Big(\underbrace{E[y_{-i}|\alpha]-E[y_i|h]}_{difference\ in\ performance}\Big)$$

3 Data

We construct a novel data set using three sources of administrative data from the French Statistical Office (INSEE): a representative survey of entrepreneurs conducted every four years that contains a wide array of entrepreneur and new firm characteristics, the French firm registry that allows us to track the exhaustive list of new firms, and accounting data from the tax files.

3.1 Data Sources

Entrepreneur survey. Our main data source is a large-scale survey of entrepreneurs in France (*Système d'Information des Nouvelles Entreprises*, or *SINE*), which is conducted by the French Statistical Office every four years from 1998 to 2014. Our sample comprises 123,511 entrepreneurs from four cohorts (1998, 2002, 2006, 2010). The two main advantages of these data for our study are that (i) they contain a large set of new firm founders' characteristics (48 questions are sent to entrepreneurs, which become more than 140 characteristics once we encode the responses) and (ii) they are representative of all new firms in the French economy and not subject to any selection biases commonly encountered in the literature.[3]

The absence of survivorship bias and selection bias is key for our analysis. In contrast to the existing literature on VC, which is restricted to standard data sources collected from VCs and hence

[3] The French Statistical Office sends questionnaires to approximately 25% of entrepreneurs who started or took over a business in France that year. Our analysis focuses on new businesses, which represent approximately 80% of the surveyed entrepreneurs. The surveyed firms are randomly selected from the exhaustive firm registry. The business owner is responsible for completing the documents. The response rate to the SINE survey is high (approximately 90%) because the tax authorities supervise the sending of questionnaires.

focuses on VC-backed firms in isolation, our sample contains both VC-backed and non-VC-backed firms.[4,5] Appendix C contains descriptions for a subset of the variables we use from the survey. The questionnaire includes questions about sources of financing, which we use to determine firms' VC-backed status.

Accounting data. Another important advantage of our data is that we can observe firm performance without relying on VC commercial data sets, which are subject to reporting biases.[6] Instead, we use 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.[7] We observe firm performance at different ages from total sales and value added reported in the tax files.

Firm registry. We use data from the firm registry (*SIRENE*) for the period 1998 to 2015.[7] 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 (e.g., Sole Proprietorship, Limited Liability Corporation, Corporation), the official creation date and geographical location. We use the firm registry to construct an exit 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. We obtain data on whether new firms get acquired before age 6 by merging the French administrative data with commercial M&A data sets from SDC platinum and Bureau Van Dijk's Zephyr. We also construct an IPO dummy equal to one in the year a firm from our sample goes public using data from Orbis.

Pitchbook data on exit valuations. Because deal-level returns data is unavailable for the French VC-backed firms in the SINE survey, we use data on exit valuations from Pitchbook. The

[4] Two notable exceptions are Chemmanur, Krishnan and Nandy (2011) and Puri and Zarutskie (2012), which use the Longitudinal Business Database (LBD), a panel data set collected by the US Census Bureau, to identify firms that do and do not receive VC financing. Because the US Census data lack information on entrepreneur characteristics, our analysis could not be conducted on US data. A few other studies examine smaller hand-collected samples of private VC-backed and non-VC-backed firms but are limited to certain geographies, time periods, industries, and firm outcomes (e.g., Hellmann and Puri, 2000, 2002).

[5] The entrepreneur survey for the 2006 cohort does not allow the identification of VC-backed firms. We therefore exclude the 2006 cohort whenever we focus the analysis on VC-backed firms.

[6] See, e.g., Gompers and Lerner (2001), for a discussion of how VCs often underreport poorly-performing deals.

[7] 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.

Pitchbook data is unique in that it reports both the operational performance and exit valuation of startups. This allows us to test the correspondence between the firms' revenue – the measure of performance our algorithm predicts – and their valuations at exit. For our sample to be large enough we keep all French and US VC-backed exits available in Pitchbook. We end up with 350 French firms and 7,593 U.S. firms in our sample.

Burgiss data on deal-level returns to investments. The Burgiss data often are described as the gold standard for fund-level VC returns (see, e.g., Kaplan and Lerner, 2016; Brown et al., 2020). To provide data on the underlying deal-level information, Burgiss gathers data from the financial reports of general partners whose investors are Burgiss clients. We use the two measure of returns available in the 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 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.

4 Algorithmic Predictions Design

We want to test whether VCs invest in the most promising firms, i.e., those for which $R(x,z) > t$. To test for deviations from that objective, we use our sample of entrepreneurs to approximate the percentile rank of rational performance predictions $R(x_i, z_i)$ using an estimator of firm performance $\widehat{m}(x_i)$ that takes characteristics of entrepreneur i as its input vector x_i. In this section, we start by explaining why we choose revenue as a performance measure $Y(x,z)$ to predict to produce a percentile rank of performance predictions $R(x,z)$. We then describe how we train and evaluate an algorithm to generate these performance predictions.

4.1 Choice of predicted performance measure

How do VCs generate returns from their investments and what does it imply for our choice of performance measure $Y(x,z)$ we ask the algorithm to predict? Gompers et al. (2020) describe VCs' decision process. This process starts with the VC observing the venture's characteristics x_i and z_i, then deciding whether to meet the management of the venture, then evaluating and conducting due diligence on the venture. If the venture goes through all these steps successfully it is offered a term sheet and decides whether to accept it.

The term sheet specifies terms and conditions of the investment (the amount invested by the VC, the VC's percentage ownership in the firm, liquidation preference, voting rights, etc.) which altogether determine the return the VC expects to receive from a given investment. Several return measures can be calculated, such as the cumulative distributed total value to paid-in capital (TVPI) or the internal rate of return (IRR).

The key advantage of our analysis is the use of administrative survey data that allow us to largely circumvent the selective labels problem by training our algorithm on all new firms, not just VC-backed ones (Kleinberg et al., 2018). To train our algorithm, we thus need a measure $Y(x, z)$ of venture performance that is available for all new firms and is highly correlated with returns to VCs. We choose firms' revenue for several reasons: exit valuations are often 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). In our main analysis we task the algorithm with predicting new firms' revenue 5 years after creation, an horizon that is between the median forecast period of 3 to 4 years documented by the VCs surveyed in Gompers et al. (2020) and the 6.6 average years between the seed and exit rounds we find in the Pitchbook data. In section 4.4 we check the robustness of our analysis to using alternative outcomes measures, including other investment horizons.

To build confidence in the fact that a venture's revenue is highly correlated with VCs' returns, and specifically for the high-returns firms VCs care about, we use Pitchbook data. Although the Pitchbook data do not contain deal-level returns, they report firms' valuation at exit, and every return measure is increasing in firms' valuation at exit. The Pitchbook data also contain two operational performance measures of the underlying ventures: revenue and ebitda at exit. We analyze the correspondence between returns and operational performance in two ways. First, we compute rank-to-rank correlations and find that a firm's rank in the distribution of revenue at exit is positively and significantly correlated with its rank in the distribution of exit valuation. Interestingly, we find the opposite result with a firm's rank in the distribution of EBITDA.[8] This finding is consistent with the fact that many VC-backed firms exit when they generate cash flow

[8] Spearman's and Kendall's rank correlation coefficients between revenue and valuation at exit are 0.45 and 0.32, respectively, with independence rejected at the 1% level. Instead, Spearman's and Kendall's rank correlation coefficients between EBITDA and valuation at exit are -0.17 and -0.14, respectively, also with independence rejected at the 1% level.

but while they are still cash flow negative (for instance, Uber, Amazon, or Airbnb). Second, we focus on firms in the top of the distribution of revenue at exit in Table 1. Rows 1, 2, and 3 of Table 1 respectively focus on firms in the top 1%, 5%, and 10% of the revenue distribution. For each set of firms, 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 firms in the top 1% of revenue end up above the 96th percentile of exit valuation. Although the percentile ranks of the average and median firm decrease in rows 2 and 3, these firms remain in the top of the distribution of exit valuation. Although our Pitchbook sample is mostly comprised of US firms, the evidence in Table 1 suggests that VCs' returns are highly correlated with ventures' revenue.

In addition, indirect pay for performance from future fundraising motives should incentivize VCs to care about their portfolio firms' 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 are 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 firms' 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 firms' revenue provide a signal to LPs to assess their skills, fund managers are incentivized to care about portfolio firms revenue beyond the direct relation between portfolio firms' revenue and fund manager compensation.

4.2 Algorithm Design

Algorithm class and train/test sets. We use Gradient Boosting Trees (*XGBoost*) to generate performance predictions (Chen and Guestrin, 2016). *XGBoost* is trained on three cohorts of entrepreneurs (1998, 2002 and 2006) representing 69% of our data (85,395 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 (37,823 observations, or 31% of our data). We index each observation by i throughout the paper. 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. Our train/test split is based on cohorts rather than a random split for three reasons. First, this approach avoids using outcomes of firms created in the future to make performance predictions. Second, it sets a level playing field for the algorithm against VCs, ensuring that both would only be able to observe the performance of past new firms before selecting new firms. 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.

Input features. To generate predictions of future operating performance, the algorithm uses a set of 140 covariates (452 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.[9]Because our objective is to study VCs' decision making, we ensure that all input features are *ex ante* covariates, i.e., the information used by the algorithm would easily be accessible to any VC during a first-pass evaluation of the venture. We therefore omit several covariates that are available in the survey that we deem not readily available to a potential investor (e.g., bank loans). Table 2 reports summary statistics for a subset of input features. We report these statistics separately for the training set and the test set.

[Insert Table 2 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 education are somewhat larger in the test set.[10]

[9] To facilitate the interpretation of our results, we focus on new firms comparable to VC-backed firms by excluding firms whose founder reported more than 20 employees in the year of creation and firms with a revenue greater than 11 million euros in their creation year (the maximum revenue observed for VC-backed firms in their creation year). The performance of our algorithm is not sensitive to this filter.

[10] Bonelli, Liebersohn and Lyonnet (2021) study the time-series evolution of entrepreneurship quality over time,

4.3 Model predictions vs. realized firm performance

All firms in test set. We begin our analysis by comparing the algorithm's predictions of operating performance $\hat{m}(x_i)$ to the observed realized performance y_i (the log of revenue generated at age 5) for the 37,823 observations in our test set. Figure 1 plots a binned scatterplot depicting the relationship between algorithmic predictions and the observed outcome among *all* new firms in our representative sample, that is, both VC-backed firms and non-VC-backed firms. Each point represents the average realized performance for new firms grouped in bins according to their predicted performance. Figure 1 illustrates the algorithm's ability to predict the distribution of new firm success reliably.

[Insert Figure 1 here]

Most promising firms in test set. In a typical year, only about 0.5% of new ventures receive VC funding, i.e., in the data, t in (1) is about 0.995 on average. This fraction is similar to that in the US (Puri and Zarutskie, 2012; Lerner and Nanda, 2020). The algorithmic policy selects firms in the top $s = 1 - t$ of predicted performance y_i. The algorithmic policy writes:

$$\alpha_i = 1 \text{ iff } M(x_i) > t, \tag{3}$$

where $M(x_i)$ is the algorithmic prediction's percentile rank for entrepreneur i. In Figure 2, we report the performance of the algorithmic policy as we increase its selectivity by increasing the investment threshold t. First, we find that the average performance of selected firms increases as we increase the value of t. Second, despite the large variance in the outcome variable in the data, Figure 2 shows that the algorithm is able to reliably rank new firms according to their potential, regardless of the fraction of new firms we ask it to select (i.e., for all values of t). Our interpretation of Figure 2 is that even within the subset of most promising firms in the test set, the algorithm is able to produce a useful ex ante ranking of firms. In other words, the algorithm demonstrates predictive ability along the entire distribution, even in the right tail. The algorithm shows promise not only by successfully avoiding to fund firms with low potential, but also by (re)allocating capital within the set of most promising ventures.

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 firm performance make it more difficult for an algorithm trained on the earlier training set to predict the performance of new firms in the later test set.

[Insert Figure 2 here]

4.4 Omitted payoffs

Because portfolio firms' valuations are commonly based on multiples of revenue, VCs' returns are highly correlated with their portfolio firms' performance. We verified the correspondence between revenue and valuation at exit using data from Pitchbook in Table 1 (see Section 4.1). In this section, we verify that our results are robust to using alternative measures of firm success.

Home run deals. We use two measures to capture VCs' preference for skewness (i.e., "home run" deals). For the first measure, we follow the literature and define home run deals as firms that are either acquired, go public, or raise funds in later stage rounds (Fazio et al., 2016; Guzman and Stern, 2020). To identify such successful exits in our sample, we match the French administrative data with data from SDC platinum and Bureau Van Dijk's Zephyr as well as Crunchbase, VentureXpert, CapitalIQ, CBInsights. We create a dummy variable *Successful Deals* equal to one for successful exits. We argue that the deals we identify as successful are unambiguously associated with a success for VCs. Not all firms that subsequently experience a successful exit are VC-backed upon their creation in the survey year. We thus ask whether an algorithm could aid VCs hit those "home run" deals.[11]

These successful exits are extremely rare. In our test set (the 2010 cohort of entrepreneurs), only 52 companies have such exits out of almost 38,000 observations. Despite the difficulty of the task, we use our training set to fit a separate model whose goal is to predict these low probability events. We use an *XGBoost* classifier which generates a probability $P[x_i]$ of successful exit for each observation i in the test set. We assign a percentile rank $M'[x_i]$ for each entrepreneur i based on these probabilities.[12] As before, we implement an algorithmic policy which selects the firms above the investment threshold t, $\alpha_i = 1$ iff $M'[x_i] > t$. Once again, to compare the performance of the algorithmic investment policy to that of VCs, we set the investment policy threshold such that the algorithm selects the top 0.5% of firms ($t = 99.5\%$). Out of the 115 VC-backed firms in the test set, 4 have a "successful exit," therefore VCs' decisions have a precision of 3.3% and a recall of 7.1%.[13]

[11] Due to data limitations, we are not able to ensure that acquisitions are made at a premium or that the initial VC (if any) indeed exits the deal. We assume that most of the successful exits we identify are viewed as a success by VCs.

[12] The percentile rank $M'[x_i]$ is different from that of performance predictions, $M[x_i]$ defined in Section 4.3.

[13] Precision is calculated by dividing true positives by the sum of true positives and false positives, and recall is calculated by dividing true positives by all positives (false negatives plus true positives). Both precision and recall are similar in previous cohorts of entrepreneurs for which we observe VC-backed status.

This is the benchmark we use to evaluate the performance of our algorithm.

The algorithmic investment policy we implement identifies *ex ante* 10 firms that subsequently experience a successful exit.[14] Therefore, our prediction model of home runs has a precision of 5.8% and a recall of 19.6%, both of which are higher than VCs'. While there are obvious data limitations to this exercise, we view this finding as highly encouraging, especially in light of the existing literature which has shown that VCs' involvement with their portfolio companies leads to higher exit rates in the form of acquisitions and IPOs, and better performance (Puri and Zarutskie, 2012; Bernstein, Giroud and Townsend, 2016).

The second measure of home run deals is a dummy equal to one for firms in the top 5% of their cohort in terms of revenue at age 5. In this exercise, instead of predicting firms' performance, our algorithm is trained to classify firms according to whether they will be among the best performers in their cohort. Table 3 contains the results. We find that 59% of algorithm-selected firms are among the best performers in their cohort, compared to only 15% of VC-backed firms (last row).

[Insert Table 3 here]

Other outcome measures. Table 3 contains the performance of predictive models trained on several outcome measures (first column). Importantly, we find that firms selected by a model that predicts one measure of success do at least as well as VCs not only based on that specific measure, but also on all other success measures.

5 Algorithmic Investment Policy

5.1 Unconstrained algorithmic investment policy

Comparing VC-backed and algorithm-selected firm performance. We continue our analysis of the algorithm's predictive ability by comparing the performance y_i of VC-backed firms to that of algorithm-selected ones.[15] Figure 3 reports the distribution of realized outcomes for all new firms, VC-backed firms, and algorithm-selected firms.

[Insert Figure 3 here]

[14] This predictive model of "successful deals" has an area under the curve (AUC) of .84. This implies that for two randomly picked firms, one a successful exit and one not, the odds that our model assigns a higher probability of being a successful exit to the one that indeed is a successful exit is 84%.

[15] Recall that we drop firms that operate in industries that never receive VC funding during our sample period to focus on firms that are more suited to receive VC funding and those that report having less than 20 employees. Our results remain qualitatively similar without this filter.

Figure 3 illustrates several interesting facts. First, the average VC-backed firm generates higher revenue at age 5 than the average firm: the average log of revenue is 2.86 for VC-backed firms, whereas it is 2.43 for the entire sample. This performance gap confirms VCs' ability to identify and invest in promising new ventures.[16] Second, we confirm the *Babe Ruth Effect* in our data: VCs bet on magnitude over frequency, and outcomes follow a power law distribution.[17] On the one hand, VCs are more likely to invest in firms that die within 5 years than the base rate. On the other hand, conditional on surviving, their portfolio firms do better than average. Third, the realized average performance of algorithm-selected firms when the algorithm can select any of the firms predicted to perform well (in red) is greater than that of VC-backed firms. This is not just an average effect. The algorithm achieves this not by selecting average-performing firms, but by avoiding more firms that fail within 5 years and identifying more super performers among surviving firms. As a result, only 3 out of 115 (3%) firms overlap between the set of algorithm-selected firms and VC-backed ones.

These distributions represent a first indication that the process by which VCs acquire and aggregate signals about a venture's prospects may be inefficient. Overall, these results suggest that VCs' policy in Equation (2) differs from the optimal policy in Equation (1), such that $\Delta(x, z) \neq 0$.

5.2 Constrained algorithmic investment policy

The algorithmic policy in Section 5.1 represents the highest possible performance that our algorithm can achieve when the algorithmic policy is allowed to select *any* new venture (as long as it is in an industry that has received VC in our sample period). Although the performance of this unconstrained policy is interesting as an assessment of the upper-bound performance gains from introducing a VC algorithmic decision aid, there are many reasons why this policy is not realistic. First, VCs are typically specialized in specific industries in which their expertise allows them to best support their portfolio firms. Second, VCs cannot always invest in companies that are geographically distant, perhaps due to their on-site involvement with their portfolio companies (e.g., Bernstein, Giroud and Townsend, 2016).

To account for VCs' constraints as well as demand-side considerations, we design a menu of *constrained* algorithmic investment policies as realistic algorithmic counterfactuals to VCs' selection. In our preferred counterfactual, the constrained algorithmic policy is restricted to investing in the

[16] This difference may in part be attributable to VCs' involvement in managing the firm they invest in.

[17] https:www.businessinsider.com/babe-ruth-grand-slams-and-startup-investing-2015-6.

most promising firms that must match both the industry and the location of the VC-backed firms it chooses to replace. Due to these constraints, we find a larger overlap between algorithm-selected firms and VC-backed ones: 29 out of 115 firms (25%) that are selected by the constrained algorithmic policy in the test set are VC-backed in the data. Despite this overlap, Figure 3 shows that the realized average performance of algorithm-selected firms when the algorithm is restricted to select any of the firms predicted to perform well in the same industry and location as VC-backed firms (in blue) is greater than that of VC-backed firms.

5.3 Counterfactual models of VC allocation

We would like to assess VCs' deviation from the objective to invest in the most promising firms defined in Equation (2), and evaluate the potential performance gains from introducing a VC algorithmic decision aid.

One way to perform this evaluation is to create a counterfactual model that sequentially drops VC-backed firms with the lowest predicted performance and replaces them with available firms with the highest predicted performance. We first run the counterfactual model without constraints ("unconstrained counterfactual"). We then impose a set of constraints that mimic the ones VCs may be subject to. For each VC-backed firms it drops, our first constrained counterfactual model lets the algorithm select a new venture only within the same industry as the VC-backed firm it drops. Our second constrained counterfactual model can only choose firms within the same location as the VC-backed firm it drops. Our third counterfactual model is constrained to pick a firm within the same industry *and* location for each VC-backed firm it drops.

[Insert Figure 4 here]

Figure 4 shows that as the counterfactual model increases the number of firms it replaces (along the x-axis), it puts more weight on the algorithm's selections and less weight on VCs' selections. The leftmost point shows the status quo of VCs' selection of firms, and the rightmost point for each line shows the algorithm's selection of firms.

This analysis yields several interesting results. First, all counterfactual models outperform VCs' selections. Second, when the counterfactual models assign full weight on the algorithm's selections, we interpret the difference in average portfolio firm performance between the unrestricted counterfactual model and each counterfactual model subject to a given constraint as the shadow cost of this constraint. As expected, the more restrictive the set of constraints, the lower the portfolio performance of the counterfactual model. We show the results for a counterfactual model constrained to entrepreneurs who have the same growth aspirations as VC-backed ones (in orange).[18,19] We then further restrict this set by adding the constraint that algorithm-selected entrepreneurs must be in the same industry as dropped entrepreneurs (in purple).

Even our most constrained algorithm significantly outperforms VC-backed firms, which suggests that VCs' constraints cannot fully explain the difference in performance between VC-backed and algorithm-selected firms. In unreported results, we find that our conclusions are unchanged when we account for demand side constraints by further constraining the algorithmic policy to investing in firms whose founders declared financial constraints to be a top concern when creating their venture.

Our results suggest that VCs pass on some firms that perform predictably well even though these firms closely resemble the firms they typically select.[20] We argue that demand-side or supply-side arguments are unlikely to explain the characteristics along which algorithm-select firms differ from VC-backed ones (e.g., gender, location). Indeed, it seems implausible that characteristics such as the gender and location of entrepreneurs fully explain why these entrepreneurs do not want VC-backing or why VCs could not back these firms. Therefore, it could be that the VCs' policy in (2) is such that $\Delta(x, y) \neq 0$ even absent constraints to VCs' investment decisions. We explore other explanations in Section 7.

[Insert Figure 5 here]

Of course, there are several caveats to this approach. We do not have data on deal size or terms and are limited to expressing potential performance gains in terms of portfolio firms' average performance. Therefore, these gains do not capture VCs' returns directly; instead, they measure gains in terms of portfolio firms' average revenue, which we have shown to be highly correlated with VCs' returns, including in the right tail. Despite these data limitations, at a minimum, our findings reveal that VCs invest in firms that perform predictably poorly.

To partially address this concern, we use Pitchbook data to construct imputed valuations and imputed investment "multiples" for the firms selected by the two algorithmic policies. To construct

[18] The growth related questions in the 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?"

[19] Catalini, Guzman and Stern (2019) study the venture growth process with versus without venture capital and show that firms with growth potential are similar to each other, irrespective of whether they are VC-backed. They find a large overlap between the firm characteristics that predict VC-backed status and those that predict IPO or M&A events without VC. Their results attenuate concerns that the algorithm-selected firms have a fundamentally different growth path from the VC-backed ones.

[20] Algorithm-selected firms are also not different from VC-backed firms in terms of their size at creation. We test for this explicitly in the last row of Table 4.

imputed valuations, we first estimate the median revenue multiple at exit (post valuation / exit revenue) at the industry level in Pitchbook. We then multiply it by the firm's revenue at age 5. The median revenue multiple across industries is approximately four. Figure C.1 in the Appendix reports the results for the unconstrained counterfactual model and for the counterfactual model constrained to match the industry and location of VC-backed firms. To impute a firm's investment "multiple" which resembles TVPI, we multiply its imputed valuation by the median fraction acquired during early deals, multiplied by 75% to account for the typical dilution in early deals. This gives us the approximate median dollar amount to early investors at exit, which we then scale by the median early deal size. Results are reported in the Appendix. Figure C.2 reports the results. While these imputation exercises obviously require several strong assumptions, we find that evaluating our counterfactual models using imputed measures designed to more closely capture VCs' returns does not change the results. A back-of-the-envelope calculation implies that to reverse our main result that the counterfactual policy improves outcomes, revenue multiples for firms selected by the constrained counterfactual policy would need to be on average approximately equal to the 5th percentile of the revenue multiples distribution.

6 VC-backed vs. Algorithm-selected Firms

6.1 Systematic differences between VC-backed and algorithm-selected firms

One way to shed light on the process by which VCs gather signals of a venture's potential is to examine how various demographic measures of entrepreneurs differ across VC-backed entrepreneurs and algorithm-selected ones. Figure 7 reports the probability densities of founders' ages, gender, education level, and geographic location for VC-backed and the unconstrained algorithm-selected entrepreneurs.

[Insert Figure 7 here]

Age. Panel A shows that although the average founder age of VC-backed and algorithm-selected firms is approximately the same, VCs select a larger fraction of young entrepreneurs than the algorithm. This result is in line with findings in Azoulay et al. (2020) that investors overemphasize youth as a key trait of successful entrepreneurs.

Gender. Panel B examines differences in founders' gender. Female entrepreneurs represent 28% of entrepreneurs in our test set. Yet, only 9% of VC-backed ventures are female-led. While there might be several explanations for this low representation of female founders among the set of VC-backed firms, the literature has recently documented possible biases against female founders (e.g., Calder-Wang and Gompers, 2021). Our results show that an algorithm with no embedded gender or other 'in-group' preferences, but simply tasked with predicting venture success would increase female VC-backed entrepreneurs by about one third.

Education. In Panel C, we find that both VCs and the algorithm select more founders with a graduate degree relative to the base rate of 15%. However, VCs select approximately one third more entrepreneurs with a graduate degree relative to the algorithmic policy.

Geography. Finally, Panel D explores the role of geographic proximity in VCs' investment decisions. In our test set, only 8% of new firms are located in the Paris region. Yet, one in five VC-backed firms is located in Paris, which is a key investors cluster. This finding is consistent with the importance of networking effects documented in Howell and Nanda (2019). In contrast, the algorithm selects Paris-based ventures at a rate below the base rate.

Taken together, the results in Figure 7 illustrate the discrepancies in demographic features for VC-backed and algorithm-selected founders. To gain a better understanding of how VC-backed firms differ from algorithm-selected firms, beyond founders' demographics, we report in Table 4 the summary statistics for a subset of features as well as t-tests for difference in means for VC-backed and algorithm-selected ventures, for two investment policy thresholds (0.5% and 1%). Table 4 reveals several interesting patterns. We highlight a few results using the unconstrained algorithmic policy and s=0.5%.

[Insert Table 4 here]

Founder experience, motivation, and other input features. While 16% of entrepreneurs in the test set reported that the motivation for starting their company was a "new idea", 39% (7%) of VC (unconstrained algorithm)-selected founders reported this was their main motivation. In addition, in the test set, 61% of founders have experience in the same activity as their new company. This experience seems to not be valued by VCs to the same extent as it is by the algorithm. Only

52% of VC-backed founders have same prior activity experience, while 90% do among algorithm-selected founders. Finally, we note that the algorithm would deploy VC to a broader set of regions. This finding has important implications as an increase in geographic diversity would change the landscape of innovation in the economy by de-emphasizing the importance of financing hubs.

Firm size. Note that while total assets at the time of the entrepreneur survey is not part of the input features that we make available to the algorithm, we check that the size of algorithm-selected firms does not significantly differ from that of VC-backed firms. The last row of Table 4 shows that the difference in total assets for firms selected by the algorithmic policy and by VCs invest is not significantly different. The standard deviations in total assets for VC-backed and algorithm-selected firms are also very similar.

6.2 Stereotypes of the most successful entrepreneurs

Can stereotypes explain why VCs' decisions differ from the algorithmic policy? Can VCs make errors in judgment that arise from oversimplifying the representation of heterogenous entrepreneurs? To answer these questions, we focus on the characteristics along which VC-backed firms differ from algorithm-selected ones (Section 6). We ask whether these differences can be explained by stereotypes, which, as in Bordalo et al. (2016), we take to form as a consequence of Kahneman and Tversky's representativeness heuristic.

Stereotypical thinking would imply that VCs select entrepreneurs with these characteristics because they are representative of the best performing firms. To test this prediction, we follow Gennaioli and Shleifer (2010*b*), Bordalo et al. (2016) and Mullainathan and Obermeyer (2022), and calculate the *representativeness* of a characteristic X_i for a certain percentile P of the performance distribution relative to the rest of the distribution $-P$ as the ratio:

$$\frac{Pr(X_i \mid P)}{Pr(X_i \mid -P)} \tag{4}$$

Stereotypes of French Successful Entrepreneurs. We study firms that are founded by male entrepreneurs, firms based in Paris, entrepreneurs with a graduate degree, and high-tech firms. The empirical prediction of the stereotype model (Tversky and Kahneman, 1974; Bordalo et al., 2016) is that for the same level of predicted performance, firms whose characteristics are *representative* of success are more likely to be VC-backed.

In Table 5, we report the representativeness of each of the characteristics of interest for the best performing firms (in column 1), defined as those in the top 5% of the performance distribution, and for the other firms in the bottom 95% of the distribution (in column 2). Column 3 of Table 5 confirms that these characteristics are representative of the best performing firms. These results can rationalize why VCs select entrepreneurs with these characteristics: Given their preference for skewness, they select entrepreneurs whose characteristics fit the stereotype of the best performing firms.

[Insert Table 5 here]

Because VCs tend to select firms that are representative of the most successful ventures, their decisions are based on accurate stereotypes (Bordalo et al., 2016). However, column 5 of Table 5 suggests that VCs might amplify these representative features in their decisions (Bordalo et al., 2016). This column shows the ratio of the representativeness of each feature for the best performers in the training set over its representativeness for VC-backed firms in the test set, which is higher than one for most features. These results raise the question of whether $\Delta(x, z) \neq 0$ because VCs make errors when predicting which firms will become the best performers. We address this question in Section 7.

Stereotypes of U.S. Successful Entrepreneurs. We complement our analysis of stereotypes using Burgiss data on U.S. VC deal-level returns for two reasons. First, we want to ask whether the stereotypes of successful entrepreneurs we identified in the French data carry through the U.S. VC context. Second, the Burgiss data uniquely allows us to verify that the stereotypes we identified in the French data remain when measuring deal-level VC returns, so that they do not depend on measuring success based on operational performance.

In Table 6, we construct the representativeness ratio of the two deal characteristics that are available in the Burgiss data: firms' location and industry. We focus on the four largest U.S. states and the four largest industries in terms of deals number. We try two different definitions of successful ventures: whether a venture is in the top 5% or top 1% of the sample distribution of deal-level TVPI.

The findings in Table 6 imply that, regardless of our definition of "successful entrepreneurs", the representativeness ratio of new ventures in California (the U.S. state with the largest number of VC-backed firms) and in the I.T. sector (the industry with the largest number of VC-backed firms) are systematically greater than one. Although each individual largest U.S. state and industry in terms of VC is not systematically representative of successful VC-backed firms on their own, when we lump them together into "VC Hubs" and "Largest Industries" we find that VCs seem to invest more often into those geographic locations and industries that are representative of the most successful entrepreneurs. Those results are very much aligned with our previous analysis showing that Paris-based and High-Tech firms were representative characteristics of successful entrepreneurs.

Appendix Table C.1 shows the results when using the internal rate of return instead of TVPI as a measure of deal performance. Overall, the results in Table C.1 and Table 6 confirm our interpretation that stereotypes of successful entrepreneurs do not heavily depend on the measure of performance, nor that they would be specific to the French context.

7 What is Driving VCs' Bias?

The above analysis raises the question of what aspects of VCs' decision making lead them to make investment decisions that differ from the algorithmic policy. The results in Section 5.3 indicate that constraints in dealflow generation cannot fully explain why VCs select different firms than the algorithm, or why algorithm-selected firms perform better than VC-backed firms.

In this section, we explore another (non-mutually exclusive) explanation based on the observation that when making investment decisions, VCs may rely on heuristics that arise in probability judgements and in the context of prediction problems (see Kahneman, 2011; Bordalo et al., 2016). For example, the founder's identity has been shown to be a first order determinant of VCs' investment decisions (Bernstein, Korteweg and Laws, 2017; Gompers et al., 2020). If such heuristics make VCs more likely to pass up certain kinds of promising ventures, they could help explain our results.

7.1 Predicting VCs' decisions

To better understand why VCs' decisions differ from the algorithmic investment policy, we develop a separate estimator, denoted $\widehat{h}(\cdot)$, that predicts for each firm 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 tested out-of-sample on the remaining 30% of observations.[21]

[21] We exclude the 2006 cohort in this test because our prediction exercise is to predict which firms are VC-backed, but the 2006 entrepreneur survey does not allow us to identify VC-backed status. We use a random split for this exercise for two reasons. The first is technical and due to the limited number of firms that are VC-backed in these three cohorts. The second reason is that we are not comparing VCs' and algorithmic selections in this exercise. We thus do not need to ensure a level-playing field for the algorithm against VCs, where both would observe the performance of past new firms. 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 reasonably well. Figure 8 shows that our model has an area under the curve (AUC) of .77. This 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 indeed is VC-backed, is 77%.

[Insert Figure 8 here]

One striking result is that if restricted to three founder demographic features, our predictive model of VCs' decisions produces an AUC of .62. This implies that a lot of the signal to predict VCs' decisions is captured by these three demographic features. We view this result as indirect evidence that VCs operate under bounded rationality as they appear to rely on a sparse model to make investment decisions. In contrast, when the estimator of firm performance $\widehat{m}(\cdot)$ takes only these three features as its input, the algorithmic policy's performance 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. To further our understanding of VCs' firm selection, we follow the approach in Ludwig and Mullainathan (2021) and test in a regression framework whether there exist factors beyond predicted performance that can predict VCs' decisions. We first regress VCs' actual decisions, *VC-backed*$_i$, on our algorithmic predictions of VCs' decisions:

$$\textit{VC-backed}_i = \beta_0 + \widehat{h}(X_i)\beta_1 + \epsilon_i \tag{5}$$

Table 7 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 firm in the third quartile of our VC-backed predictions is 1.1 p.p. more likely to be VC-backed compared to a firm in the first quartile, a 127% increase relative to the mean.[22] 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):

$$\textit{VC-backed}_i = \beta_0 + \widehat{m}(X_i)\beta_1 + \epsilon_i \tag{6}$$

[22] There are 26,098 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 did not care about portfolio firms' revenue, we would expect revenue predictions to not load significantly ($\beta_1 = 0$ in Equation (6)). This is not the case: Column 2 shows that predictions for being in the top of the cohort in terms of revenue correlates with VC-backed status. A firm in the third quartile of our best performer predictions is 0.26 p.p. more likely to be VC-backed compared to a firm in the first quartile, a 29% increase relative to the mean. Column 3 shows that successful exit predictions also correlate with VCs' decisions. Next, we test whether our predictions of VCs' decisions remain significant once we control for performance predictions by estimating:

$$\textit{VC-backed}_i = \beta_0 + \widehat{h}(X_i)\beta_1 + \widehat{m}(X_i)\beta_2 + \epsilon_i \tag{7}$$

Columns 4 through 6 of Table 7 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 6 where we add venture performance predictions. This result implies that our model predicting VCs' behavior picks up signal 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 7 here]

Which entrepreneurs are more likely to be casting errors? We compare the characteristics of entrepreneurs who have low predicted performance but high chances of being VC-backed (low $\widehat{m}(x_i)$ and high $\widehat{h}(x_i)$) to those who have high predicted performance but low chances of being VC-backed (high $\widehat{m}(x_i)$ and low $\widehat{h}(x_i)$). This comparison allows us to identify the profile of entrepreneurs who are more likely to be "casting errors."

First, we sort firms into quintiles according to their predicted performance ($\widehat{m}(x_i)$) and their predicted chance of being VC-backed ($\widehat{h}(x_i)$). Second, we keep firms that are in the first and fifth quintiles of these distributions and create two groups of firms: one group containing those firms both in the first quintile of $\widehat{m}(x_i)$ *and* the fifth quintile of $\widehat{h}(x_i)$, and one group containing those firms both in the fifth quintile of $\widehat{m}(x_i)$ *and* the first quintile of $\widehat{h}(x_i)$. Third, we run a t-test of the difference in characteristics between these two groups. Table 8 contains the results.

[Insert Table 8 here]

The results in Table 8 suggest that entrepreneurs who are more likely to be casting errors are male, are based in Paris and whose venture is in the tech industry.

7.2 Stereotypical thinking by VCs

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" (Lerner and Nanda, 2020). The evidence in Sections 7.1 and 6.2 motivates our exploration of bias: Figure 8 shows that a large part of the signal that predicts VCs' investment decisions lies in three demographic features, and Table 5 suggests that some representative features of successful entrepreneurs are over-represented among VC-backed firms. In this section, we ask whether VCs exaggerate the representative features of successful entrepreneurs in their selection of firms.

To assess whether VCs' emphasis on certain features is efficient, we follow the approach in Mullainathan and Obermeyer (2022) and create simple models $\widehat{m}_{simple}(\cdot)$ to predict whether a firm will be among the best performers of its cohort. In these simple models, the only departure from the estimator $\widehat{m}(\cdot)$ is that we restrict the set of input features to variables that have drawn most attention from the existing literature.[23] We regress VCs' decisions on our full model predicting which firms are most likely to be among the best performers, as well as our simple models:

$$\textit{VC-backed}_i = \beta_0 + \widehat{m}(X_i)\beta_1 + \widehat{m}_{simple}(X_i)\beta_2 + \epsilon_i \tag{8}$$

Under the null hypothesis, $\beta_2 = 0$, so that the variables used in $\widehat{m}_{simple}(\cdot)$ do not matter for VCs' decisions over and above their effect on firms' performance. Alternatively, $\beta_2 \neq 0$ would imply

$$\frac{\mathrm{Cov}(M_{\widehat{m}}\,\textit{VC-backed}, M_{\widehat{m}}\widehat{m}_{simple})}{\mathrm{Var}(M_{\widehat{m}}\,\textit{VC-backed})} \neq 0, \tag{9}$$

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). Intuitively, $\beta_2 \neq 0$ implies that the variables used in $\widehat{m}_{simple}(x_i)$ contain signal to predict VCs' decisions beyond their effect on predicted performance $\widehat{m}(x_i)$.

We interpret the sign of the coefficient β_2 as in Mullainathan and Obermeyer (2022). If the

[23] 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. Given our findings in Section 6.2 that VCs select entrepreneurs representative of the best performing firms, we train the estimators $\widehat{m}(\cdot)$ and $\widehat{m}_{simple}(\cdot)$ to predict $top5rev_5$, a dummy variable equal to one for firms in the top 5% of their cohort's in terms of revenue at age 5. We report results using observations in our test set, the 2010 cohort of entrepreneurs.

coefficient β_2 on a simple model's prediction is positive, the covariance term in (9) is positive. In this case, the simple model's features affect VCs' probability to back a firm ($\textit{VC-backed}_i$) in the same direction as they affect the firm's predicted performance ($\widehat{m}(x_i)$). In the words of Mullainathan and Obermeyer (2022), VCs *overweight* the features used in a simple model when β_2 is positive, that is, they exaggerate the signal contained in these features to predict performance. Instead, if the coefficient β_2 is negative, the covariance term in (9) is negative and in that case, we say that VCs *underweight* the features used in the simple model.

Panel A of Table 9 contains the results of Equation (8) for several simple models that take entrepreneur features as inputs. Since potential investors are highly responsive to information about the founding team (Bernstein, Korteweg and Laws, 2017; Gompers et al., 2020), our first simple 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 our VC-backed variable on $\widehat{m}(x_i)$, our full estimator that predicts whether a firm will be among the best performers of its cohort ($top5rev_5$). 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 decisions. The interquartile range of $\widehat{m}_{simple}(\text{personal features}_i)$ is 0.0291 , translating to a shift of about 0.12 p.p. in the probability of being VC-backed, which represents an 38% increase relative to the baseline average.[24] In Columns 3 to 8, we test other simple models focusing on one personal characteristic in isolation.

[Insert Table 9 here]

We find that VCs exaggerate several features that are representative of the most successful entrepreneurs. First, column 4 shows that VCs overweight the entrepreneur's gender in their decision to back a firm. Although female entrepreneurs perform worse than male entrepreneurs on average (Figure 1), we find that controlling for performance predictions, female entrepreneurs are 0.21 p.p. less likely to be VC-backed than they would if VCs' decisions were solely based on the effect of gender on firm performance, a 66% decrease 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, 2020; Calder-Wang and Gompers, 2021).

Second, VCs exaggerate the entrepreneur's education in their decisions (see, e.g., Queiró, 2021,

[24] Approximately 0.3% of firms are VC-backed in our test set.

on the importance of education in new firms' 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 to receive VC-backing to a greater extent than justified by its effect on predicted performance. Column 6 shows that VCs overweight the "Grande Ecole" feature, which is a dummy equal to one if the entrepreneur graduated from an elite French school.[25] We find that VCs are more than three times more likely to back an entrepreneur who graduated from a Grande Ecole, even when controlling for performance predictions.

Third, VCs overweight optimism and past entrepreneurial experience. In Column 9, the simple model uses features that capture the entrepreneurs' optimism: whether they want to grow and whether they expect to hire over the next 12 months. In Column 10, we find that serial entrepreneurs are also more likely to be VC-backed than if VCs' decisions were solely based on expected performance.

We do not find evidence that VCs exaggerate the entrepreneur's age in their decisions (column 3), or that they are biased towards the entrepreneur's nationality (column 7) or her family's entrepreneurial background (column 8).

Fourth, Panel B of Table 9 shows that VCs overweight the new ventures' location and industry. Column 2 shows that Paris-based firms are on average 0.40 p.p. more likely to be VC-backed compared to what performance would imply, which represents a 126% increase relative to the mean. In contrast, we do not find evidence that VCs either overweight or underweight Marseille-, Lyon- or Bordeaux-based firms (columns 3 to 5). In Columns 6 through 8, we focus on industries that are most VC-backed in our data. We find that VCs overweight firms in the high tech industry, which are four times more likely to be VC-backed when controlling for performance.

Finally, in Column 9, we focus on proxies for the venture's traction: the total number of workers, the number of clients, and the clients' location. Consistent with Bernstein, Korteweg and Laws (2017), we do not find evidence that VCs exaggerate new firms' traction in their decisions.

8 Conclusion

This paper uses machine learning to study how venture capitalists (VCs) make investment decisions. Our approach is to contrast VCs' decisions to an algorithmic policy that selects the most promising new ventures based on predictions of operating performance. We find that VCs invest in some firms

[25] The Grande Ecole 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.

that perform predictably poorly and pass on others that perform predictably well. This approach does not rely on the assumption that algorithmic predictions are correct. Rather, we use these predictions to isolate potential errors by VCs and we rely on realized outcomes to evaluate actual errors. The interpretation of our results is facilitated by the representativeness and completeness of our data, which include both VC-backed and non-VC-backed firms, circumventing selection issues that are prevalent in both the venture capital and the machine learning literature.

We estimate the shadow cost of the constraints faced by VCs by comparing the performance of the (unconstrained) algorithmic policy to that of algorithmic policies constrained to selecting firms similar to VC-backed ones. Even our most constrained algorithm significantly outperforms VC-backed firms, which implies that VCs' constraints cannot fully explain the higher performance of algorithm-selected firms compared to VC-backed firms.

To understand why VCs do not select the most promising entrepreneurs, we use an algorithmic model that predicts for each new firm whether it is VC-backed. One striking result is that almost half of the predictable component of VCs' decisions can be attributed to three founder demographics (gender, age, education). Consistent with stereotypical thinking (Tversky and Kahneman, 1974; Bordalo et al., 2016), we show that VCs are more likely to back firms whose characteristics are representative of the most successful entrepreneurs (i.e., characteristics that occur more frequently among the best performing entrepreneurs relative to the other ones). We follow the approach in Mullainathan and Obermeyer (2022) and create simple models that predict VCs' decisions based on these representative features. We find that VCs exaggerate some representative features of success in their decisions (e.g., male, highly educated, Paris-based, and high-tech entrepreneurs). True, entrepreneurs with these characteristics have better chances of becoming the very best performers of their cohort, but representativeness exaggerates these features and induces VCs to neglect predictably good performers with different features.

Figures and Tables

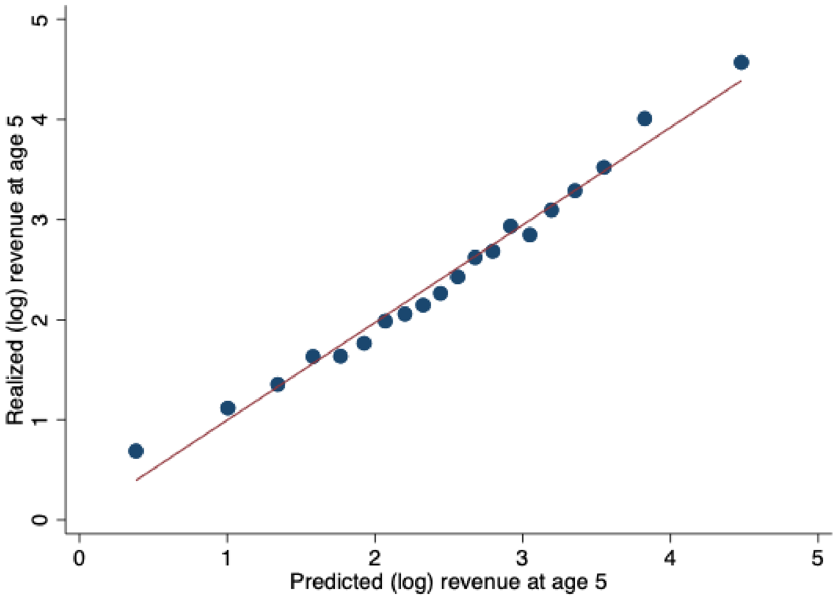

Figure 1: Algorithm Performance: All New Firms in Test Set. This figure shows the average observed performance (y-axis) across 20 bins of predicted performance (x-axis) among the 37,823 new firms in the 2010 test set. The performance measure is the log revenue at age 5. The predictive model was trained using 10-fold cross validation on the sample of all firms in the 1998, 2002 and 2006 cohorts.

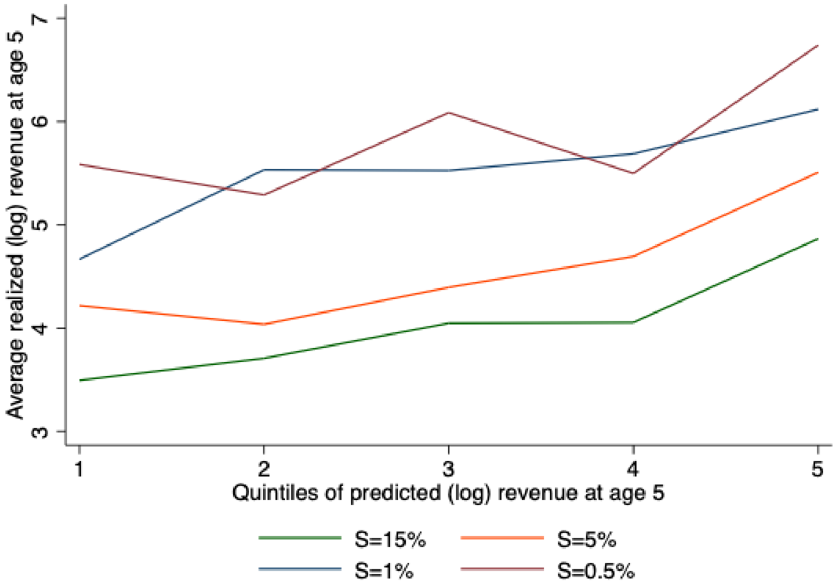

Figure 2: Algorithm Performance: Algorithm-selected New Firms in Test Set for Various Selectivity Thresholds ($S = 1-t$). This figure shows the average observed performance (y-axis) across five quintiles of predicted performance (x-axis) for the set of firms selected by the algorithmic policy under various selectivity thresholds. The performance measure is the log revenue at age 5. The predictive model was trained using 10-fold cross validation on the sample of all firms in the 1998, 2002 and 2006 cohorts.

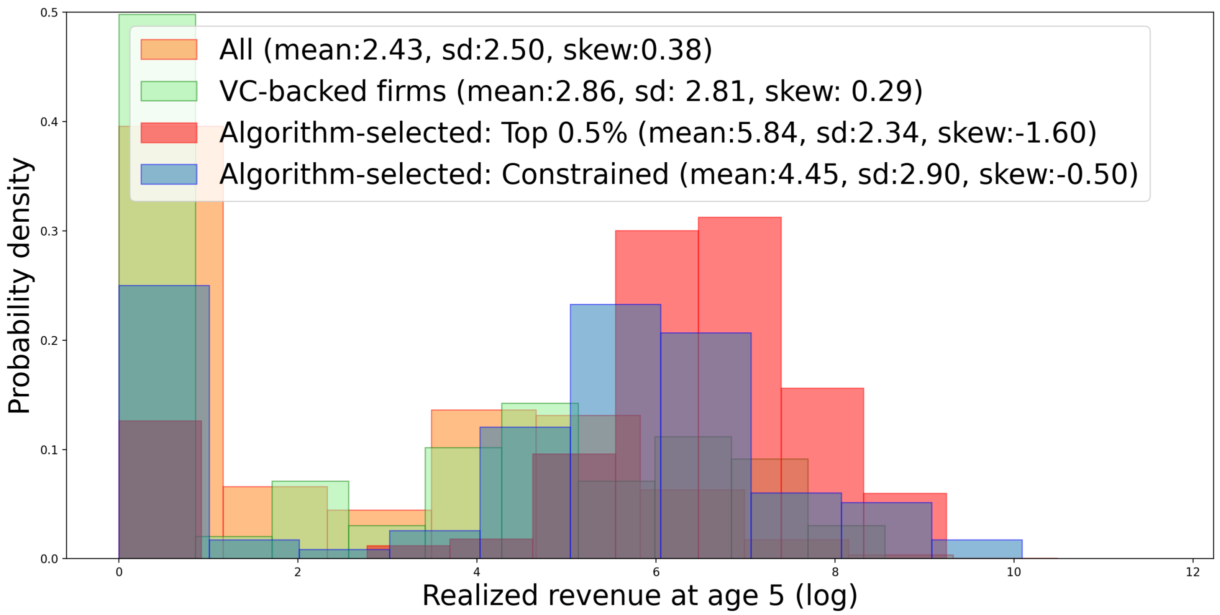

Figure 3: Realized Performance of Ventures in Test Set when the Algorithm Predicts the log of Revenue at Age 5. This figure shows the distribution of firm performance for all firms in the 2010 cohort (our test set in orange) as well as the breakdown for VC-backed firms (in green), for algorithm-selected firms using the unconstrained policy at the $s = 0.5\%$ threshold (in red), and for algorithm-selected firms selected by the policy restricted to investing in firms that match the industry and the location of VC-backed firms (in blue). The predictive model is trained on the sample of all new firms in the 1998, 2002 and 2006 cohorts using 10-fold cross validation. We report the mean, standard deviation and skewness of revenue at age 5 (log) for each group.

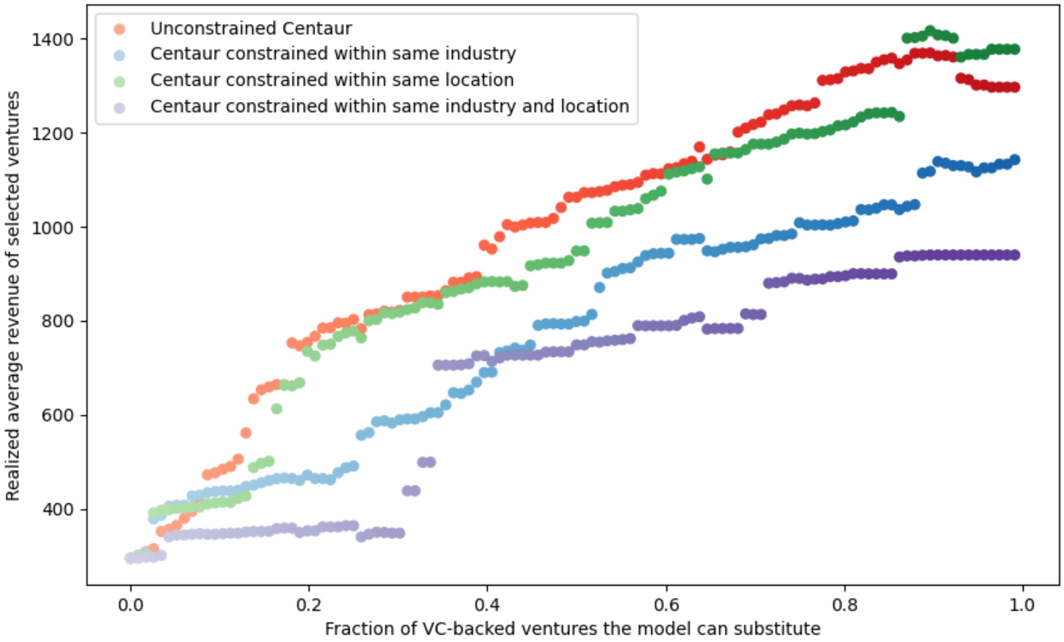

Figure 4: Potential Performance Gains: Counterfactual Models. This figure shows the average realized revenue at age 5 for several counterfactual models that replace VC-backed firms that are predicted to become poor performers with firms that are predicted to become good performers by the algorithm. The origin represents the status quo: it includes the full set of VC-backed firms in the test set and their observed average performance at age 5. The red line shows the performance of the unconstrained counterfactual model. Each line below it represents the performance of a counterfactual model constrained to replace VC-backed firms with firms that are in the same industry (in blue), the same location (in green), or the same industry *and* location (in purple).

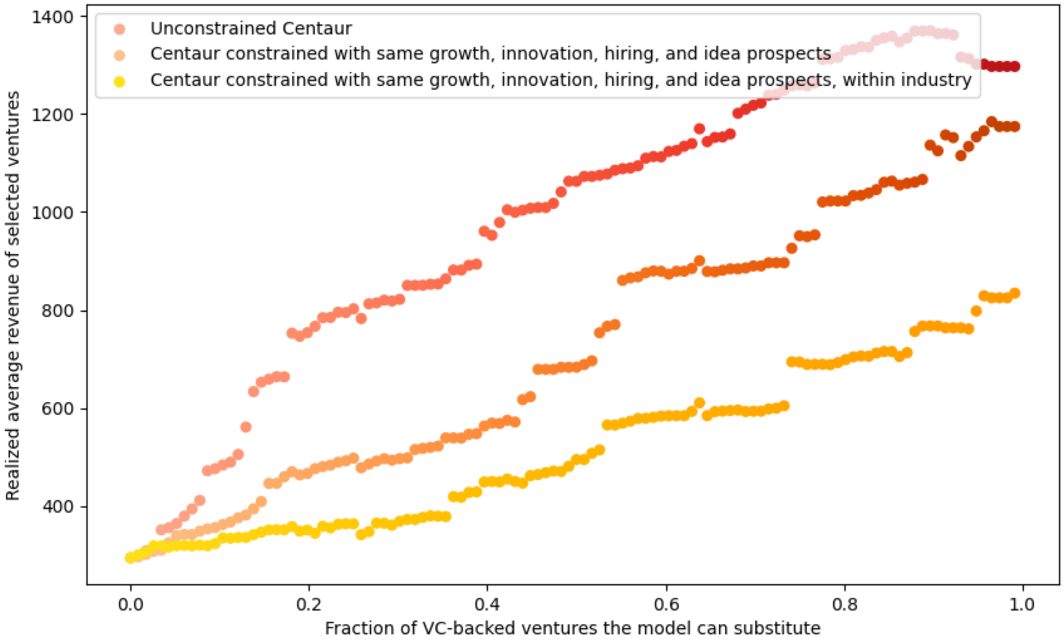

Figure 5: Counterfactual Models Restricted to VC-prone Ventures. This figure shows the average realized performance at age 5 for several counterfactual models that replace VC-backed firms that are predicted to become poor performers with firms that are predicted to become good performers by the algorithm. The origin represents the status quo: it includes the full set of VC-backed firms in the test set and their observed average performance at age 5. The red line shows the performance of the unconstrained counterfactual model. The orange line represents the performance of a counterfactual model constrained to replace VC-backed firms with firms founded by an entrepreneur whose responses to growth related questions in the entrepreneur survey match those of the entrepreneur whose firm was dropped by the counterfactual model (same growth prospects, expectation to hire, innovate and motivated by a new idea). The purple line further restricts the counterfactual model to selecting a firm in the same industry as the firm it drops.

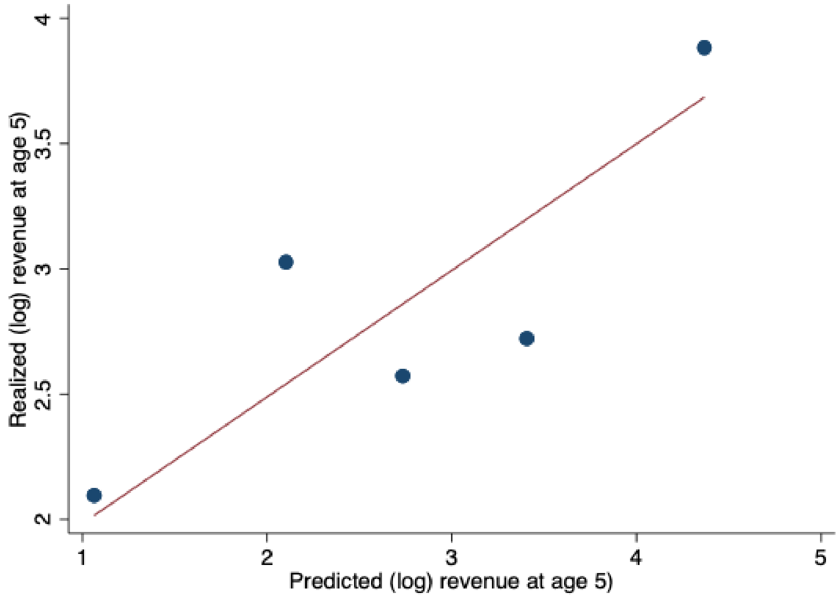

Figure 6: Algorithm Performance: VC-backed Firms in Test Set. This figure shows the average observed performance (y-axis) across 5 bins of predicted performance (x-axis) for the VC-backed firms in our 2010 cohort (our test set). The predictive model is trained on the sample of all new firms in the 1998, 2002, and 2006 cohorts.

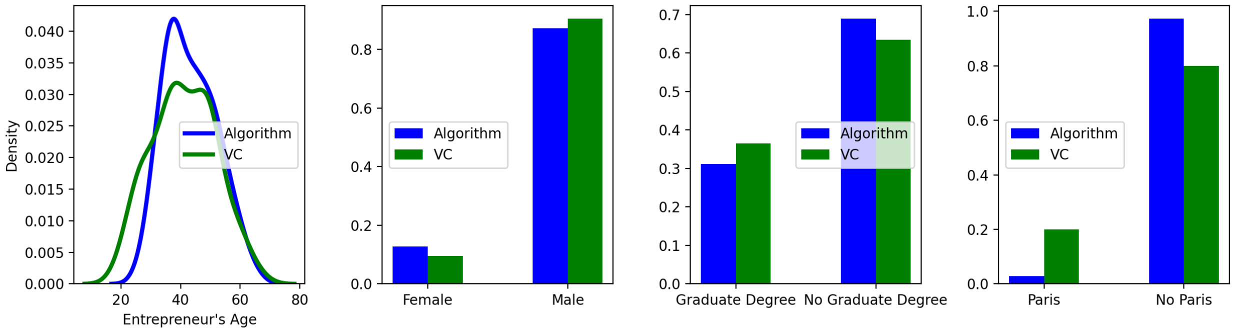

Figure 7: Entrepreneur Demographics for VC-backed and Algorithm-selected Ventures. This figure shows the probability densities of founders' ages as well as the breakdown of entrepreneurs' gender, education level and geographic location in the 2010 cohort (our test set) for VC-backed and algorithm-selected firms at the $s = 0.5\%$ threshold. The predictive model is trained on the sample of all new firms in the 1998, 2002 and 2006 cohorts using 10-fold cross validation.

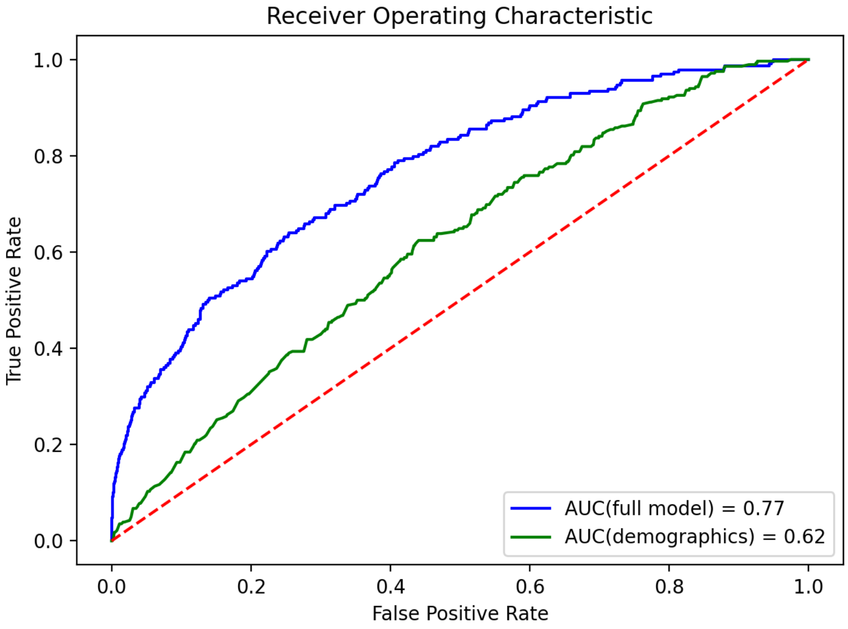

Figure 8: Area Under the Curve (AUC) of a Predictive Model of VCs' Decisions. This figure presents the AUC of a predictive model of VCs' decisions. The AUC of .77 for the full model implies that for two randomly picked ventures, one VC-backed and one not, the odds that our model assigns a higher probability 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).

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

Table 1: Correspondence Between Revenue and Valuation at Exit. This table reports the correspondence between VC-backed firms' revenue and valuation, both at exit. Rows 1, 2, and 3 focus on firms in the top 1%, 5%, and 10% of the revenue distribution, respectively. For each set of firms, 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 350 French firms and 7,593 US firms.

Table 2: 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 four-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 C 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)	2.31	2.46	2.20	5.67	7.70	85,171	2.43	2.50	2.08	5.81	7.69	35,928
	Revenue at Age 5	164.29	1,816.30	8.00	290.40	2,205.00	85,171	165.14	1,193.23	7.03	332.26	2,175.07	35,928
	Alive at Age 5	0.62	0.48	1.00	1.00	1.00	85,171	0.65	0.48	1.00	1.00	1.00	35,928
Demographics													
	Entrepreneur's Age	37.79	10.13	37.00	52.00	64.00	83,032	39.78	10.65	39.00	54.00	66.00	35,928
	Female	0.29	0.45	0.00	1.00	1.00	83,042	0.28	0.45	0.00	1.00	1.00	35,928
	Entrepreneur's Nationality (FR)	0.87	0.34	1.00	1.00	1.00	85,171	0.92	0.27	1.00	1.00	1.00	35,928
	Entrepreneurial Family	0.68	0.47	1.00	1.00	1.00	81,297	0.70	0.46	1.00	1.00	1.00	35,928
Professional Background													
	Self-employed	0.37	0.48	0.00	1.00	1.00	85,171	0.31	0.46	0.00	1.00	1.00	35,928
	Previously Employed	0.47	0.50	0.00	1.00	1.00	85,171	0.55	0.50	1.00	1.00	1.00	35,928
	Part-time Entrepreneur	0.19	0.40	0.00	1.00	1.00	81,051	0.21	0.41	0.00	1.00	1.00	35,928
	Same Prior Industry	0.52	0.50	1.00	1.00	1.00	81,258	0.60	0.49	1.00	1.00	1.00	35,928
	Serial Entrepreneur	0.27	0.45	0.00	1.00	1.00	85,171	0.29	0.45	0.00	1.00	1.00	35,928
	Previously Employed in Small Firm	0.40	0.49	0.00	1.00	1.00	85,171	0.42	0.49	0.00	1.00	1.00	35,928
	Previously Inactive	0.10	0.30	0.00	0.00	1.00	85,171	0.00	0.00	0.00	0.00	0.00	35,928
	Below High School Degree	0.44	0.50	0.00	1.00	1.00	85,171	0.41	0.49	0.00	1.00	1.00	35,928
	Undergraduate Degree	0.16	0.37	0.00	1.00	1.00	85,171	0.26	0.44	0.00	1.00	1.00	35,928
	Graduate Degree	0.14	0.34	0.00	1.00	1.00	85,171	0.15	0.35	0.00	1.00	1.00	35,928
	Grande Ecole	0.04	0.21	0.00	0.00	1.00	22,034	0.06	0.24	0.00	0.00	1.00	34,995
	Completed Required Training	0.21	0.40	0.00	1.00	1.00	85,171	0.22	0.41	0.00	1.00	1.00	35,928
Motivation and Expectations													
	Expectation: Growth	0.53	0.50	1.00	1.00	1.00	85,171	0.42	0.49	0.00	1.00	1.00	35,928
	Expectation: Sustain	0.27	0.45	0.00	1.00	1.00	85,171	0.39	0.49	0.00	1.00	1.00	35,928
	Expectation: Rebound	0.07	0.25	0.00	0.00	1.00	85,171	0.08	0.28	0.00	0.00	1.00	35,928
	Motivation: Peer Entrepreneurs	0.11	0.31	0.00	1.00	1.00	83,225	0.09	0.28	0.00	0.00	1.00	35,928
	Expect to Hire	0.24	0.43	0.00	1.00	1.00	85,171	0.26	0.44	0.00	1.00	1.00	35,928
	Motivation: New Idea	0.18	0.39	0.00	1.00	1.00	83,225	0.16	0.37	0.00	1.00	1.00	35,928
	Motivation: Opportunity	0.33	0.47	0.00	1.00	1.00	83,225	0.43	0.50	0.00	1.00	1.00	35,928
	Innovation	0.39	0.49	0.00	1.00	1.00	85,171	0.48	0.50	0.00	1.00	1.00	35,928

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	85,171	0.09	0.28	0.00	0.00	1.00	35,928
	Marseille-based	0.02	0.14	0.00	0.00	1.00	85,171	0.03	0.18	0.00	0.00	1.00	35,928
	Lyon-based	0.02	0.13	0.00	0.00	1.00	85,171	0.02	0.13	0.00	0.00	1.00	35,928
	Bordeaux-based	0.02	0.14	0.00	0.00	1.00	85,171	0.02	0.13	0.00	0.00	1.00	35,928
	Business Services Industry	0.16	0.36	0.00	1.00	1.00	85,171	0.15	0.35	0.00	1.00	1.00	35,928
	Health and Social Work Industry	0.04	0.20	0.00	0.00	1.00	85,171	0.04	0.20	0.00	0.00	1.00	35,928
	Construction Industry	0.18	0.39	0.00	1.00	1.00	85,171	0.18	0.38	0.00	1.00	1.00	35,928
	High tech Industry	0.01	0.12	0.00	0.00	1.00	85,171	0.02	0.14	0.00	0.00	1.00	35,928
	Energy Industry	0.00	0.02	0.00	0.00	0.00	85,171	0.03	0.16	0.00	0.00	1.00	35,928
	B2B	0.33	0.47	0.00	1.00	1.00	85,171	0.33	0.47	0.00	1.00	1.00	35,928
	B2C	0.58	0.49	1.00	1.00	1.00	85,171	0.58	0.49	1.00	1.00	1.00	35,928
	International Customers	0.07	0.25	0.00	0.00	1.00	85,171	0.05	0.22	0.00	0.00	1.00	35,928
	Local Customers	0.48	0.50	0.00	1.00	1.00	85,171	0.54	0.50	1.00	1.00	1.00	35,928
	Domestic Customers	0.14	0.35	0.00	1.00	1.00	85,171	0.16	0.36	0.00	1.00	1.00	35,928
Venture Organization													
	Co-founders	0.12	0.32	0.00	1.00	1.00	85,171	0.14	0.35	0.00	1.00	1.00	35,928
	Outsourcing: Accounting	0.63	0.48	1.00	1.00	1.00	82,518	0.74	0.44	1.00	1.00	1.00	34,623
	Number of Employees	1.59	1.52	1.00	3.00	8.00	85,171	1.64	1.61	1.00	3.00	9.00	34,623
	10+ Clients	0.58	0.49	1.00	1.00	1.00	85,171	0.59	0.49	1.00	1.00	1.00	35,928
	Number of Paid Managers	0.15	0.46	0.00	1.00	2.00	85,171	0.18	0.43	0.00	1.00	2.00	34,623
	Customers from Prior Job	0.30	0.46	0.00	1.00	1.00	85,171	0.26	0.44	0.00	1.00	1.00	35,928
	Suppliers from Prior Job	0.22	0.42	0.00	1.00	1.00	85,171	0.21	0.41	0.00	1.00	1.00	35,928
	Help from Professionals	0.03	0.18	0.00	0.00	1.00	85,171	0.10	0.30	0.00	0.00	1.00	35,928
	Help from Family	0.27	0.44	0.00	1.00	1.00	85,171	0.18	0.38	0.00	1.00	1.00	35,928
	No External Help	0.41	0.49	0.00	1.00	1.00	85,171	0.27	0.44	0.00	1.00	1.00	35,928
Financial Characteristics													
(not included as input features)	Total Assets	144.68	4,532.02	29.00	183.00	1,157.00	56,935	268.18	1,362.92	42.90	435.58	3,261.47	35,461
	Bank Loan	0.35	0.48	0.00	1.00	1.00	84,000	0.41	0.49	0.00	1.00	1.00	35,928
	Other Loan	0.08	0.27	0.00	0.00	1.00	84,000	0.10	0.29	0.00	0.00	1.00	35,928
	No Outside Financing	0.54	0.50	1.00	1.00	1.00	84,000	0.52	0.50	1.00	1.00	1.00	35,928
	Other Firm Financing	0.05	0.21	0.00	0.00	1.00	51,062	0.04	0.19	0.00	0.00	1.00	35,928
	Grant	0.21	0.41	0.00	1.00	1.00	84,000	0.08	0.27	0.00	0.00	1.00	35,928
	Future VC Financing	0.01	0.11	0.00	0.00	1.00	85,171	0.02	0.16	0.00	0.00	1.00	35,928
Industries-Locations													
	Number of Industries	-	-	-	-	-	37	-	-	-	-	-	37
	Number of Regions	-	-	-	-	-	321	-	-	-	-	-	321

Algorithm trained on	Algorithm evaluated on						
	Revenue$_5$ (log)	Revenue$_7$ (log)	Top 5% Revenue$_5$	Top 5% Revenue$_7$	Imputed Valuation	Revenue Growth	Successful Deals
Revenue_5 (log)	5.95	5.50	0.54	0.51	5012.35	0.12	2
Revenue_7 (log)	6.08	5.65	0.53	0.50	5702.83	0.15	3
Top 5% Revenue_5	5.67	5.02	0.59	0.54	6063.83	0.09	3
Top 5% Revenue_7	5.68	5.08	0.60	0.54	6969.52	0.10	3
Successful Deals	3.95	3.51	0.23	0.24	3243.05	0.03	10
	Comparison: Average performance measures						
	Revenue$_5$ (log)	Revenue$_7$ (log)	Top 5% Revenue$_5$	Top 5% Revenue$_7$	Imputed Valuation	Revenue Growth	Successful Deals
All firms in test set	2.43	2.02	0.05	0.05	745.75	-0.05	52
VC-backed firms	2.86	2.48	0.15	0.17	1498.34	0.01	4

Table 3: Performance of Algorithmic Policy Using Various Measures of Firm Performance. This table reports the average observed outcome for algorithm-selected firms at the $s = 0.5\%$ threshold for different predictive models that predict various measures of firm success. In the first panel, we train the algorithm to predict various outcome measures: the new firm's revenue in 5 and 7 years , whether the firm will be in the top 5% of its cohort in terms of revenue in 5 and 7 years, and whether the firm exits through an acquisition or IPO or receives VC funding in later years). For comparison, in the second panel we show the mean of each performance measure for the 2010 cohort (test set) in row 1, and for VC-backed firms only in row 2.

		Test Set										
	variable		VC-backed		Algorithm-selected ($s = 0.5\%$)			Algorithm-selected ($s = 1\%$)			Difference ($s = 1\%$)	Difference ($s = 0.5\%$)
		Mean	SD	N	Mean	SD	N	Mean	SD	N	T-Test	T-Test
Outcomes												
	Revenue at Age 5 (log)	2.86	2.83	115	5.84	2.35	180	5.51	2.53	360	-2.65***	-2.98***
	Revenue at Age 5	293.61	699.37	115	1138.23	1603.37	180	1066.94	1910.88	360	-773.33***	-844.63***
	Alive at Age 5	0.69	0.47	115	0.91	0.29	180	0.89	0.31	360	-0.21***	-0.22***
Founder Demographics												
	Entrepreneur's Age	41.11	10.59	115	43.32	8.73	180	42.57	9.08	360	-1.46	-2.20*
	Founder's Nationality (FR)	0.94	0.24	115	0.99	0.11	180	0.98	0.14	360	-0.04*	-0.05**
	Female	0.10	0.30	115	0.13	0.33	180	0.13	0.34	360	-0.03	-0.03
Founder Professional Background												
	Same Prior Industry	0.52	0.50	115	0.90	0.30	180	0.87	0.33	360	-0.35***	-0.38***
	Serial Entrepreneur	0.40	0.49	115	0.41	0.49	180	0.38	0.49	360	0.02	-0.01
	Previously Employed in Small Firm	0.29	0.45	115	0.33	0.47	180	0.38	0.49	360	-0.09*	-0.04
	Graduate Degree	0.37	0.48	115	0.31	0.46	180	0.29	0.45	360	0.08	0.05
	Grande École	0.26	0.44	115	0.06	0.24	179.00	0.09	0.28	358.00	0.17***	0.20***
Founder Motivation and Expectations												
	Expectation: Growth	0.58	0.50	115	0.55	0.50	180	0.61	0.49	360	-0.03	0.03
	Motivation: Successful Peer Entrepreneurs	0.06	0.24	115	0.09	0.29	180	0.10	0.30	360	-0.04	-0.03
	Expect to Hire	0.50	0.50	115	0.59	0.49	180	0.61	0.49	360	-0.11**	-0.09
	Motivation: New Idea	0.39	0.49	115	0.09	0.29	180	0.12	0.32	360	0.27***	0.30***
	Motivation: Opportunity	0.37	0.48	115	0.57	0.50	180	0.57	0.50	360	-0.21***	-0.20***
	Innovation	0.73	0.45	115	0.43	0.50	180	0.47	0.50	360	0.26***	0.30***
Venture Characteristics												
	Paris-based	0.20	0.40	115	0.03	0.16	180	0.03	0.17	360	0.17***	0.17***
	High tech industry	0.10	0.30	115	0.01	0.11	180	0.01	0.12	360	0.08***	0.08***
Organization												
	Outsourcing: Accounting	0.90	0.30	109.00	0.92	0.27	180	0.93	0.26	360	-0.03	-0.02
	Outsourcing: Management	0.09	0.29	109.00	0.23	0.42	180	0.20	0.40	360	-0.11***	-0.14***
	Outsourcing: Logistics	0.15	0.36	109.00	0.34	0.47	180	0.31	0.46	360	-0.16***	-0.19***
	Number of Employees	2.40	2.93	109.00	6.59	4.89	180	5.69	4.45	360	-3.29***	-4.19***
Industries-Locations												
	Number of Industries	-	-	26	-	-	24	-	-	28	-	-
	Number of Regions	-	-	62	-	-	97	-	-	137	-	-
Financial Characteristics (not included in input features)	Total Assets (k euros)	527.29	1985.20	114.00	684.65	2228.96	179.00	592.01	2037.95	358.00	-64.72	-157.36

Table 4: Differences Between VC-backed and Algorithm-selected New Ventures. This table reports selected summary statistics for VC-backed and algorithm-selected firms at the $s = 0.5\%$ and $s = 1\%$ thresholds. We report t-tests for the difference in means between VC-backed and algorithm-selected firms. 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). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Feature	Top 5%	Bottom 95%	Representativeness of best performers $\frac{Pr(X_i \mid \text{Top5})}{Pr(X_i \mid \text{Bottom95})}$
	(1)	(2)	(3)
Male	85%	72%	1.17
Graduate Degree	18%	15%	1.19
Grande Ecole	8%	5%	1.59
Optimism	50%	20%	2.50
Serial Entrepreneur	42%	23%	1.88
Paris-based	15%	8%	1.80
High tech	1%	1%	1.05

Table 5: Stereotypes of the Most Successful Entrepreneurs. This table reports the fraction of entrepreneurs with a given characteristic among the best performing firms (top 5% of revenue at age 5, column 1) and among the other firms (bottom 95% of revenue at age 5, column 2). A given characteristic is representative (or stereotypical) of the best performing firms if it scores high on the representativeness ratio (column 3) of the percentage in column 1 over that in column 2. Revenue at age 5 for the calculations in columns 1 and 2 is measured in the training sample, and representativeness of VC-backed firms in column 3 is measured in the test set.

Feature	Top 5%	Bottom 95%	Representativeness	Top 1%	Bottom 99%	Representativeness	Representativeness
			of best performers			of best performers	*among* VC-backed firms
			$\frac{Pr(X_i \mid \text{Top5})}{Pr(X_i \mid \text{Bottom95})}$			$\frac{Pr(X_i \mid \text{Top5})}{Pr(X_i \mid \text{Bottom95})}$	$Pr(X_i \mid \text{VC-backed})$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VC Hub	65.6%	61.8%	1.06	76.7%	61.8%	1.24	65%
California	44.6%	38.8%	1.15	60.2%	38.8%	1.55	39%
Massachusetts	8.4%	8.9%	0.94	5.6%	8.9%	0.63	8.89%
New York	7.4%	7.3%	1.01	6.0%	7.3%	0.82	7.32%
Texas	5.3%	6.8%	0.78	4.9%	6.7%	0.73	3%
Largest Industries	79.1%	75.5%	1.05	79.3%	75.7%	1.05	89.8%
Information Technology	49.7%	44.6%	1.11	61.7%	44.7%	1.38	44.8%
Health Care	19.9%	21.6%	0.92	8.6%	21.6%	0.40	21.5%
Consumer Discretionary	9.5%	9.4%	1.01	9.0%	9.4%	0.96	9.4%
Industrials	6.1%	7.7%	0.79	2.6%	7.7%	0.34	7.6%
Communication	6.5%	6.5%	1.00	9.4%	6.5%	1.45	6.5%

Table 6: Stereotypes of the Most Successful Entrepreneurs in U.S. Burgiss Data (Using TVPI). This table reports the fraction of entrepreneurs with a given characteristic among the best performing firms and among the other firms. The two deal characteristics available in the Burgiss data are the firm 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 firms are in the top 5%, and the other firms in the bottom 95%, in terms of TVPI. In columns 4 and 5, the best performing firms are in the top 1%, and the other firms in the bottom 99%, in terms of TVPI. A given characteristic is representative (or stereotypical) of the best performing firms 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.

	VC-backed					
	(1)	(2)	(3)	(4)	(5)	(6)
$\widehat{h}(X)$	1.689***			1.720***	1.774***	1.783***
	(0.0600)			(0.0639)	(0.0626)	(0.0654)
$\widehat{m}(X)_{top5va_5}$		0.0528***		-0.00967		-0.00352
		(0.00646)		(0.00678)		(0.00691)
$\widehat{m}(X)_{homerun}$			0.354***		-0.499***	-0.489***
			(0.103)		(0.105)	(0.107)
Observations	26,098	26,098	26,098	26,098	26,098	26,098
R-squared	0.029	0.003	0.000	0.030	0.030	0.030

Standard errors in parentheses

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

Table 7: VCs' Decision Model is not Subsumed by Performance or Home Run 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 three estimators. $\widehat{h}(X)$ is a vector of predicted probabilities for whether a firm is VC-backed; $\widehat{m}(X)_{top5rev_5}$ is a vector of predicted probabilities for whether a firm will be in the top 5% of its cohort in terms of revenue at age 5; $\widehat{m}(X)_{homerun}$ is a vector of predicted probabilities for whether a firm will be a home run. All three estimators are built using a random 70/30 split using the 1998, 2002, and 2010 cohorts of entrepreneurs. We set the algorithmic selection policy for the second and third predictive models at the $s = 0.5\%$ threshold.

Female	-0.245***
	(-6.86)
Grande Ecole	-0.0176
	(-1.56)
High Tech	0.0323***
	(3.04)
Paris-based	0.215***
	(8.28)

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 8: Which Entrepreneurs Are More Likely to be Casting Errors? We sort firms into quintiles according to their predicted performance (log of revenue at age 5) and their predicted likelihood of being VC-backed. We keep firms in the top and bottom quintiles of these distributions and end up with two groups of firms: one group containing firms with the lowest predicted performance and the highest chances of being VC-backed, and one group containing firms with the highest predicted performance and the lowest chances of being VC-backed. This table reports t-tests for the difference in means of entrepreneur and venture characteristics between these two groups.

Table 9: Full vs. Simple Models. All estimators in this table predict $top5va_5$, a dummy equal to one for firms in the top 5% of their cohort in terms of (log) value added at age 5. The algorithms are trained on the sample of all new firms in the 1998, 2002 and 2006 cohorts. We report results using our test set which is the 2010 cohort of entrepreneurs. This table reports the results of a regression of VC-backed status on predictions of $top5va_5$ from our full model $\widehat{m}_{full}(X)$ and from the simple models $\widehat{m}_{simple}(X)$, which take as inputs only a subset X of features. Estimator $\widehat{m}_{simple}$(personal features) is trained 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)
$\widehat{m}_{full}(X)$	0.0282***	0.0259***	0.0281***	0.0264***	0.0268***	0.0230***	0.0281***	0.0283***	0.0233***	0.0265***
	(0.00343)	(0.00353)	(0.00345)	(0.00348)	(0.00344)	(0.00329)	(0.00344)	(0.00344)	(0.00365)	(0.00349)
$\widehat{m}_{simple}$(personal features)		0.0422***								
		(0.0149)								
$\widehat{m}_{simple}$(age)			0.00692							
			(0.0237)							
$\widehat{m}_{simple}$(male)				0.0682***						
				(0.0217)						
$\widehat{m}_{simple}$(graduate degree)					0.461***					
					(0.0763)					
$\widehat{m}_{simple}$(grande ecole)						0.426***				
						(0.0503)				
$\widehat{m}_{simple}$(French nationality)							0.0467			
							(0.124)			
$\widehat{m}_{simple}$(relatives)								-0.0150		
								(0.0589)		
$\widehat{m}_{simple}$(optimism)									0.0360***	
									(0.00916)	
$\widehat{m}_{simple}$(serial entrepreneur)										0.0393***
										(0.0146)
R^2	0.002	0.002	0.002	0.002	0.003	0.004	0.002	0.002	0.002	0.002
Observations	35,928	35,928	35,928	35,928	35,928	35,928	35,928	35,928	35,928	35,928

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)
$\widehat{m}_{full}(X)$	0.0282***	0.0273***	0.0282***	0.0282***	0.0282***	0.0279***	0.0283***	0.0300***	0.0191***
	(0.00343)	(0.00344)	(0.00343)	(0.00343)	(0.00343)	(0.00343)	(0.00343)	(0.00357)	(0.00422)
$\widehat{m}_{simple}$(Paris-based)		0.191***							
		(0.0502)							
$\widehat{m}_{simple}$(Marseille-based)			0.0240						
			(6.526)						
$\widehat{m}_{simple}$(Lyon-based)				-0.975					
				(1.410)					
$\widehat{m}_{simple}$(Bordeaux-based)					1.840				
					(3.082)				
$\widehat{m}_{simple}$(high tech)						1.656***			
						(0.279)			
$\widehat{m}_{simple}$(business services)							-0.0923		
							(0.118)		
$\widehat{m}_{simple}$(energy)								-0.0137*	
								(0.00764)	
$\widehat{m}_{simple}$(startup traction)									0.0257***
									(0.00700)
R^2	0.002	0.002	0.002	0.002	0.002	0.003	0.002	0.002	0.002
Observations	35,928	35,928	35,928	35,928	35,928	35,928	35,928	35,928	35,928

Standard errors in parentheses

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

Bibliography

Acharya, Viral V, Oliver F Gottschalg, Moritz Hahn, and Conor Kehoe. 2013. "Corporate governance and value creation: Evidence from private equity." *Review of Financial Studies*, 26(2): 368–402.

Arroyo, Javier, Francesco Corea, Guillermo Jimenez-Diaz, and Juan A Recio-Garcia. 2019. "Assessment of machine learning performance for decision support in venture capital investments." *Ieee Access*, 7: 124233–124243.

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.

Bernstein, Shai, Xavier Giroud, and Richard R Townsend. 2016. "The impact of venture capital monitoring." *Journal of Finance*, 71(4): 1591–1622.

Bonelli, Maxime, Jack Liebersohn, and Victor Lyonnet. 2021. "The Rising Bar to Entrepreneurship: Evidence from France."

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

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, 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.

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

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

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.

Gennaioli, Nicola, and Andrei Shleifer. 2010*a*. "What comes to mind." *The Quarterly journal of economics*, 125(4): 1399–1433.

Gennaioli, Nicola, and Andrei Shleifer. 2010*b*. "What Comes to Mind." *Quarterly Journal of Economics*, 125(4): 1399–1433.

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.

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.

Hebert, Camille. 2020. "Mind the Gap: 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.

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. 2020. "Human interactions and financial investment: A video-based approach."

Kahneman, Daniel. 2011. *Thinking, Fast and Slow.* Macmillan.

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 N, Berk A Sensoy, and Per Strömberg. 2009. "Should investors bet on the jockey or the horse? Evidence from the evolution of firms from early business plans to public companies." *Journal of Finance*, 64(1): 75–115.

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.

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.

Ludwig, Jens, and Sendhil Mullainathan. 2021. "Automated Discovery of Human Biases."

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

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." *The 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.”

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.

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 firm 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 firm.
Part-time Entrepreneur	Dummy equal to one if the entrepreneur is working for another firm while creating the new firm.
Same Prior Industry	Dummy equal to one if the entrepreneur has worked in the same industry the new firm 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 firm.
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).
Grande école	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 a required training to create the new firm.
Entrepreneur motivation and expectations	
Expectation: growth	Dummy equal to one if the entrepreneur expects the new firm's business to grow over the next 12 months.
Expectation: sustain	Dummy equal to one if the entrepreneur expects to sustain the new firm's business at its current level over the next 12 months.
Expectation: rebound	Dummy equal to one if the entrepreneur expects the new firm'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 firm is located in Paris.
Marseille-based	Dummy equal to one if the new firm is located in Marseille.
Lyon-based	Dummy equal to one if the new firm is located in Lyon.
Bordeaux-based	Dummy equal to one if the new firm is located in Bordeaux.
Business services industry	Dummy equal to one if the new firm is in the business services industry (naf1 code 74).
Health and social work industry	Dummy equal to one if the new firm is in the health and social work industry (naf1 code 85).
Construction industry	Dummy equal to one if the new firm is in the construction industry (naf1 code 45).
High-tech industry	Dummy equal to one if the new firm is in the high-tech industry industry (naf1 code 72).
Energy industry	Dummy equal to one if the new firm is in the energy industry industry (naf1 code 40).
B2B	Dummy equal to one if the new firm is business-to-business.
B2C	Dummy equal to one if the new firm is business-to-customer.
International customers	Dummy equal to one if the new firm has international customers.
Local customers	Dummy equal to one if the new firm has local customers.
Domestic customers	Dummy equal to one if the new firm has domestic customers.
Co-founders	Dummy equal to one if the entrepreneur has co-founders.
Outsourcing: Accounting	Dummy equal to one if the new firm outsources accounting services.
Number of employees	The number of employees in the new firm.
10+ clients	Dummy equal to one if the new firm has more than 10 customers.
Number of unpaid managers	The number of managers in the new firm who are not employed.
Number of paid managers	The number of managers in the new firm 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 firms to finance their firm.
Public grant	Dummy equal to one if the entrepreneur received a public grant to finance their firm.
Future VC financing	Dummy equal to one if the entrepreneur receives VC-backing up to 5 years after creation (this variable is constructed from other SINE survey waves following entrepreneurs over time).

Appendix B Model Interpretability

We would like to make our models more transparent, first to improve their interpretability by documenting which features matter in generating the predictions, and second, to compare the features that are relevant for performance predictions versus VCs' decisions. Lundberg and Lee (2017) develop an approach to improve model interpretability 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 captures the direction and extent to which it moves the model's output away from its unconditional expectation. It is the change in expected model output, averaged across all possible orderings of all other features. SHAP values can be aggregated across observations to facilitate the model's global interpretability by yielding a ranking of features that contribute the most to the predictions. We wish to emphasize that SHAP values do not allow any causal interpretation, but they are helpful to understand how our algorithmic models generate predictions.[26]

[26] Erel et al. (2021) use the SHAP method to better understand their model's predictions of corporate directors' performance.

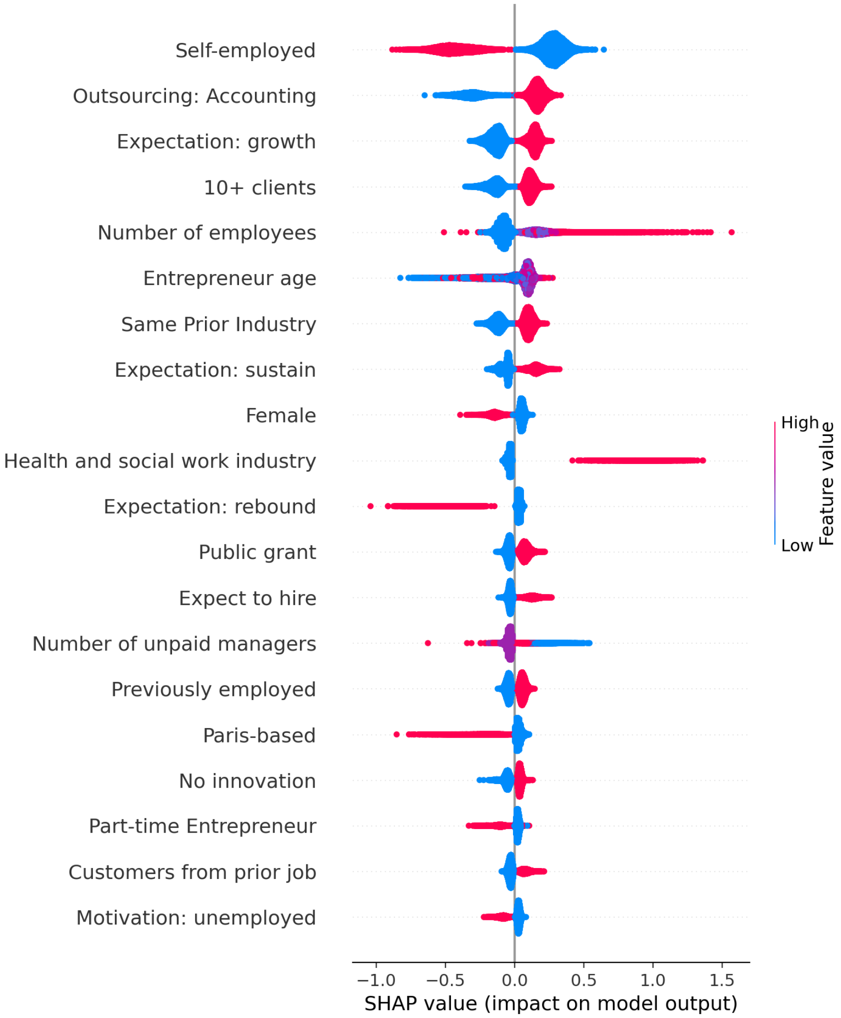

Figure 1: SHAP Values of Most Important Input Features to Predict Operating Performance. This figure reports the SHAP values for the top-20 features that are most important in predicting operating performance (log of revenue at age 5). Features are ranked in decreasing order of importance. For each feature, each point represents one observation and its location on the x-axis indicates its SHAP value. Positive (negative) SHAP values indicate that feature's value for this observation increased (lowered) the prediction of operating performance. Colors capture the value of the feature for each observation. The predictive model is trained on all new firms in the 1998, 2002, and 2006 cohorts using ten-fold cross validation.

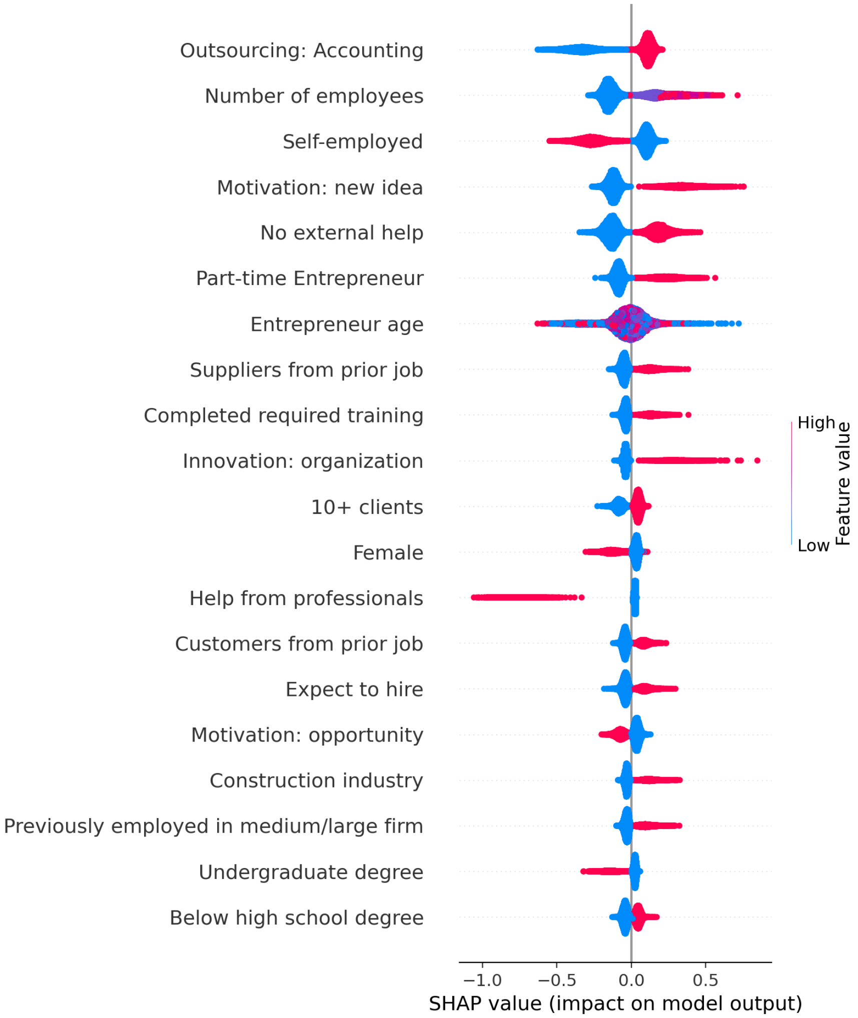

Figure 2: SHAP Values of Most Important Input Features to Predict VC backing. This figure reports the SHAP values for the top-20 features that are most important in predicting whether a firm will receive VC financing. The predictive model is trained on a random sample of all new firms in the 1998, 2002, and 2010 cohorts using five-fold cross validation. Features are ranked in decreasing order of importance. For each feature, each point represents one observation and its location on the x-axis indicates its SHAP value. Positive (negative) SHAP values indicate that feature's value for this observation increased (lowered) the prediction of operating performance. Colors capture the value of the feature for each observation.

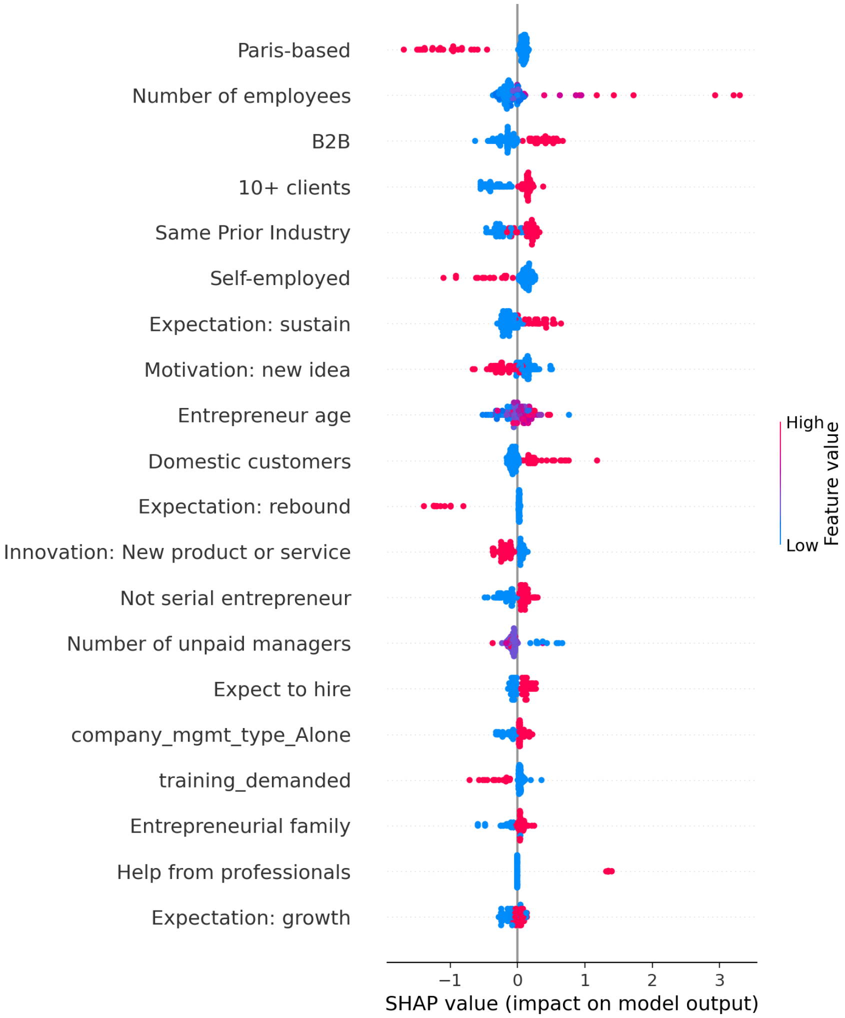

Figure 3: SHAP Values of Most Important Input Features to Predict Operating Performance When the Model is Trained on VC-backed Firms Only. This figure reports the SHAP values for the top-20 features that are most important in predicting operating (log of revenue at age 5). Features are ranked in decreasing order of importance. For each feature, each point represents one observation and its location on the x-axis indicates its SHAP value. Positive (negative) SHAP values indicate that feature's value for this observation increased (lowered) the prediction of operating performance. Colors capture the value of the feature for each observation. The predictive model is trained on new VC-backed firms in the 1998 and 2002 cohorts using ten-fold cross validation.

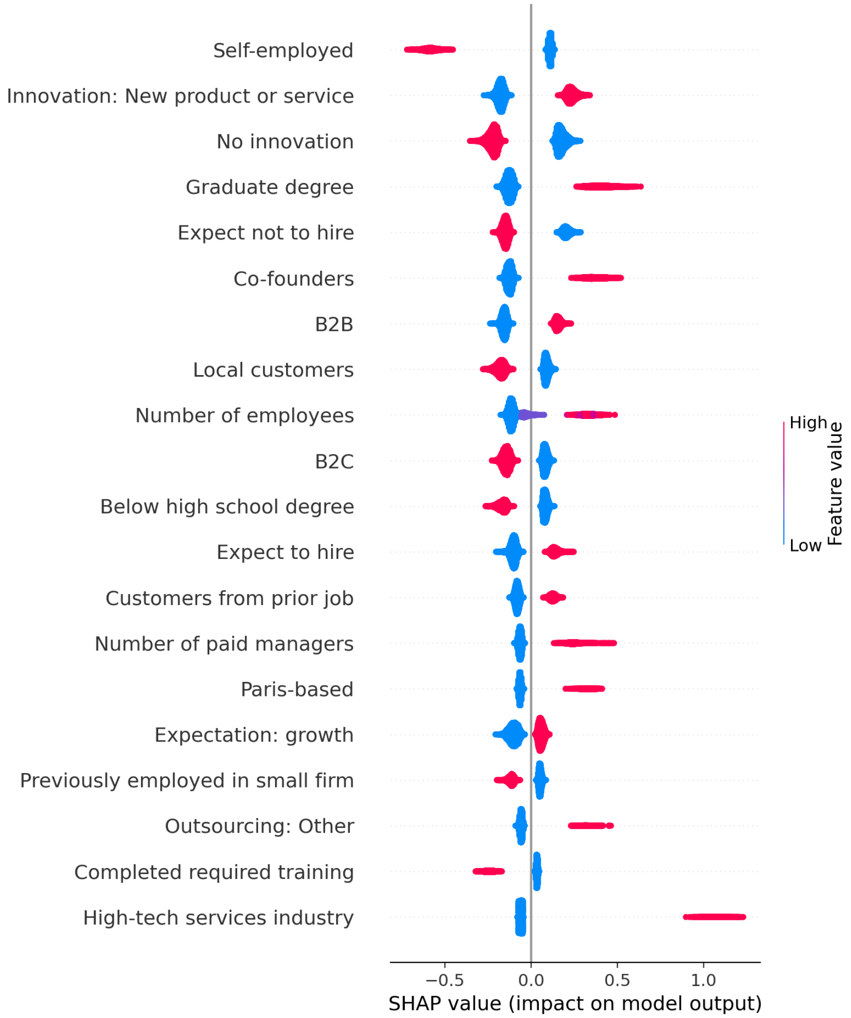

Figure 4: SHAP Values of Most Important Input Features to Predict Home Runs. This figure reports the SHAP values for the top-20 features that are most important in predicting successful deals. The dummy variable *successful deal* is a proxy for a successful exit by the VC, i.e., a "home run." Whether a new firm is VC-backed or not, *successful deal* takes a value of one if the firm receives a (new) round of VC funding (e.g., series B funding), if it is acquired by another firm, or if it goes public. We use data from Crunchbase, Capital IQ, CB Insights, Preqin, Venture Xpert, SDC and Zephyr to construct the successful deal measure. Features are ranked in decreasing order of importance. For each feature, each point represents one observation and its location on the x-axis indicates its SHAP value. Positive (negative) SHAP values indicate that feature's value for this observation increased (lowered) the prediction of operating performance. Colors capture the value of the feature for each observation. The predictve model is trained on the sample of all new firms in the 1998, 2002 and 2006 cohorts using ten-fold cross validation.

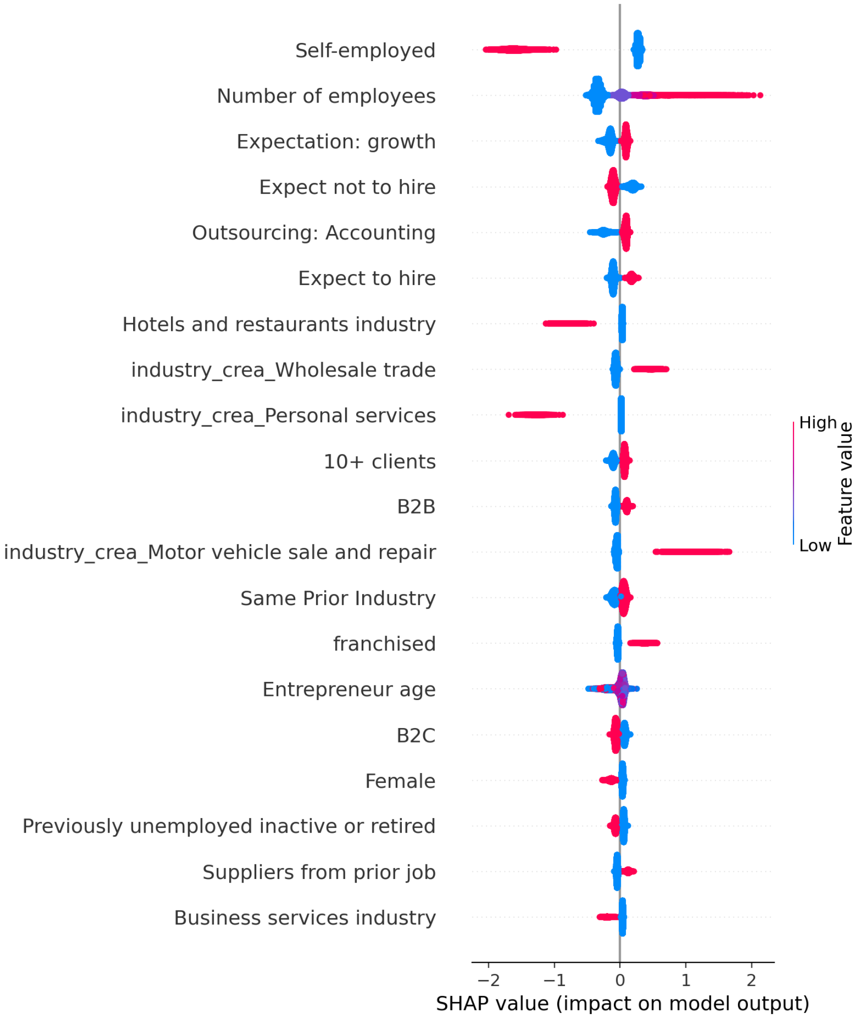

Figure 5: SHAP Values of Most Important Input Features to Predict Best Performers. This figure reports the SHAP values for the top-20 features that are most important in predicting whether a firm will be in the top 5% of its cohort in terms of operating performance (revenue at age 5). Features are ranked in decreasing order of importance. For each feature, each point represents one observation and its location on the x-axis indicates its SHAP value. Positive (negative) SHAP values indicate that feature's value for this observation increased (lowered) the prediction of operating performance. Colors capture the value of the feature for each observation. The predictve model is trained on the sample of all new firms in the 1998, 2002 and 2006 cohorts using ten-fold cross validation.

Appendix C Additional Tables and Figures

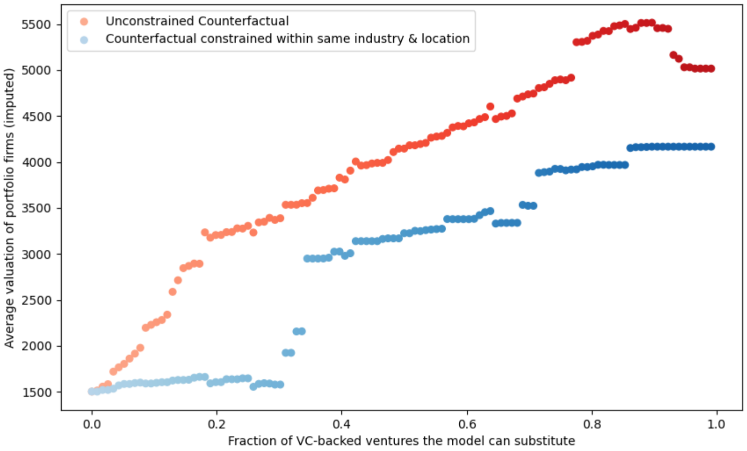

Figure C.1: Counterfactual policy evaluated with imputed valuations. This figure reports the average imputed valuation at age 5 for two counterfactual models that replace VC-backed firms that are predicted to become poor performers with firms that are predicted to become good performers by the algorithm. The origin represents the status quo: it includes the full set of VC-backed firms in the test set and their average imputed valuation at age 5. The red line shows the performance of the unconstrained counterfactual model. The blue line shows the performance of a counterfactual model constrained to replace VC-backed firms with firms that are in the same industry and location. We calculate industry-level median exit valuation multiples for early deals in Pitchbook data starting in 2000. Imputed valuations are constructed by multiplying the firms' observed revenue at age 5 by their respective industry's median exit valuation multiple. Numbers on the y-axis are in thousands.

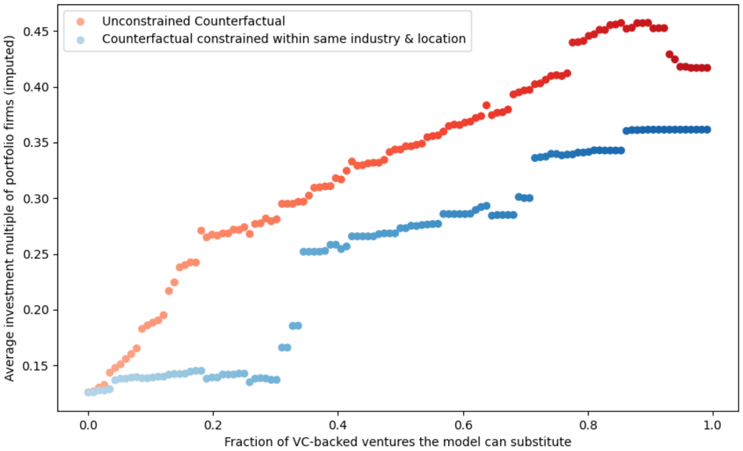

Figure C.2: Counterfactual policy evaluated with imputed investment multiples. This figure reports the average imputed investment multiples at age 5 for two counterfactual models that replace VC-backed firms that are predicted to become poor performers with firms that are predicted to become good performers by the algorithm. The origin represents the status quo: it includes the full set of VC-backed firms in the test set and their average imputed investment multiple at age 5. The red line shows the performance of the unconstrained counterfactual model. The blue line shows the performance of a counterfactual model constrained to replace VC-backed firms with firms that are in the same industry and location. We calculate industry-level median exit valuation multiples for early deals in Pitchbook data starting in 2000. Imputed valuations are constructed by multiplying the firms' 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 fraction acquired for early deals, times 75% to account for dilution, and scaled by the industry's median deal size for early deals.

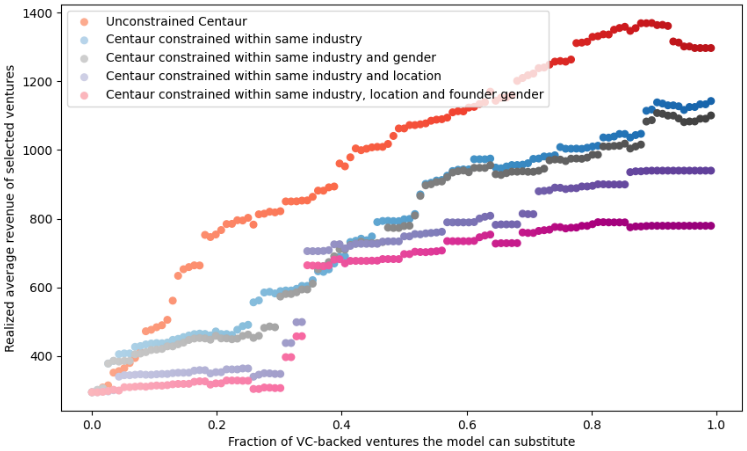

Figure C.3: The shadow cost of backing too few female entrepreneurs. Our findings suggest that VCs pass on promising founders with particular demographics. Recent work has placed special emphasis on the role of the founder's gender in VCs' decisions (e.g., Calder-Wang and Gompers, 2021; Hebert, 2020), and interest groups are trying to raise awareness of the inequity of VC finance between male and female entrepreneurs.[27] We therefore use our counterfactual model approach to provide an estimate of the shadow cost of backing too few female entrepreneurs. The setup is the same as in Section 5.3 but now the algorithm faces one additional constraint: in addition to having to select a new venture within the same industry (or industry and location), it must pick one with a founder of the same gender as the VC-backed firm it drops. When all firms in the portfolio are selected by the algorithm (rightmost point), the average performance of portfolio firms increases by 6% (5%) when the gender constraint is relaxed, relative to when the counterfactual model is constrained to gender and industry (gender, industry, and location). While this approach is subject to the same interpretational limitations and caveats as those described in Section 5.3, this estimate of the shadow cost of gender preferences contributes to broadening our understanding of the extent to which frictions affect VC allocation efficiency.

Feature	Top 5%	Bottom 95%	Representativeness	Top 1%	Bottom 99%	Representativeness	Representativeness
Feature			of best performers			of best performers	*among* VC-backed firms
			$\frac{Pr(X_i \mid \text{Top5})}{Pr(X_i \mid \text{Bottom95})}$			$\frac{Pr(X_i \mid \text{Top5})}{Pr(X_i \mid \text{Bottom95})}$	$Pr(X_i \mid \text{VC-backed})$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VC Hub	64.2%	61.8%	1.04	67.0%	61.9%	1.08	65%
California	40.1%	38.8%	1.03	44.2%	38.8%	1.14	39%
Massachusetts	10.2%	8.6%	1.19	11.7%	8.6%	1.36	8.9%
New York	8.0%	7.5%	1.07	7.1%	7.5%	0.95	7.3%
Texas	5.9%	7.0%	0.84	4.1%	6.9%	0.59	6.7%
Largest Industries	77.7%	75.3%	1.03	82.2%	75.4%	1.09	89.8%
Information Technology	44.4%	45.7%	0.97	41.6%	45.7%	0.91	44.8%
Health Care	26.3%	20.2%	1.30	35.5%	20.3%	1.75	21.5%
Consumer Discretionary	7.0%	9.5%	0.74	5.1%	9.4%	0.54	9.4%
Industrials	6.1%	8.1%	0.75	5.1%	8.1%	0.63	7.6%
Communication	5.9%	6.1%	0.97	6.1%	6.1%	1.00	6.5%

Table C.1: Stereotypes of the Most Successful Entrepreneurs in U.S. Burgiss Data (Using IRR). This table reports the fraction of entrepreneurs with a given characteristic among the best performing firms and among the other firms. The two deal characteristics available in the Burgiss data are the firm 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 firms are in the top 5%, and the other firms in the bottom 95%, in terms of IRR. In columns 4 and 5, the best performing firms are in the top 1%, and the other firms in the bottom 99%, in terms of IRR. A given characteristic is representative (or stereotypical) of the best performing firms 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.