Venture Capital (Mis)Allocation in the Age of AI

Victor Lyonnet Léa H. Stern*

13th May, 2022

Abstract

We use machine learning to study how 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, machine learning, entrepreneurship, stereotypes.

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

*For helpful comments, we thank Will Cong, Camille Hebert, Johan Hombert, Chris Hrdlicka, Hyiek Kim, David Robinson, Daniel Rock, Amin Shams, Andrei Shleifer, Michael Weisbach, Karen Wruck, and conference and seminar participants at University of Colorado Boulder, Ohio State, University of Washington, and the 2022 RFS-GSU Fintech conference. Samin Jalali, George Nurisso, Boyu Wang, Daisy Wang, and especially Danial Salman provided excellent research assistance. 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. A typical venture capitalist (VC) considers approximately two hundred of them and ends up backing around four (Gompers et al., 2020). Without any historical data on new firms and a complex set of entrepreneur and new firm characteristics, 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? We use machine learning (ML) prediction methods to answer these questions.

We use French administrative data to identify 123,511 new firms from four cohorts of entrepreneurs who create a firm between 1998 and 2010. Our sample is representative of the entire population of entrepreneurs and free from survivorship and selection bias. The data contain detailed information on more than 400 features of entrepreneurs and new firm characteristics. We use these features to train a Gradient Boosting Trees algorithm (*XGBoost*) on the first three cohorts of entrepreneurs to predict new firms' operating performance five years after creation, and evaluate the algorithm's out-of-sample predictions in the last cohort of entrepreneurs. All results pertain exclusively to the algorithm's performance in this test set, which was left untouched during training. Because we observe firms' operating performance regardless of whether they are VC-backed, we circumvent the selection problem in which data on outcomes can be missing in a nonrandom manner (the "selective labels" issue).

We define an algorithmic investment policy that selects the most promising ventures based on algorithmic predictions of firm performance. To train our model, we need a measure of firm performance that is available for *all* new firms, and one that is a good proxy for private investors' return. We therefore use firms' operating performance at a typical VC investment horizon (Acharya et al., 2013). Our approach is to contrast VCs' decisions to this algorithmic policy. Importantly, we do not assume that the 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. We study the differences between VC-backed and algorithm-selected ventures to uncover sources of errors and we build a separate model that predicts VCs' decisions to broaden our understanding of their decision making process.

Although VC-backed firms perform much better than the average firm in our sample, our approach reveals two potential types of errors in VCs' investment decisions. First, VCs invest in some firms that perform predictably poorly. If VCs passed on the bottom half of their portfolio firms in

terms of predicted performance, the average performance of portfolio firms would increase by 48%. Second, VCs pass on some new firms that perform predictably well. If they selected predictably good performers instead of predictably poor performers, the average performance of portfolio firms would be an order of magnitude higher.

Why do firms selected by the algorithmic policy perform better? To answer this question, we start by proposing a method to estimate the shadow cost of constraints faced by VCs: We compare the performance of the (unconstrained) algorithmic policy to that of algorithmic policies constrained to selecting firms similar to VC-backed ones (e.g., firms in the same industry, location, or industry-location). For example, VCs tend to invest locally, within a given commuting zone. Lifting this constraint would increase the average performance of portfolio firms by 13%, which we interpret as the shadow cost of the location constraint. 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. To better explain this difference, we compare the entrepreneur characteristics of VC-backed firms to those of algorithm-selected firms. Focusing on entrepreneur characteristics that have drawn most attention from the existing literature, we show that compared to the algorithmic policy, VCs tend to select fewer female entrepreneurs, more young entrepreneurs, and more Paris-based entrepreneurs.

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.

Our interpretation of this first set of results relies on the assumption that VCs could, in principle, have invested in the algorithm-selected firms they did not back, and that these firms would have accepted VC. Our empirical design and findings support these assumptions and attenuate concerns that demand-side or supply-side considerations explain why algorithm-selected firms were not VC-backed.[1] First, our analysis is restricted to firms that are created in industries that receive VC-

[1]The fact that VCs support the entrepreneurs they select is not an issue for our approach. Indeed, this support

backing in our data. This restriction helps alleviate concerns that algorithm-selected firms are not suitable candidates for VC. Second, 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 and to the two-sided matching process of VC investment. Third, we construct an algorithmic policy that is constrained to selecting entrepreneurs with the same growth aspirations as VC-backed ones, and show that even this highly constrained algorithm significantly outperforms VC-backed firms. This exercise helps establish that demand or supply-side effects are unlikely to explain our main result. Fourth, we argue that the characteristics along which algorithm-selected firms differ from VC-backed ones (e.g., gender, age, location) are unlikely to fully explain why these firms could not receive VC. Finally, two additional findings support the idea that VCs pass on predictably good performers due to prediction errors: our algorithm can predict performance and improve VC allocation even *within* the set of VC-backed firms (by dropping predictably poor performers), and it can still improve VC allocation when trained on VC-backed firms only (instead of all new firms).

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 is much less accurate when it is restricted to these three features. We view this result 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 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

implies that a VC-backed firm performs better on average than an otherwise similar non-VC-backed firm (Chemmanur, Krishnan and Nandy, 2011; Puri and Zarutskie, 2012; Bernstein, Giroud and Townsend, 2016). Therefore, it effectively raises the bar for our algorithm to identify outperforming non-VC-backed firms.

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). Controlling for a firm's probability of being among the best performers in its cohort, we find that moving from the first to the third quartile of representativeness increases an entrepreneur's chances of being backed by 51% relative to the baseline average.[2] These results imply that VCs make prediction errors as they exaggerate some representative features of success in their decisions.

To make the algorithm more transparent and understand how it maps entrepreneur and firm characteristics into performance predictions, we use the SHapley Additive exPlanations (SHAP) method. We find a strong overlap in the features that matter the most when predicting firm performance and when predicting VCs' decisions, suggesting that the algorithmic model uncovers patterns in firm performance that match VCs' investment decisions.

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 approach helps reconcile a series of existing evidence on VCs' investment decisions. 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

[2]The features that capture representativeness include the entrepreneur's gender, whether the entrepreneur has a graduate degree, whether the firm is based in Paris, and whether it operates in the high-tech industry.

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 (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 is based on Mullainathan and Obermeyer (2022). In our model, VCs can invest in many new firms that are characterized by a vector of features (X, Z) drawn from a fixed distribution. Both X and Z are observed by VCs, but only X is recorded in the data. The performance of new firms is a random variable denoted $Y(X, Z)$. Based on the features (X, Z), *rational* predictions $\tilde{Y}(X, Z)$ of firms' performance can be formed.[3] We denote $R(X, Z) \in [1 : 100]$ the percentile rank of these predictions.

The VCs' contractual payoff depends on the performance of their selected firms (Gompers and Lerner, 1999). Therefore, we define the VCs' *optimal* policy as investing in the most promising new firms:

$$I = 1 \text{ iff } R(X, Z) > t, \tag{1}$$

where $I = 1$ if the VCs invest, $I = 0$ otherwise, and t is the percentile threshold under which the VCs do not invest.[4]

We investigate whether VCs' decisions differ from the optimal policy described in (1). We write the VCs' actual policy as

$$I = 1 \text{ iff } R(X, Z) > t + \Delta(X, Z), \tag{2}$$

where $\Delta(X, Z)$ captures shifts in the investment threshold for firms with characteristics (X, Z). These shifts may arise due to VCs' biases, constraints, or private benefits, for example. They affect VCs' willingness to invest in these firms, which makes them depart from the optimal investment policy in (1).

[3] Sterk, Sedláček and Pugsley (2021) show that ex ante differences in (X, Z) across firms is a key determinant of firm performance $Y(X, Z)$, and that ex post shocks do not matter as much.

[4] The threshold t is determined outside our model and depends on VCs' financing and operational constraints. Empirically, $t = 99.5$, so that VCs invest in only 0.5% of all new firms (Lerner and Nanda, 2020).

If $\Delta(X,Z) < 0$, there are firms in percentile ranks $R(X,Z) < t$ which receive VC-backing, in contradiction with the optimal policy in (1). We denote the set $\nu^{\mathbf{L}}$ of *over-backed* firms such that

$$\underbrace{E(I|(X,Z) \in \nu^{\mathbf{L}}) > 0}_{\text{VC-backed}} \text{ and } \underbrace{E(R|(X,Z) \in \nu^{\mathbf{L}}) < t}_{\text{but below threshold}}.$$

If $\Delta(X,Z) > 0$, there are firms in percentile ranks $R(X,Z) > t$ which do not receive VC-backing. We denote the set $\nu^{\mathbf{H}}$ of *under-backed* firms such that

$$\underbrace{E(I|(X,Z) \in \nu^{\mathbf{H}}) < 1}_{\text{not VC-backed}} \text{ and } \underbrace{E(R|(X,Z) \in \nu^{\mathbf{H}}) > t}_{\text{but above threshold}}.$$

To establish inefficiencies in VCs' investment decisions, it is enough to observe a non-empty set $\nu^{\mathbf{L}}$ or $\nu^{\mathbf{H}}$ of firms in the data. We use machine learning prediction methods to identify inefficiencies such that VCs shift the investment threshold ($\Delta(X,Z) \neq 0$), and understand the sources of these inefficiencies.

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.[5]

[5]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 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 focuses on VC-backed firms in isolation, our sample contains both VC-backed and non-VC-backed firms.[6,7] Appendix B contains descriptions for a subset of the variables we use from the survey.

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.[8] 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.[9] 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.[9] 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.

---

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.

[6]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).

[7]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.

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

[9]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.

4 Algorithmic Policy 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. This estimator produces a percentile rank of performance predictions denoted $M(X_i)$.[10] We do not assume that the algorithmic predictions are correct, i.e., that $R(X_i, Z_i) = M(X_i)$. Rather, we use $M(X_i)$ to isolate potential errors by VCs, and we rely on realized outcomes Y_i to evaluate actual errors.

Choice of objective function. The estimator $\widehat{m}(X_i)$ is trained to predict new firms' performance Y_i. Importantly, the outcome should be a measure of firm performance that is available for all new firms, not only VC-backed ones. Our main measure is the firms' (log) value added at age 5, which is a typical VC investment horizon (Gompers et al., 2020).[11] Value added is very similar to EBITDA, with a correlation of 0.91.[12] VCs' compensation structure has remained largely unchanged over the years (Lerner and Nanda, 2020) and their profits are contractually related to their portfolio firms' operating performance. Therefore, selecting new firms based on predictions of their future performance is at the core of VCs' business (Bernstein, Korteweg and Laws, 2017; Gompers et al., 2020). Critical for our interpretation of the results, we observe firms' operating performance for VC-backed firms as well as for non-VC-backed firms, which allows us to circumvent the selective labels problem (Kleinberg et al., 2018).[13] Our results are robust to predicting various measures of venture success, and in particular, measures that explicitly account for VCs' preference for skewness (see Section 5.2 for alternative outcomes, including "home run" deals).

[10] Recall that only the features in X are recorded in the data, which puts the algorithm at a disadvantage to predict performance, compared to VCs who observe the vector of features (X, Z). While VCs do not have access to the entrepreneur survey, we ensure that the features the algorithm uses as inputs would easily be part of VCs' information set when conducting a first pass evaluation of the ventures.

[11] We assign a zero as the (log) value added at age 5 for firms that do not survive or have a negative value added at age 5 (the value of the first percentile of the distribution). Our results are not sensitive to the outcome value we choose for these firms.

[12] We train our algorithm to predict the (log of) value added rather than EBITDA for two reasons. First, value added is measured after taxes, depreciation and amortization, and after all investors (debtholders and shareholders) are paid their interest. Because value added is the relevant measure of profit that directly discounts to value, it can explain variation in stock values better than EBITDA (Stewart, 2019). Second, the value added reported in the tax files is *pro forma* and serves as the basis for the computation of the value added tax by the French tax authority. Our results are very similar using EBITDA.

[13] We recognize that outcomes may be subject to treatment effects and assume treatment effect homogeneity. Note, however, that if the outcome for VC-backed firms is inflated due to VCs' involvement (e.g., Puri and Zarutskie (2012)), this would raise the bar for the algorithm to identify outperforming non-VC-backed firms.

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 (85,658 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 (37,853 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 452 covariates X 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.[14]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. Table 1 reports summary statistics for a subset of input features. We report these statistics separately for the training set and the test set.

[Insert Table 1 here]

Although most input features (i.e., entrepreneur and firm characteristics) are similar across the training and test sets, we observe that average realized performance is slightly higher in the test set compared to the training set, and some founder characteristics such as the entrepreneur's age and education are somewhat larger in the test set.[15]

[14]To facilitate the interpretation of our results, we exclude firms whose founder reported more than 20 employees in the year of creation. Our results are not sensitive to this filter.

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

5 Model Performance

5.1 Predicting the distribution of firm outcomes

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 value added at age 5 – $log(va_5)$) for the 37,853 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, we find that only about 0.5% of new ventures receive VC funding, i.e., the empirical value of t in (1) is 99.5. This fraction is similar to that in the US (Puri and Zarutskie, 2012; Lerner and Nanda, 2020). We set a threshold t and ask the algorithm to select firms in the top $s = 1 - t\%$ of predicted performance Y_i. The algorithmic policy writes:

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

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).[16] 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

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 would play *against* us, making it more difficult for an algorithm trained on the earlier training set to predict the performance of new firms in the later test set.

[16] Although it appears constant to the eye in Figure 2, the average realized performance is not exactly equal across the first four quintiles of the top 0.5% firms. There also is variation in realized performance within quintile.

by successfully avoiding to fund firms with low potential, but also by (re)allocating capital within the set of most promising ventures.

[Insert Figure 2 here]

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.[17] Figure 3 reports the distribution of realized outcomes for three sets of firms: 1- the entire test set, 2- VC-backed firms, and 3- algorithm-selected firms.

[Insert Figure 3 here]

Figure 3 illustrates several interesting facts. First, the average VC-backed firm is much more profitable than the average firm: $log(va_5)$ for VC-backed firms is 2.26, whereas it is 1.88 for the entire sample. This performance gap confirms VCs' ability to identify and invest in promising new ventures.[18] Second, we confirm the *Babe Ruth Effect* in our data: VCs bet on magnitude over frequency, and outcomes follow a power law distribution.[19] 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 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.

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, we find that the algorithmic policy improves on VCs' decisions, suggesting that VCs' policy in Equation (2) differs from the optimal policy in Equation (1), such that $\Delta(X, Z) \neq 0$.[20]

5.2 Alternative outcomes: "home run" deals and other measures of success

Because portfolio firms' valuations are commonly based on multiples of accounting performance, VCs' returns are highly correlated with their portfolio firms' performance for the vast majority of

[17] 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. Our results remain qualitatively similar without this filter.

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

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

[20] We obtain similar results when we restrict the analysis to Paris-based firms only, and to firms that operate in the five most VC-intensive industries.

deals. Although more the exception than the rule, there are cases where a portfolio company does not necessarily have to generate profits for VCs to earn large returns (e.g., the company goes public while it is still losing money). In this section, we verify that our results are robust to using other 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.[21]

These successful exits are extremely rare. In our test set (the 2010 cohort of entrepreneurs), only 56 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.[22] As before, we implement an algorithmic policy which selects the firms above the investment threshold t, $I = 1$ 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 120 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%.[23] This is the benchmark we use to evaluate the performance of our algorithm.

[Insert Table 2 here]

[21] 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.

[22] The percentile rank $M'[X_i]$ is different from that of performance predictions, $M[X_i]$ defined in Section 5.1.

[23] 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 positives plus true positives). Both precision and recall are similar in previous cohorts of entrepreneurs for which we observe VC-backed status.

We report in Table 2 that the algorithmic investment policy we implement identifies *ex ante* 11 firms that subsequently experience a successful exit.[24] 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 $top5va_5$, a dummy equal to one for firms in the top 5% of their cohort in terms of operating performance 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. Row 3 of Table 3 contains the results. We find that 60% of algorithm-selected firms are among the best performers in their cohort, compared to only 14% of VC-backed firms (last row).[25]

Other outcome measures. Table 3 contains the performance of predictive models trained on several other outcome measures: (log) value added at age 7, a dummy equal to one for firms in the top 5% of their cohort in terms of operating performance at age 5, at age 7, and EBITDA at age 5 scaled by capital at creation. 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. For example, when the algorithmic policy ($s = 0.5\%$) is trained to predict $log(va_5)$ or $log(va_7)$, it selects four firms classified as home runs, the same number as VCs.

[Insert Table 3 here]

5.3 Potential performance gains and the shadow cost of VCs' constraints

Contracting the set of VC-backed firms. We would like to assess VCs' deviation from the objective to invest in the most promising firms defined in Equation (1), and evaluate the potential performance gains from introducing a VC algorithmic decision aid. One way to perform this evaluation is to observe how the average realized performance of VCs' portfolio companies changes as the set of VC-backed firms is contracted when we exclude those ventures the algorithm predicts

[24] 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%.

[25] Figures B.4 and B.5 further depict the performance of this classification algorithm.

to perform poorly. We drop firms one at a time, starting with the one with the lowest percentile rank $M(X_i)$. Figure 4 reports the results of this first exercise. The origin represents the status quo: the average performance at age 5 of the full set of VC-backed firms in the test set. The figure illustrates how portfolio firms' average performance changes as firms are dropped out of the set. The rightmost observation reports the realized performance of the VC-backed firm with the highest algorithmic percentile rank $M(X_i)$.

[Insert Figure 4 here]

The average value added of portfolio firms would increase by 48% if VCs dropped the bottom half of their portfolio firms in terms of predicted performance. Of course, there are several caveats to this approach. We do not have data on deal size and we can only express potential performance gains in terms of portfolio firms' average performance. These gains do not capture VCs' returns directly; instead, they measure gains in terms of portfolio firms' average profits, which is highly correlated with VCs' returns. Despite these data limitations, our findings reveal that VCs invest in firms that perform predictably poorly. Therefore, the set $\nu^{\mathbf{L}}$ of "over-backed" firms is not empty.

A Centaur model of VC allocation. Another way to assess VCs' deviation from the policy in Equation (1) is to document the potential performance gains from dropping firms in the set $\nu^{\mathbf{L}}$ (the "over-backed" firms) and investing instead in firms in the set $\nu^{\mathbf{H}}$ (the "under-backed" firms).

This Centaur model sequentially drops VC-backed firms with bad predicted performance and replaces them with firms with good predicted performance. We first run the Centaur model without constraints ("unconstrained Centaur"). 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 Centaur model lets the algorithm select a new venture only within the same industry as the VC-backed firm it drops. Our second constrained Centaur model can only choose firms within the same location as the VC-backed firm it drops. Our third Centaur model is constrained to pick a firm within the same industry *and* location for each VC-backed firm it drops.

[Insert Figure 5 here]

Figure 5 shows that as the Centaur 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 Centaur models outperform VCs' selections. Second, when the Centaur models assign full weight on the algorithm's selections, we interpret the difference in average portfolio firm performance between the unrestricted Centaur model and each Centaur 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 Centaur model. We find that broadening investments in terms of industries or location can increase the average performance of portfolio firms by 30% and 13%, respectively. 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. Therefore, it could be that the VCs' policy in (2) is such that $\Delta(X, Z) \neq 0$ even absent constraints to VCs' investment decisions. We explore other explanations in Section 6.

5.4 Discussion of assumptions

Our approach is to contrast VCs' decisions to the algorithmic policy that selects the most promising ventures. This approach relies on the assumption that VCs could, in principle, have invested in the algorithm-selected firms they did not back, and that these firms would have accepted VC.

General comments about our approach. Our empirical design reduces concerns about the validity of this assumption. First, our analysis is restricted to firms that are created in industries that receive VC-backing in our data. This restriction alleviates concerns that algorithm-selected firms are not suitable candidates for VC, at least with respect to their industry. Note also that VCs' support to entrepreneurs and their complex two-sided matching process are not an issue for our approach: VCs' support effectively raises the bar for our algorithm to identify outperforming non-VC-backed firms, and we observe a firm's VC-backed status in the aggregate so that we are not limited to observing whether a firm matches with one particular VC. Therefore, our approach mitigates concerns related to negotiations and to the two-sided matching process of VC investment.

Reducing the pool from which the algorithm can select firms. To further alleviate potential concerns with our assumption, we design a test to explore demand-side and supply-side reasons that could explain why algorithm-selected firms are not VC-backed. It could be that entrepreneurs did not want VC, for example if they do not intend to grow. It could also be that VCs choose not to invest in these firms because they are different from the ones they typically consider.

To address the concern that algorithm-selected firms are not a good fit for VC, we propose a new Centaur model in Figure 6. This model drops VC-backed firms in the set $\nu^{\mathbf{L}}$ of predictably poor performers, and replaces them with firms from a restricted subset of the set $\nu^{\mathbf{H}}$ of predictably good performers. We first restrict the set $\nu^{\mathbf{H}}$ to entrepreneurs who have the same growth aspirations as VC-backed ones (in orange).[26,27] 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). As expected, the more restrictive this Centaur model, the lower its average performance compared to the unconstrained Centaur. Importantly, for both specifications, we find that the Centaur models outperform VCs' selection. Therefore, VCs pass on some firms that perform predictably well even though these firms closely resemble the firms they typically select.[28]

[Insert Figure 6 here]

Analysis using VC-backed firms only. Two additional findings lend further support to the idea that VCs pass on the predictably best performers. First, we evaluate the algorithm's performance in the test set on VC-backed firms only. Figure 7 shows that the algorithm predicts the distribution of outcomes well, even for the 120 VC-backed ventures in our test set. This finding suggests that the algorithm is useful to not only help identify *ex ante* successful firms that, for reasons we do not observe, are not part of VCs' dealflow, but also to reallocate capital *within* the set of firms that, by revealed preferences, have been approved by VCs.

[Insert Figure 7 here]

Second, we estimate a separate model which instead of being trained on all new firms, is trained using only VC-backed firms. In this way, the algorithm learns patterns to predict success only within the set of firms that have been selected by VCs. Figure 8 shows that an algorithmic policy tasked with selecting the top 15% of firms would outperform the allocation of VCs.[29] While this is

[26] 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?"

[27] 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.

[28] 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.

[29] We increase the threshold s to 15% for this exercise as there are 120 VC-backed firms in our test set. Results are not sensitive to the choice of threshold. In addition, because the 2006 entrepreneur survey does not allow us to identify VC-backed firms, we do not use this cohort to train the algorithm when we restrict the analysis to VC-backed firms only.

not our preferred specification (we wish to take advantage of having data on all new firms to learn about VCs' decision making), these findings address potential concerns about unobservables and further support the idea that VCs make errors in their investment decisions.

[Insert Figure 8 here]

Profiles of non-VC-backed best performers. 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. Moreover, our finding that VCs invest in some firms that perform predictably poorly (the set $\nu^{\mathbf{L}}$ of "over-backed" firms, see Section 5.3) suggests that VCs make errors that lead to them to invest in these firms, such that $\Delta(X, Z) \neq 0$ in their policy (2). This finding makes it plausible that VCs might *also* pass on firms that perform predictably well. Section 6 investigates VCs' decisions in details.

5.5 Model interpretability

We would like to make our model more transparent for two main reasons. First, we wish to improve its interpretability by documenting which features matter in generating the predictions. Second, we want 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. SHAP values do not allow any causal interpretation, but they are helpful to understand how our algorithmic models generate predictions.[30]

Figure 9 reports the SHAP summary plot when the algorithm predicts operating performance ($log(va_5)$). It shows the input features with the highest impact on the model's predictions (input

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

features are listed in descending order of importance), as well as the effect of each feature. Each row shows an input feature and the Shapley value for each observation's feature is shown on the x-axis. Observations with a dummy variable equal to one (zero) are shown in red (blue). For continuous variables, observations with a high (low) value are shown in red (blue), while intermediate values are shown in purple. For each feature, overlapping points are stacked vertically, showing the distribution of Shapley values for each feature. Features with SHAP values larger (less) than 0 push the prediction above (below) the unconditional expectation.

[Insert Figure 9 here]

The top 5 features include whether the founder was self-employed prior to starting the business, whether the firm pays for external administrative and accounting services, the total number of employees, whether the founder has prior experience in the industry the new venture operates in, and the entrepreneur's age. The analysis of SHAP values shows that a large number of founders were working in the same industry prior to creating the venture (stacked red observations for *Same Prior Industry*), and this increases the prediction of performance five years out (red observations have a positive SHAP value). The founder's age has a large distribution of SHAP values. Many founders are middle aged (represented in the stacked purple observations), and younger founders tend to be associated with lower SHAP values, with a long left tail. Importantly, the results strongly point to the importance of feature interactions. While on average, it appears that higher values of age tend to push performance predictions above the unconditional mean (purple and red observations lay in the positive SHAP value region), in some cases the algorithm views older entrepreneurs as a very negative signal, as evidenced by the very negative SHAP values for some red observations. Conversely, there are some young founders for whom age has a positive impact on the prediction.

In Figure B.1, we report the SHAP summary plot when the algorithm is trained only using VC-backed firms, as in Section 5.4. This allows us to observe what features are most predictive of operating performance *within the set of VC-backed firms*. There is a very strong overlap not only in the most important variables between our main algorithm and this one, but also in the relationship between the value of a feature and the impact on the prediction. However, some notable differences emerge. For example, while the number of companies the founder created in the past does not appear to be first order for performance predictions using the entire test set, it is an important determinant of performance predictions among the set of VC-backed firms. Interestingly, VC-backed serial entrepreneurs are predicted to perform *worse* than other VC-backed founders. In

addition, among the set of VC-backed founders, those who answer that their motivation to start the business is a new idea are predicted to perform worse on average than those who do not.

In Figure B.2, we show the SHAP summary plot for our predictive model of our first measure of home runs, i.e., deals associated with a successful exit such as a later round of VC financing, an acquisition, or an IPO. We document in Section 5.2 that our model shows promise to identify these low probability events. The model suggests that innovative and B2B businesses are more likely to hit home runs, as are those located in Paris and those that operate in the high-tech industry. Businesses with entrepreneurs who have co-founders, those who bring in customers from their prior job, and those with a graduate degree are also more likely to hit home runs.

Finally, Figure B.3 shows the SHAP summary plot for our second measure of home runs, $top5va_5$, a dummy equal to one for firms in the top 5% of their cohort in terms of operating performance.

6 Understanding Venture Capitalists' Decision Making through Algorithmic Predictions

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.

6.1 VC-backed vs. 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 10 reports the probability densities of founders' ages, gender, education level, and geographic location for VC-backed and algorithm-selected entrepreneurs.

[Insert Figure 10 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 almost double the proportion of VC-backed female founders.

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%. The results show that the algorithmic policy selects more founders with a graduate degree relative to VCs. Note that if we decrease the selectivity of the algorithm to $s = 15\%$, we find the opposite result: VCs back a higher proportion of founders with a graduate degree relative to the algorithmic policy. This is the only change we observe in Figure 10 as we decrease the selectivity of the algorithmic policy; all other results are similar.

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 10 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 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% (8%) of VC (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 91% 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.

6.2 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.[31] We therefore use our Centaur 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. Figure 11 reports the results graphically.

[Insert Figure 11 here]

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 Centaur 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.

[31] For instance, https://www.wearesista.com/ is a French special interest group promoting public policies aimed at leveling the playing field in VC finance between male and female entrepreneurs.

6.3 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.[32]

Model performance. Our predictive model predicts VCs' investment decisions reasonably well. Figure 12 shows that our model has an area under the curve (AUC) of .78. 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 78%.

[Insert Figure 12 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 almost half of the signal 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.

Model interpretability. In Figure 13, we report the SHAP summary plot for our predictive model of VCs' investment decisions. Figure 13 indicates which types of entrepreneurs are more likely to be VC-backed. We note a strong overlap in the most relevant features that predict VCs' decisions and those that predict operating performance, lending further support to the idea that the model that predicts VC-backed status is similar to the one that predicts operating performance. We find that VCs tend to finance entrepreneurs who were not self-employed independent workers prior to creating the business, those who hire more workers and hire administrative and accounting

[32] 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.

services. Interaction effects are important especially for the founder's age. When a founder lists "new idea" as the motivation for starting the business, this increases our algorithm's prediction that she will be VC-backed. Note that this feature also had a positive SHAP value when predicting home runs, consistent with VCs trying to identify future home runs. Interestingly, Figure B.1 shows that within the set of VC-backed firms, founders who listed "new idea" as a motivation on average are associated with *lower* performance predictions. Female founders are associated with either very low or very high predictions of being VC-backed. This finding suggests that interactions between female and other features, such as industry for example (Hebert, 2020), are important.

[Insert Figure 13 here]

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, $\textit{VC-backed}_i$, on our algorithmic predictions of VCs' decisions:

$$\textit{VC-backed}_i = \beta_0 + \widehat{h}(X_i)\beta_1 + \epsilon_i \tag{4}$$

Table 5 confirms that our model of VCs' decisions indeed performs well. The model's estimates are correlated with VCs' actual decisions (column 1) and imply that a firm in the third quartile of our VC-backed predictions is 1.4 p.p. more likely to be VC-backed compared to a firm in the first quartile, a 145% increase relative to the mean.[33] We then regress VCs' actual decisions on our performance predictions $\widehat{m}(X_i)$ using our two home run measures ($top5va_5$ and successful exits):

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

If VCs did not care about portfolio firms' performance, we would expect performance predictions to not load significantly ($\beta_1 = 0$ in Equation (5)). This is not the case: Column 2 shows that predicted operating performance correlates with VC-backed status. A firm in the third quartile of our best performer predictions is 0.24 p.p. more likely to be VC-backed compared to a firm in the first quartile, a 25% 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

[33]There are 26,776 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).

significant once we control for performance predictions by estimating:

$$VC\text{-}backed_i = \beta_0 + \widehat{h}(X_i)\beta_1 + \widehat{m}(X_i)\beta_2 + \epsilon_i \tag{6}$$

Columns 4 through 6 of Table 5 show that there remains significant predictability in VCs' behavior even when controlling for algorithmic predictions of venture performance. The coefficient on our predictions of VCs' decisions, β_1, remains virtually unchanged from column 1 to column 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. The coefficients on predicted performance are negative in these specifications, suggesting that VCs' decisions negatively correlate with performance once we control for predictions of VC-backed status.

[Insert Table 5 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 6 contains the results.

[Insert Table 6 here]

The results in Table 6 imply that entrepreneurs are more likely to be casting errors when they are male, without a graduate degree, older, optimistic (i.e., they expect to grow and hire), motivated by new ideas and have successful peer entrepreneurs. Entrepreneurs who are casting errors are also more likely to innovate and be serial entrepreneurs.

6.4 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 6.1). 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. Specifically, 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.

[Insert Figure 14 here]

We start with Figure 14, which plots the performance distribution of firms in the training set according to the characteristics of interest. We note two things. First, the distribution of firm performance is shifted toward higher performances for male entrepreneurs relative to female entrepreneurs, Paris-based versus not Paris-based firms, entrepreneurs with a graduate degree versus no graduate degree, and high-tech versus non high-tech firms.[34] Second, the average performance of entrepreneurs whose characteristics are more common among VC-backed firms relative to algorithm-selected ones is higher than that of their comparison group. Entrepreneurs in the high-tech industry are an exception, though their performance has higher variance than non high-tech firms. These findings suggest that entrepreneurs with these characteristics are more likely to reach higher levels of performance and to end up in the right tail of the performance distribution.

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), 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:

$$\frac{Pr(X_i \mid P)}{Pr(X_i \mid -P)} \tag{7}$$

[34] Figure 14 drops firms that no longer exist or have negative value added at age 5. Since many Paris-based and high-tech firms are in this case, the average performance of these firms looks larger in Figure 14 than it would without survivorship bias. In particular, the average performance of Paris-based firms is lower than that of non-Paris based firms in the entire distribution.

Figure 15 plots the representativeness of each of the characteristics of interest across percentiles of the performance distribution. We find that all of these characteristics are representative of the best performing firms, that is, their representativeness ratio is higher than one in the right tail of the performance distribution. Moreover, the representativeness of each of these characteristics trends upward from the 50th to the 99th percentiles of the performance distribution. One notable fact is that male entrepreneurs are representative of all percentiles above the 57th.

[Insert Figure 15 here]

In Table 7, 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 7 confirms the finding in Figure 15 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 7 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 7 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 the next section.

6.5 Do stereotypes bias VCs' decisions?

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 6.3 and 6.4 motivates our exploration of bias: Figure 12 shows that a large part of the signal that predicts VCs' investment decisions lies in three demographic features, and Table 7 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.[35] 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

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

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 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 8 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

[35] 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 6.4 that VCs select entrepreneurs representative of the best performing firms, we train the estimators $\widehat{m}(\cdot)$ and $\widehat{m}_{simple}(\cdot)$ to predict $top5va_5$, a dummy variable equal to one for firms in the top 5% of their cohort's operating performance. We report results using observations in our test set, the 2010 cohort of entrepreneurs.

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 ($top5va_5$). 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 .0375, translating to a shift of about 0.26 p.p. in the probability of being VC-backed, which represents an 81% increase relative to the baseline average.[36] In Columns 3 to 8, we test other simple models focusing on one personal characteristic in isolation.

[Insert Table 8 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 9), we find that female entrepreneurs are 0.22 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, an 81% 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, 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.[37] 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

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

[37] 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.

model uses features that capture the entrepreneurs' optimism: whether they want to grow and whether they expect to hire over the next 12 months. Controlling for performance predictions, VCs are 58% more likely to back an entrepreneur in the third quartile of the "optimism" distribution relative to an entrepreneur in the first quartile. In Column 10, we find that serial entrepreneurs are 48% 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 8 shows that VCs overweight the new ventures' location and industry. Column 2 shows that Paris-based firms are on average 0.46 p.p. more likely to be VC-backed compared to what performance would imply, which represents a 144% 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 185% 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.

7 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 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

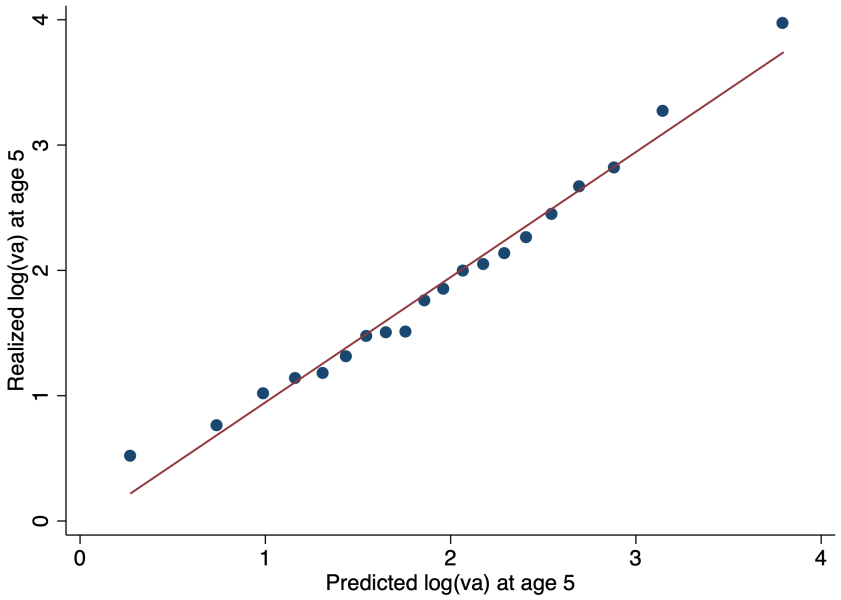

Figure 1: Algorithm Performance: All New Firms in Test Set. This figure shows the average observed performance (y-axis) across 20 bins of predicted performance (x-axis) among all new firms in the 2010 test set. The performance measure is the log value added at age 5. The predictive model was trained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

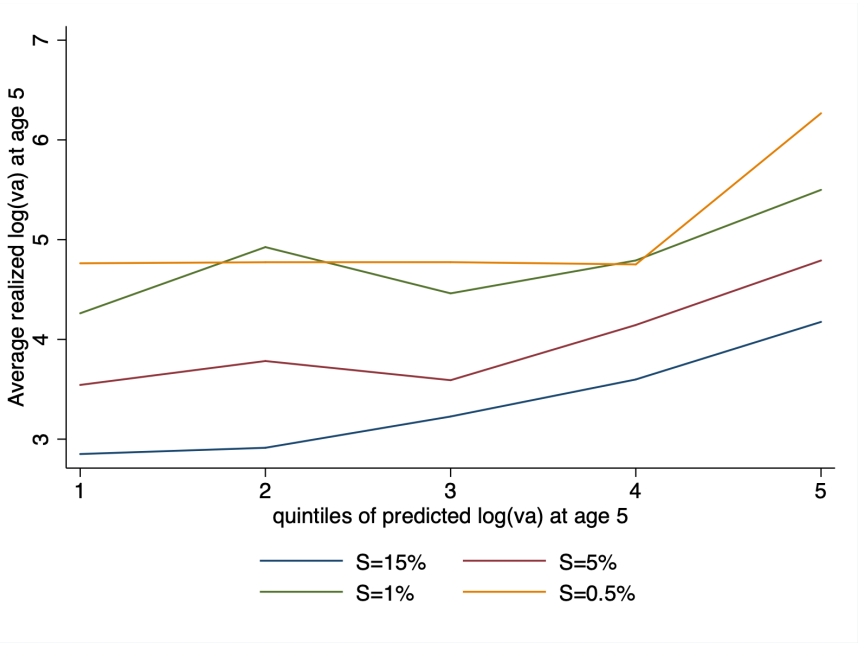

Figure 2: Algorithm Performance: Algorithm-selected New Firms in Test Set for Various Selectivity Thresholds. This figure shows the average observed performance (y-axis) across five quintiles of predicted performance (x-axis) for various selectivity thresholds among all new firms in the 2010 test set. The performance measure is the log value added 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.

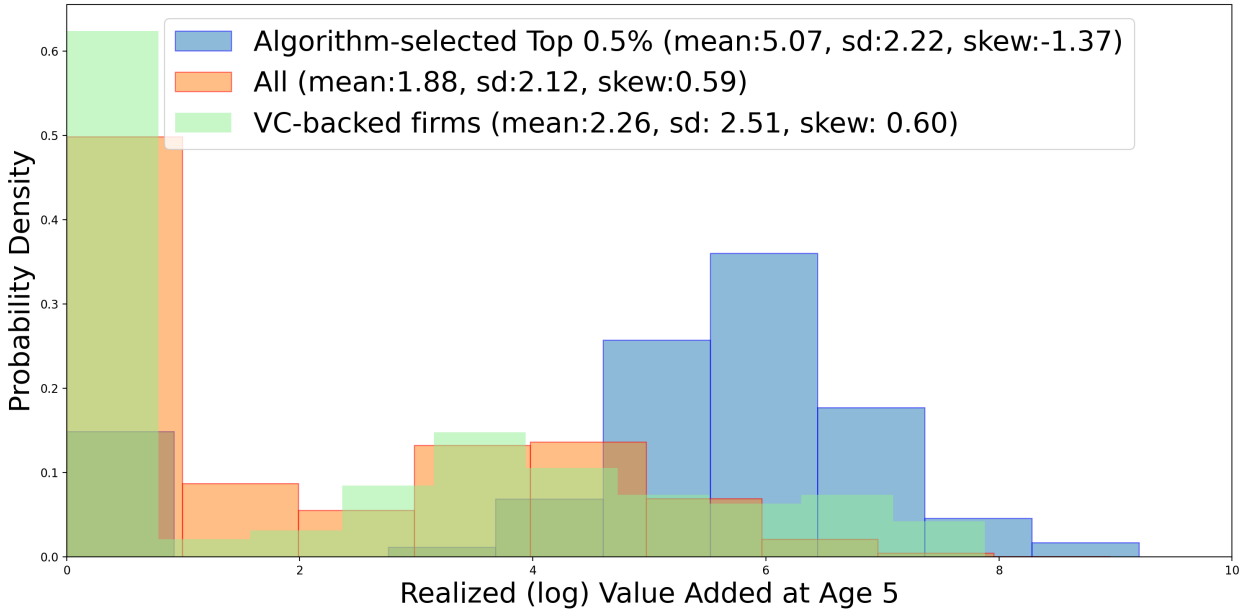

Figure 3: Realized Performance of Ventures in Test Set when the Algorithm Predicts the log of Value Added at Age 5. This figure shows the probability density of firm performance for all firms in the 2010 cohort (our test set) as well as the breakdown for VC-backed firms and for algorithm-selected firms using 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. We report the mean, standard deviation and skewness of value added at age 5 (log) for each group.

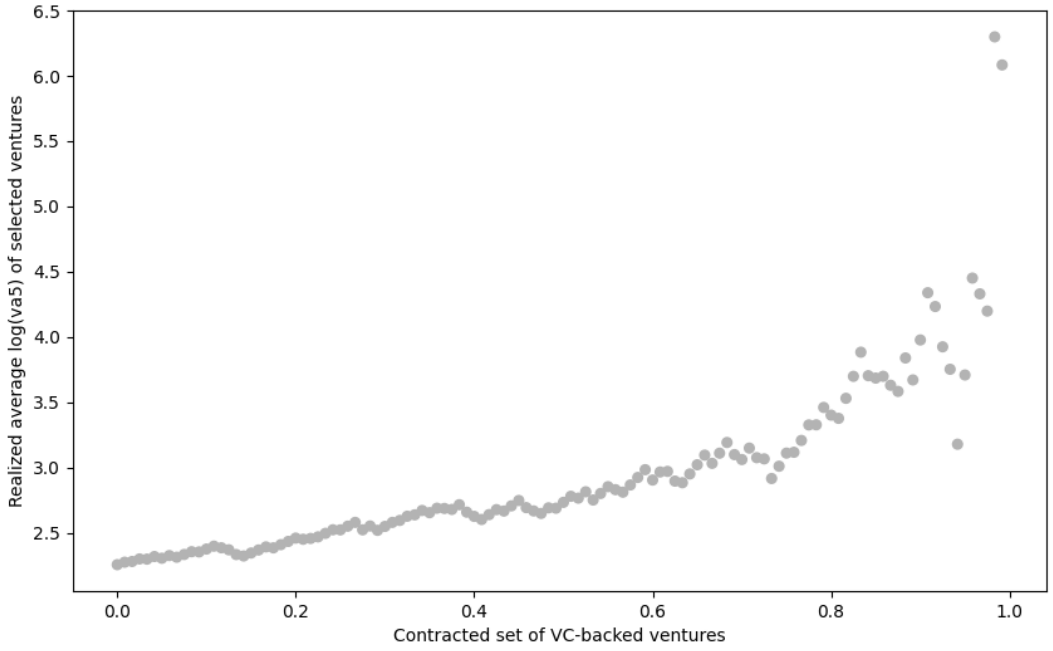

Figure 4: Potential Performance Gains: Contracting the Set of VC-backed Firms. This figure shows the average of portfolio firms' performance (log of value added at age 5) if VCs were to drop one by one the firms with the lowest algorithm-predicted performance. The origin represents the status quo: it includes the full set of VC-backed firms in the test set and their observed average performance at age 5. The fraction of portfolio firms dropped from the set of VC-backed firms is shown on the x-axis. The predictive model was trained on the sample of all firms in the 1998, 2002 and 2006 cohorts. Results are shown for the test set (the 2010 cohort of entrepreneurs.)

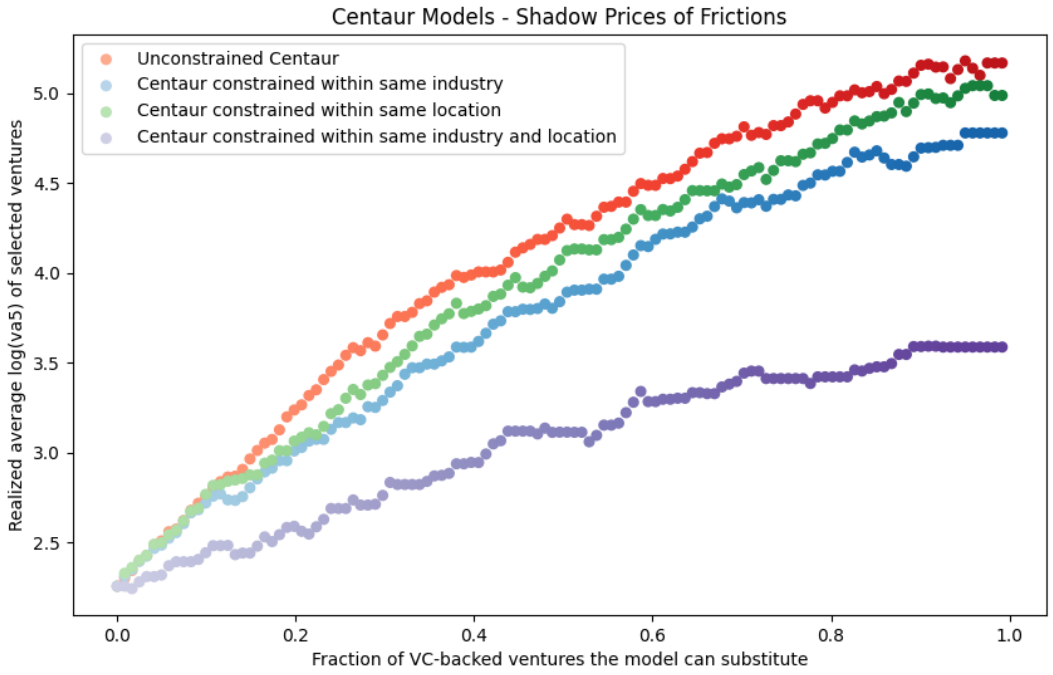

Figure 5: Potential Performance Gains: Centaur Models. This figure shows the average realized performance at age 5 (the log of value added) for several Centaur models that replace VC-backed firms that are predicted to become poor performers with firms that are predicted to become good performers by the 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 Centaur model. Each line below it represents the performance of a Centaur 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).

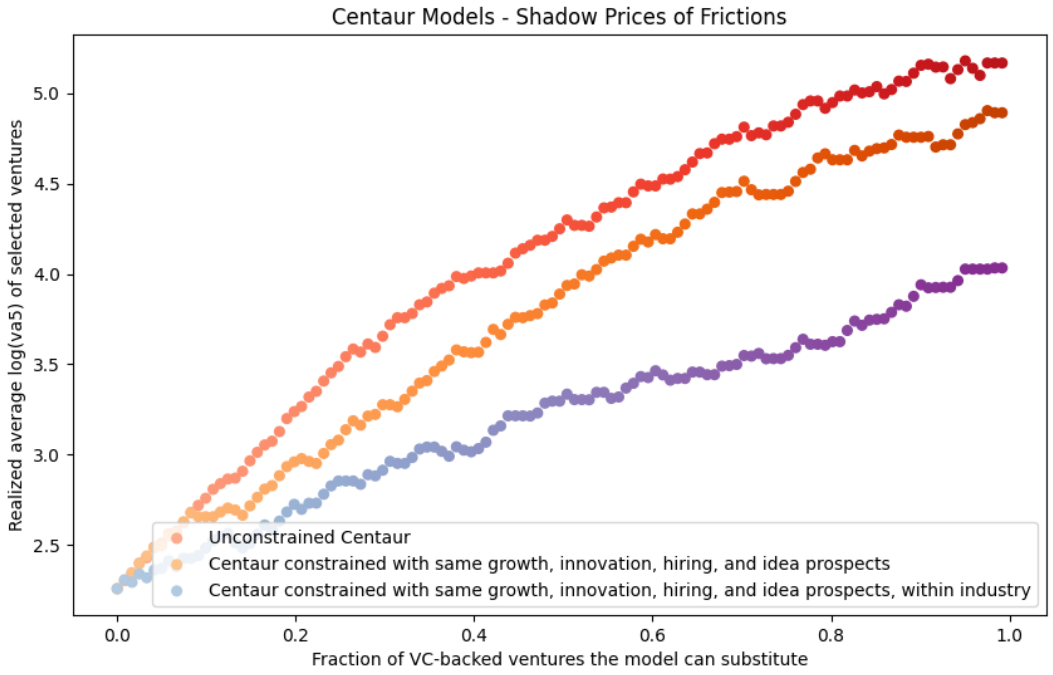

Figure 6: Centaur Models Restricted to VC-prone Ventures. This figure shows the average realized performance at age 5 for several Centaur models that replace VC-backed firms that are predicted to become poor performers with firms that are predicted to become good performers by the algorithm. The origin 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 Centaur model. The orange line represents the performance of a Centaur 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 Centaur model (same growth prospects, expectation to hire, innovate and motivated by a new idea). The purple line further restricts the Centaur model to selecting a firm in the same industry as the firm it drops.

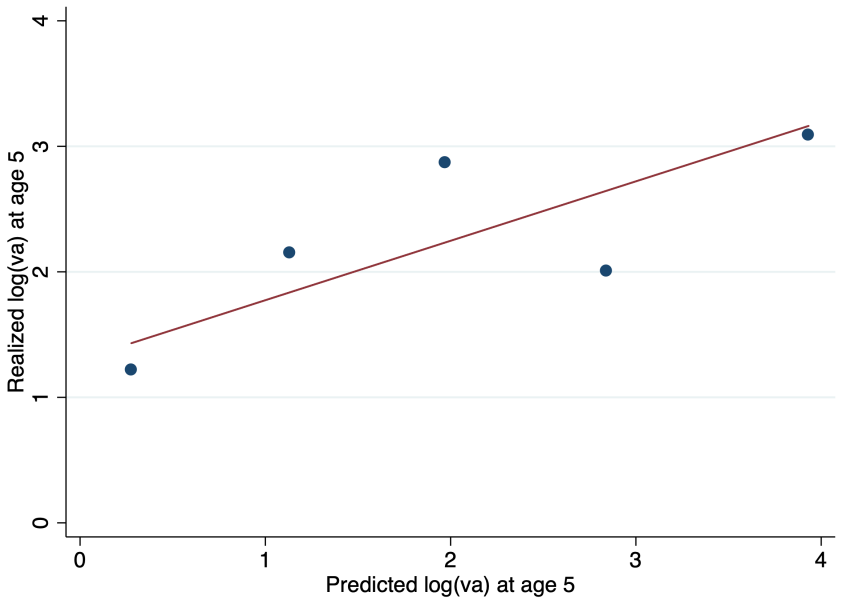

Figure 7: Algorithm Performance: VC-backed Firms in Test Set. This figure shows the average observed performance (y-axis) across 5 bins of predicted performance (x-axis) for the VC-backed firms in our 2010 cohort (our test set). The predictive model is trained on the sample of all new firms in the 1998, 2002, and 2006 cohorts.

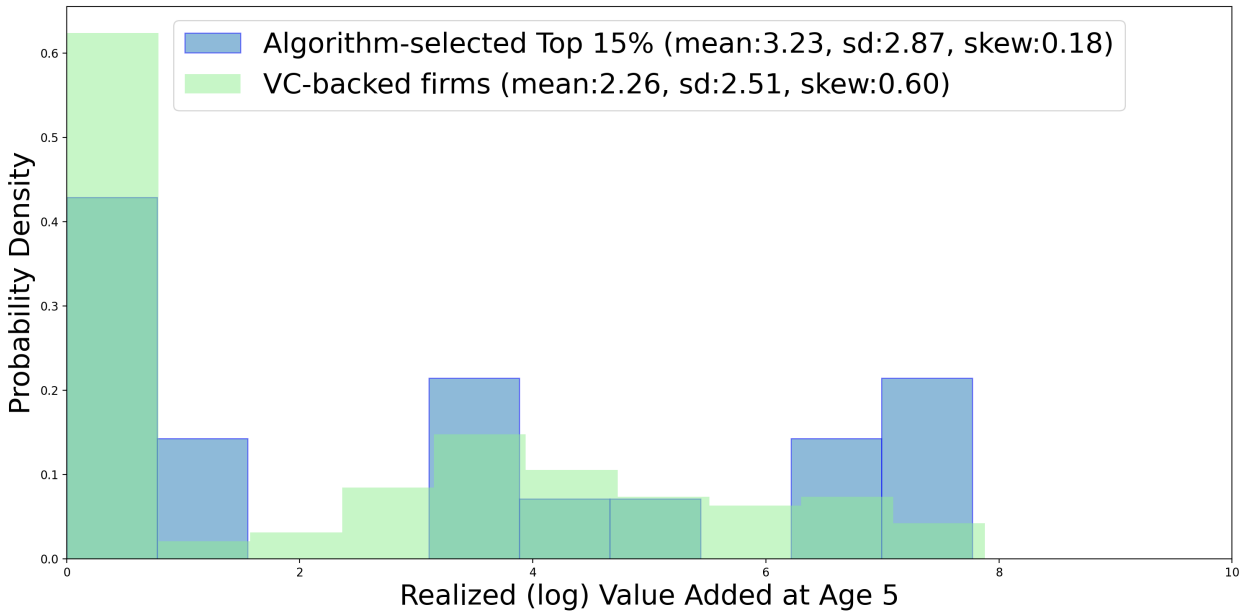

Figure 8: Realized Performance of Ventures in Test Set when the algorithm is Trained on VC-backed Firms Only. This figure shows the probability density of firm performance for the VC-backed firms in the 2010 cohort. We train a predictive model using VC-backed firms only and set the algorithmic financing policy threshold at $s = 15\%$. The predictive model is trained on the sample of VC-backed firms in the 1998 and 2002 cohorts. We report the mean, standard deviation and skewness of value added at age 5 (log) for each group.

Figure 9: 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. 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.

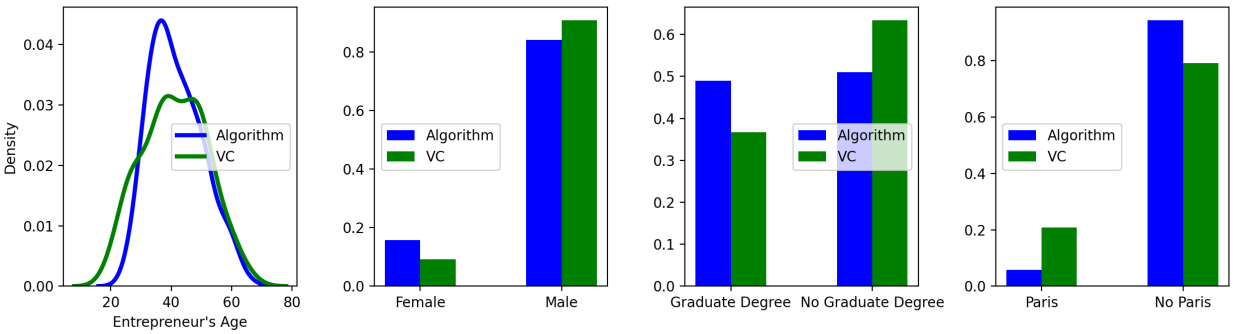

Figure 10: Entrepreneur Demographics for VC-backed and Algorithm-selected Ventures. This figure shows the probability densities of founders' ages as well as the breakdown of entrepreneurs' gender, education level and geographic location in the 2010 cohort (our test set) for VC-backed and algorithm-selected 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.

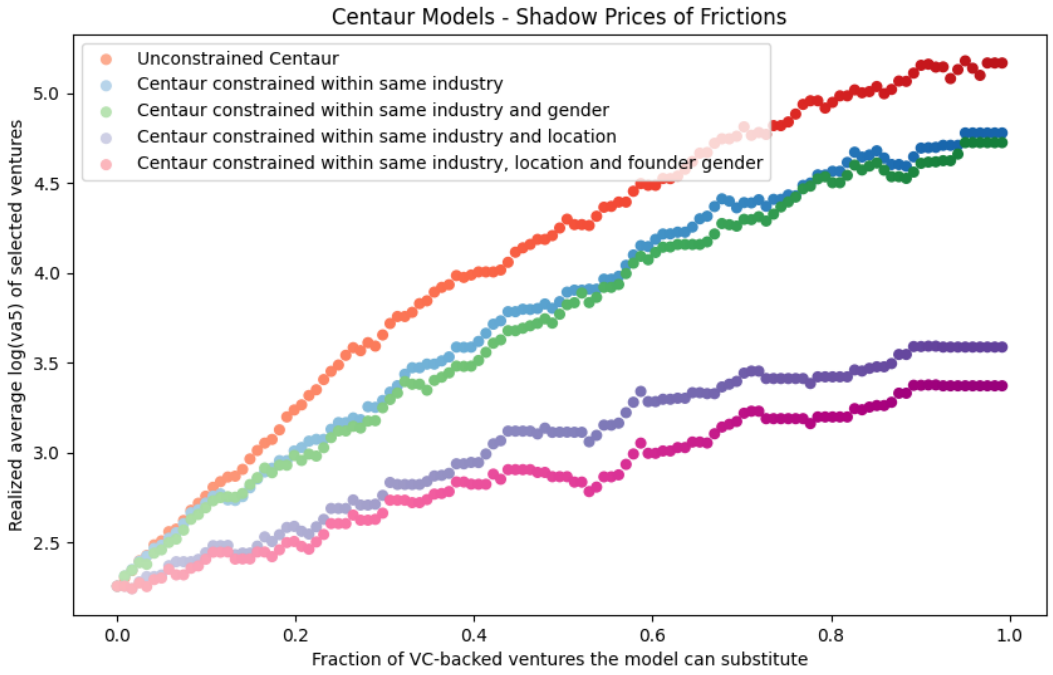

Figure 11: Centaur Models with Gender Constraints. This figure shows the average realized performance at age 5 for several Centaur models that replace VC-backed firms that are predicted to become poor performers with firms that are predicted to become good performers by the algorithm. The origin represents 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 Centaur model. Each line below it represents the performance of a Centaur model constrained to replace VC-backed firms with firms that are in the same industry (in blue), the same industry and gender (in green), the same industry and location (in purple), or the same industry, location and gender (in pink).

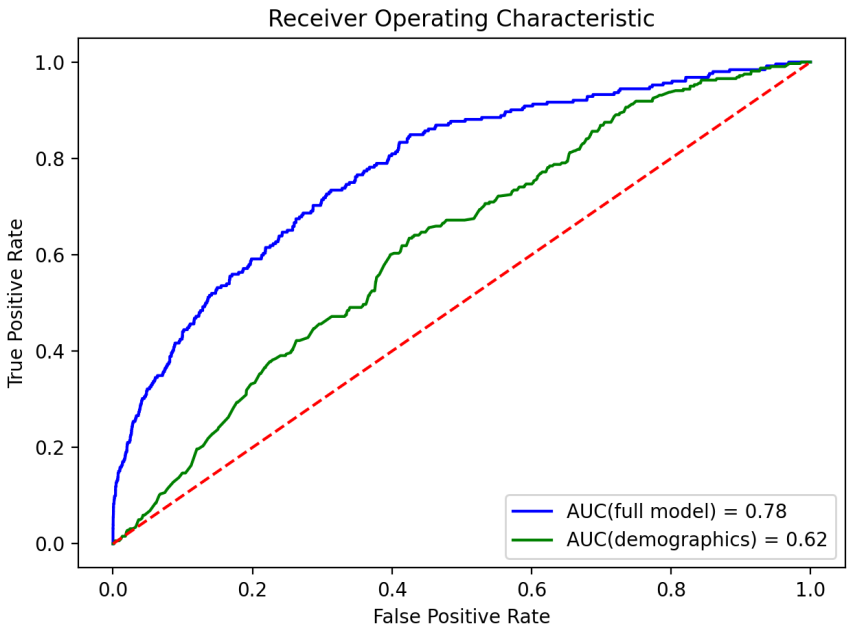

Figure 12: Area Under the Curve (AUC) of a Predictive Model of VCs' Decisions. This figure presents the ROC and AUC of a predictive model of VCs' decisions. The AUC of .78 for the full model implies that for two randomly picked ventures, one VC-backed and one not, the odds that our model assigns a higher probability of being VC-backed to the one that is indeed VC-backed is 78%. We also report the ROC and AUC of a model that only includes entrepreneurs' demographic features (age, gender and education).

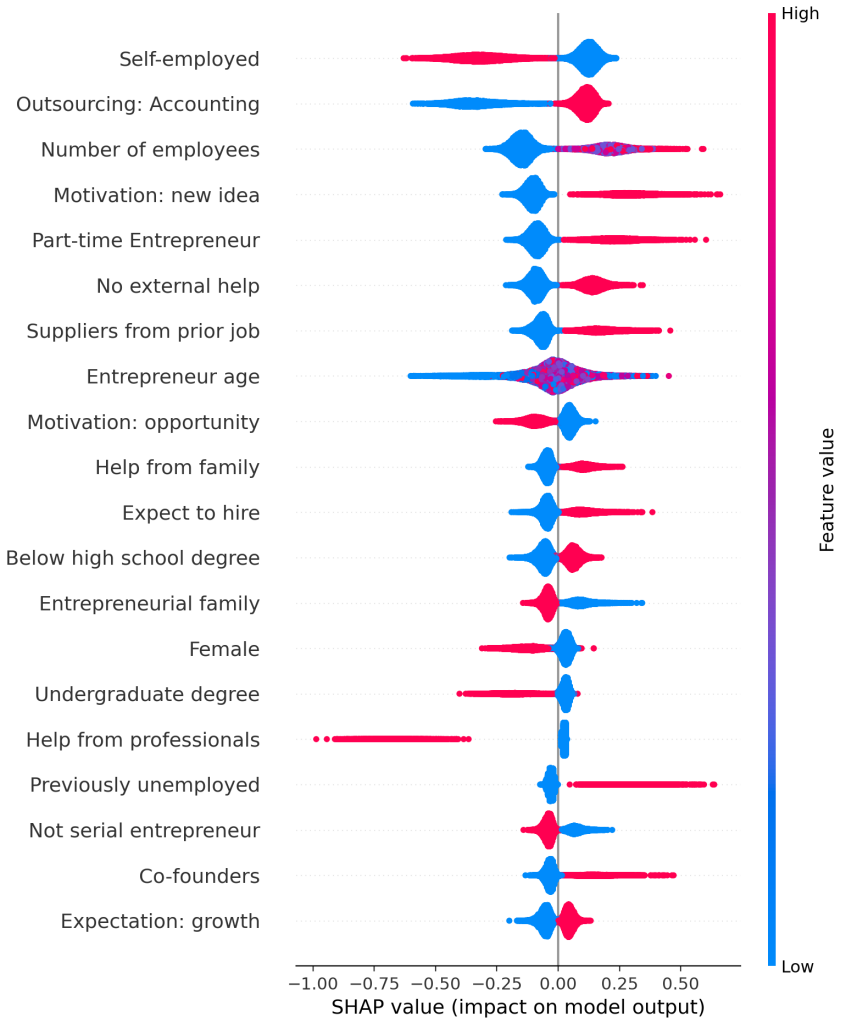

Figure 13: SHAP Values of Most Important Input Features to Predict VC backing. This figure reports the SHAP values for the top-20 features that are most important in predicting whether a firm will receive VC financing. The predictive model is trained on a random sample of all new firms in the 1998, 2002, 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.

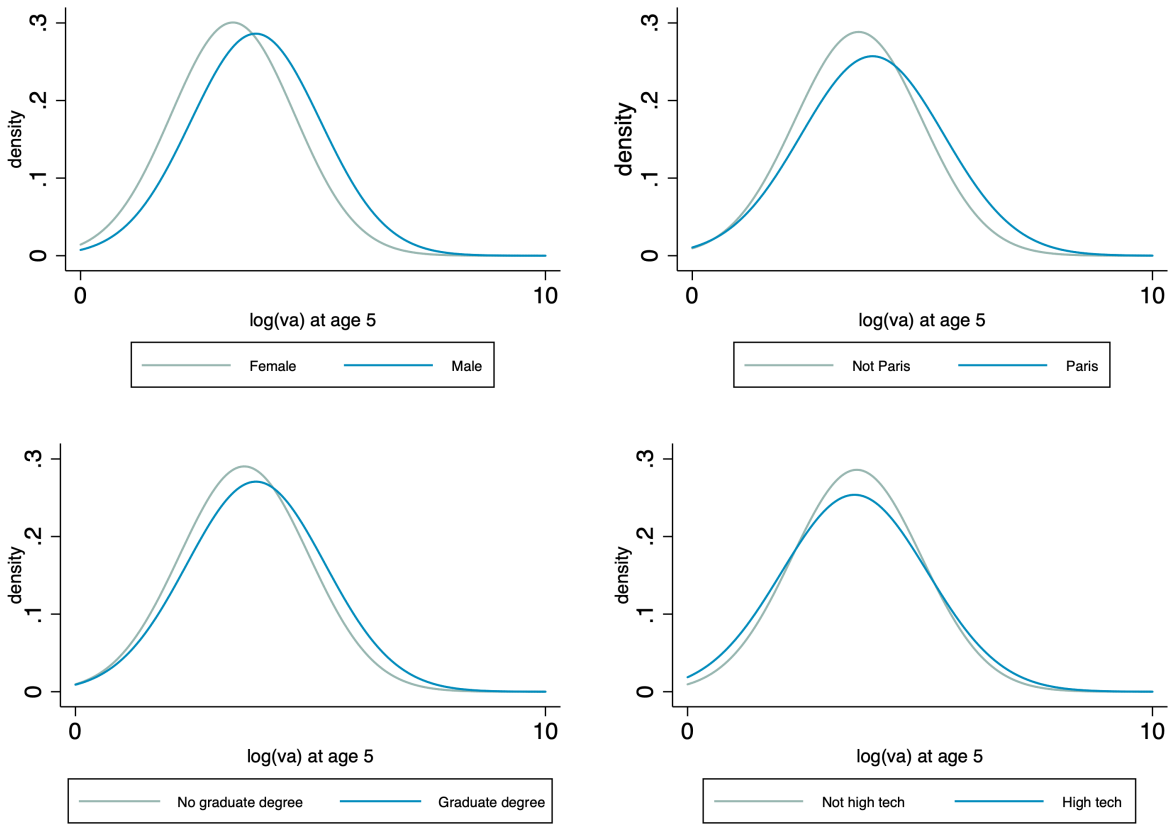

Figure 14: Distribution of Predicted Performance for Entrepreneurs Whose Characteristics are Overweighted or Underweighted by VCs. This figure superimposes the distributions of firm performance (log of value added at age 5) for firms with male vs. female entrepreneurs in the top left panel, firms in Paris vs. those outside of Paris in the top right panel, firms with entrepreneurs with a graduate degree vs. those without a graduate degree in the bottom left panel, and firms in the high-tech vs. those not in the high-tech industry in the bottom right panel. The sample includes surviving firms in the training set that have a positive value added at age 5.

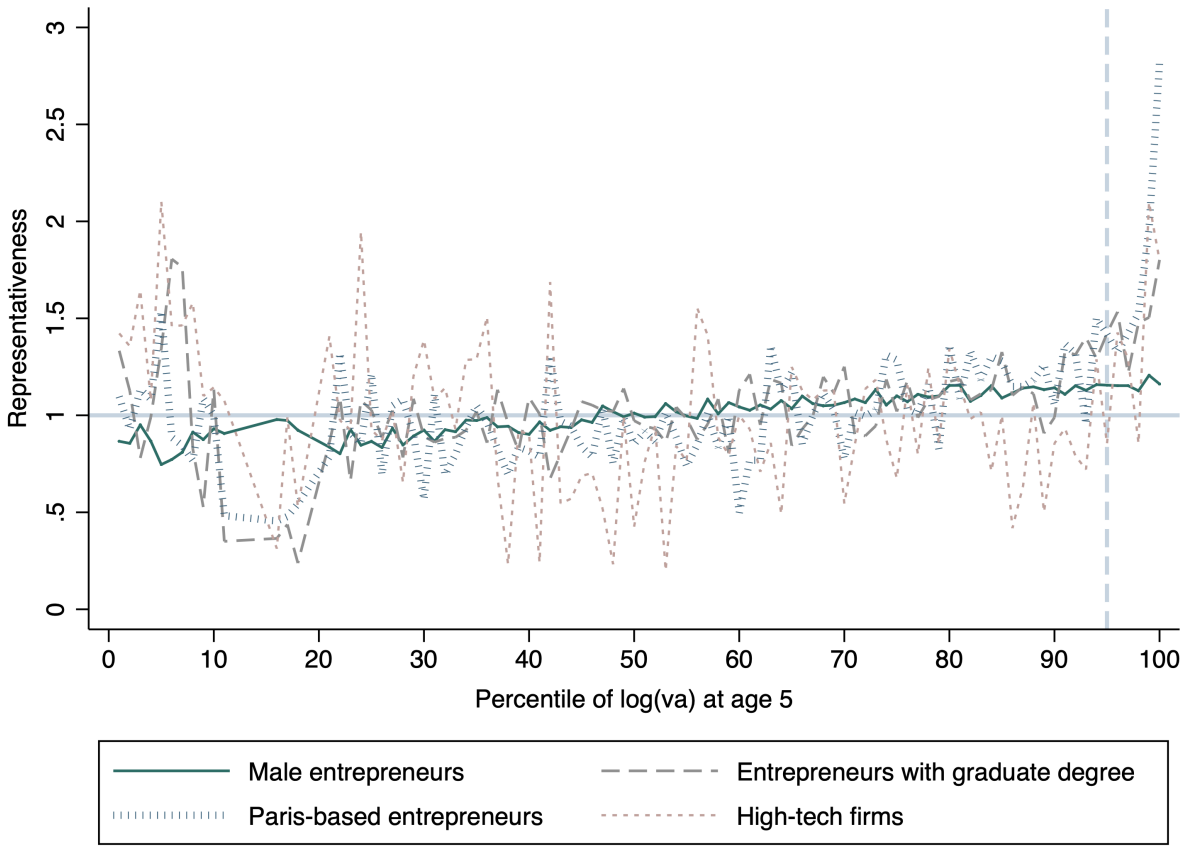

Figure 15: Representativeness of Certain Characteristics Overweighted by VCs. This figure shows the representativeness of selected entrepreneur and firm characteristics for each percentile P of the performance distribution relative to the rest of the distribution $-P$: $Pr(X_i \mid P)/Pr(X_i \mid -P)$. The sample includes firms in the 1998, 2002, and 2006 cohorts. We assign a zero as the (log) value added at age 5 of firms that do not survive or have a negative value added at age 5.

Table 1: Summary Statistics: Entrepreneur and Venture Characteristics. This table reports summary statistics for the outcome measure (Value Added 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) value added at age 5 of firms that do not survive or have a negative value added at age 5. Alive at Age 5 is presented as an alternative outcome measure. 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 B describes the variables in the entrepreneur survey.

		Training						Test					
	Variable	Mean	SD	p50	p90	p99	N	Mean	SD	p50	p90	p99	N
Outcomes													
	Value Added at Age 5 (log)	1.80	2.07	0	4.77	6.56	85658	1.88	2.12	1	4.93	6.56	37853
	Value Added at Age 5	46.09	716.16	0	117	706	85658	47.15	493.67	2	136.886	707.057	37853
	Alive at Age 5	0.62	0.48	1	1	1	85658	0.66	0.48	1	1	1	37853
Demographics													
	Entrepreneur's Age	37.79	10.13	37	52	63	83500	39.74	10.66	39	54	66	37853
	Female	0.29	0.45	0	1	1	83510	0.28	0.45	0	1	1	37853
	Entrepreneur's Nationality (FR)	0.86	0.34	1	1	1	85658	0.92	0.27	1	1	1	37853
	Entrepreneurial Family	0.68	0.47	1	1	1	81743	0.71	0.46	1	1	1	37853
Professional Background													
	Self-employed	0.37	0.48	0	1	1	85658	0.32	0.46	0	1	1	37853
	Previously Employed	0.47	0.50	0	1	1	85658	0.55	0.50	1	1	1	37853
	Part-time Entrepreneur	0.19	0.40	0	1	1	81487	0.21	0.41	0	1	1	37853
	Same Prior Industry	0.52	0.50	1	1	1	81704	0.61	0.49	1	1	1	37853
	Serial Entrepreneur	0.27	0.45	0	1	1	85658	0.29	0.46	0	1	1	37853
	Previously Employed in Small Firm	0.40	0.49	0	1	1	85658	0.43	0.49	0	1	1	37853
	Previously Inactive	0.10	0.30	0	0	1	85658	0.00	0.00	0	0	0	37853
	Below High School Degree	0.44	0.50	0	1	1	85658	0.41	0.49	0	1	1	37853
	Undergraduate Degree	0.16	0.37	0	1	1	85658	0.26	0.44	0	1	1	37853
	Graduate Degree	0.14	0.34	0	1	1	85658	0.15	0.35	0	1	1	37853
	Grande Ecole	0.04	0.21	0	0	1	34260	0.06	0.24	0	0	1	37853
	Completed Required Training	0.21	0.41	0	1	1	85658	0.21	0.41	0	1	1	37853
Motivation and Expectations													
	Expectation: Growth	0.53	0.50	1	1	1	85658	0.42	0.49	0	1	1	37853
	Expectation: Sustain	0.27	0.44	0	1	1	85658	0.39	0.49	0	1	1	37853
	Expectation: Rebound	0.07	0.25	0	0	1	85658	0.08	0.28	0	0	1	37853
	Motivation: Peer Entrepreneurs	0.11	0.31	0	1	1	83695	0.09	0.28	0	0	1	37853
	Expect to Hire	0.24	0.43	0	1	1	85658	0.26	0.44	0	1	1	37853
	Motivation: New Idea	0.18	0.39	0	1	1	83695	0.16	0.37	0	1	1	37853
	Motivation: Opportunity	0.33	0.47	0	1	1	83695	0.44	0.50	0	1	1	37853
	Innovation	0.39	0.49	0	1	1	85658	0.47	0.50	0	1	1	37853

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	1	1	85658	0.08	0.28	0	0	1	37853
	Marseille-based	0.02	0.14	0	0	1	85658	0.03	0.18	0	0	1	37853
	Lyon-based	0.02	0.13	0	0	1	85658	0.02	0.13	0	0	1	37853
	Bordeaux-based	0.02	0.14	0	0	1	85658	0.02	0.13	0	0	1	37853
	Business Services Industry	0.16	0.36	0	1	1	85658	0.14	0.35	0	1	1	37853
	Health and Social Work Industry	0.04	0.20	0	0	1	85658	0.04	0.19	0	0	1	37853
	Construction Industry	0.18	0.39	0	1	1	85658	0.17	0.37	0	1	1	37853
	High tech Industry	0.03	0.18	0	0	1	85658	0.03	0.18	0	0	1	37853
	Energy Industry	0.00	0.02	0	0	0	85658	0.03	0.16	0	0	1	37853
	B2B	0.33	0.47	0	1	1	85658	0.33	0.47	0	1	1	37853
	B2C	0.58	0.49	1	1	1	85658	0.59	0.49	1	1	1	37853
	International Customers	0.07	0.25	0	0	1	85658	0.05	0.21	0	0	1	37853
	Local Customers	0.48	0.50	0	1	1	85658	0.55	0.50	1	1	1	37853
	Domestic Customers	0.14	0.35	0	1	1	85658	0.15	0.36	0	1	1	37853
Venture Organization													
	Co-founders	0.12	0.32	0	1	1	85658	0.14	0.35	0	1	1	37853
	Outsourcing: Accounting	0.63	0.48	1	1	1	82975	0.73	0.44	1	1	1	36466
	Number of Employees	1.59	1.52	1	3	8	85658	1.63	1.58	1	3	9	36466
	10+ Clients	0.58	0.49	1	1	1	85658	0.59	0.49	1	1	1	37853
	Number of Paid Managers	0.15	0.46	0	1	2	85658	0.18	0.43	0	1	2	36466
	Customers from Prior Job	0.30	0.46	0	1	1	85658	0.27	0.44	0	1	1	37853
	Suppliers from Prior Job	0.22	0.42	0	1	1	85658	0.21	0.41	0	1	1	37853
	Help from Professionals	0.03	0.18	0	0	1	85658	0.10	0.30	0	0	1	37853
	Help from Family	0.27	0.44	0	1	1	85658	0.17	0.38	0	1	1	37853
	No External Help	0.40	0.49	0	1	1	85658	0.27	0.44	0	1	1	37853
Financial Characteristics													
(not included as input features)	Total Assets	142.37	5179.29	24	143	966	42241	392.76	11587.65	41	444.01	3870.14	36398
	Bank Loan	0.35	0.48	0	1	1	84478	0.41	0.49	0	1	1	37853
	Other Loan	0.08	0.27	0	0	1	84478	0.09	0.29	0	0	1	37853
	No Outside Financing	0.54	0.50	1	1	1	84478	0.52	0.50	1	1	1	37853
	Other Firm Financing	0.05	0.21	0	0	1	51398	0.04	0.19	0	0	1	37853
	Grant	0.21	0.41	0	1	1	84478	0.08	0.27	0	0	1	37853
	Future VC Financing	0.01	0.11	0	0	1	85658	0.02	0.15	0	0	1	37853
Industries-Locations													
	Number of Industries	-	-	-	-	-	38	-	-	-	-	-	41
	Number of Regions	-	-	-	-	-	321	-	-	-	-	-	321

	All Successful Deals				Acquisition Events			
cohort	All	VC-backed	Algorithm-selected $s = 0.5\%$ (190 firms)	Algorithm-selected $s = 1\%$ (379 firms)	All	VC-backed	Algorithm-selected $s = 0.5\%$ (190 firms)	Algorithm-selected $s = 1\%$ (379 firms)
2010	56	4	11	16	37	3	6	9

Table 2: VC-backed vs. Algorithm-selected Firms: Successful Deals. This table presents the total number of successful deals and acquisition events (which are a subset of successful deals) for firms in the 2010 cohort of the *SINE* survey, which is our test set. For all firms, regardless of whether they receive VC, *successful deal* equals 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 report algorithm-selected firms in the test set (2010 cohort) at the $= 0.5\%$ threshold. The data come from Crunchbase, Capital IQ, CB Insights, Preqin, Venture Xpert, SDC and Zephyr.

Algorithm trained on	Algorithm evaluated on					
	VA_5 (log)	VA_7 (log)	Top 5% VA_5	Top 5% VA_7	$Ebitda_5$ / $capital_0$(log)	Successful Deals
VA_5 (log)	5.07	4.86	0.55	0.55	3.09	4
VA_7 (log)	5.01	4.89	0.49	0.50	2.92	4
Top 5% VA_5	4.97	4.60	0.60	0.56	3.09	4
Top 5% VA_7	4.90	4.64	0.56	0.55	3.09	4
$Ebitda_5$ /$capital_0$(log)	4.69	4.50	0.45	0.44	2.86	3
Successful Deals	3.07	2.66	0.28	0.26	1.78	11
	Comparison: Average performance measures					
	VA_5 (log)	VA_7 (log)	Top 5% VA_5	Top 5% VA_7	$Ebitda_5$ / $capital_0$(log)	Successful Deals
All firms in test set	1.88	1.61	0.05	0.05	1.01	56
VC-backed firms	2.26	1.94	0.14	0.13	1.06	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 average performance in 5 and 7 years (rows 1 and 2), whether the firm will be in the top 5% of its cohort in terms of value added in 5 and 7 years (rows 3 and 4), the new firm's ratio of Ebitda over initial capital (row 5), and whether the firm exits through an acquisition or IPO or receives VC funding in later years (row 6). 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 ($s = 0.5\%$)	Difference ($s = 1\%$)
		Mean	SD	N	Mean	SD	N	Mean	SD	N	T-Test	T-Test
Outcomes												
	Value Added at Age 5 (log)	2.26	2.52	120	5.07	2.23	190	4.79	2.23	379	-2.81***	-2.53***
	Value Added at Age 5	140.66	432.54	120	540.51	1019.67	190	435.12	869.09	379	-399.85***	-294.47***
	Alive at Age 5	0.69	0.46	120	0.91	0.29	190	0.89	0.31	379	-0.22***	-0.20***
Founder Demographics												
	Entrepreneur's Age	41.26	10.58	120	41.68	8.70	190	41.27	8.70	379	-0.42	-0.02
	Entrepreneur's Nationality (FR)	0.94	0.24	120	0.98	0.14	190	0.98	0.12	379	-0.04	-0.04*
	Female	0.09	0.29	120	0.16	0.37	190	0.18	0.38	379	-0.07*	-0.09***
Founder Professional Background												
	Same Prior Industry	0.52	0.50	120	0.91	0.29	190	0.88	0.33	379	-0.39***	-0.36***
	Serial Entrepreneur	0.42	0.50	120	0.35	0.48	190	0.32	0.47	379	0.07	0.10*
	Previously Employed in Small Firm	0.29	0.46	120	0.37	0.48	190	0.35	0.48	379	-0.08	-0.05
	Graduate Degree	0.37	0.48	120	0.49	0.50	190	0.41	0.49	379	-0.12**	-0.04
	Grande École	0.27	0.44	120	0.11	0.31	190	0.10	0.30	379	0.16***	0.17***
Founder Motivation and Expectations												
	Expectation: Growth	0.58	0.50	120	0.52	0.50	190	0.57	0.50	379	0.06	0.01
	Motivation: Successful Peer Entrepreneurs	0.06	0.24	120	0.11	0.31	190	0.11	0.31	379	-0.05*	-0.05*
	Expect to Hire	0.51	0.50	120	0.54	0.50	190	0.56	0.50	379	-0.03	-0.05
	Motivation: New Idea	0.39	0.49	120	0.08	0.27	190	0.09	0.29	379	0.31***	0.30***
	Motivation: Opportunity	0.38	0.49	120	0.56	0.50	190	0.54	0.50	379	-0.19***	-0.17***
	Innovation	0.73	0.44	120	0.41	0.49	190	0.43	0.50	379	0.32***	0.31***
Venture Characteristics												
	Paris-based	0.21	0.41	120	0.06	0.23	190	0.05	0.22	379	0.15***	0.16***
	High-tech Services Industry	0.10	0.30	120	0.01	0.10	190	0.02	0.13	379	0.09***	0.08***
Organization												
	Outsourcing: Accounting	0.90	0.30	114	0.89	0.31	190	0.91	0.29	379	0.01	-0.01
	Outsourcing: Management	0.10	0.30	114	0.28	0.45	190	0.23	0.42	379	-0.19***	-0.14***
	Outsourcing: Logistics	0.16	0.37	114	0.25	0.44	190	0.23	0.42	379	-0.09**	-0.08*
	Number of Employees	2.37	2.87	114	6.01	4.78	190	5.41	4.47	379	-3.64***	-3.04***
Industries-Locations												
	Number of Industries	-	-	29	-	-	23	-	-	29	-	-
	Number of Regions	-	-	68	-	-	96	-	-	145	-	-
Financial Characteristics (not included in input features)	Total Assets (k euros)	660.58	2233.84	118	584.20	1805.17	187	626.75	2166.54	375	76.38	33.83

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) value added at age 5 of firms that do not survive or have a negative value added at age 5. The data come from the entrepreneur survey (SINE) conducted by the French Statistical Office, tax files from the Ministry of Finance and the firm registry (SIRENE). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	VC-backed					
	(1)	(2)	(3)	(4)	(5)	(6)
$\widehat{h}(X)$	2.078***			2.165***	2.132***	2.198***
	(0.0592)			(0.0625)	(0.0612)	(0.0637)
$\widehat{m}(X)_{top5va_5}$		0.0485***		-0.0302***		-0.0267***
		(0.00679)		(0.00702)		(0.00714)
$\widehat{m}(X)_{homerun}$			0.544***		-0.334***	-0.266***
			(0.0967)		(0.0979)	(0.0995)
Observations	26,776	26,776	26,776	26,776	26,776	26,776
R-squared	0.044	0.002	0.001	0.045	0.044	0.045

Standard errors in parentheses

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

Table 5: 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; $\widehat{m}(X)_{top5va_5}$ is a vector of predicted probabilities for whether a firm will be in the top 5% of its cohort in terms of operating performance 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	-0.276***
	(-8.19)
Graduate Degree	-0.149***
	(-5.74)
High Tech	0.0150*
	(1.88)
Paris-based	0.223***
	(8.80)
Entrepreneur's Age	6.153***
	(7.84)
Expectation: Growth	0.164***
	(4.74)
Expect to Hire	0.220***
	(7.71)
Motivation: New Idea	0.454***
	(16.05)
Motivation: Successful Peer Entrepreneurs	0.0503**
	(2.12)
Innovation	0.288***
	(8.70)
Serial Entrepreneur	0.392***
	(13.02)

t statistics in parentheses

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

Table 6: Which Entrepreneurs Are More Likely to be Casting Errors? We sort firms into quintiles according to their predicted performance 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.

Feature	Top 5%	Bottom 95%	Representativeness of best performers $\frac{Pr(X_i \mid \text{Top5})}{Pr(X_i \mid \text{Bottom95})}$	Representativeness of VC-backed firms $\frac{Pr(X_i \mid \text{VC-backed})}{Pr(X_i \mid \text{non-VC-backed})}$	Overreaction $\frac{\frac{Pr(X_i \mid \text{VC-backed})}{Pr(X_i \mid \text{non-VC-backed})}}{\frac{Pr(X_i \mid \text{Top5})}{Pr(X_i \mid \text{Bottom95})}}$
	(1)	(2)	(3)	(4)	(5)
Male	85%	72%	1.17	1.26	1.08
Graduate Degree	22%	15%	1.53	2.5	1.63
Grande Ecole	10%	5%	2.12	4.52	2.13
Optimism	52%	20%	2.63	2.31	.88
Serial Entrepreneur	39%	22%	1.77	1.42	.8
Paris-based	15%	8%	1.81	2.48	1.37
High tech	5%	3%	1.41	2.95	2.09

Table 7: 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 operating performance, column 1) and among the other firms (bottom 95% of operating performance, 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. Column 4 contains the representativeness of VC-backed firms along each characteristic, and column 5 contains the overreaction ratio of the percentage in column 4 over that in column 3. Operating performance 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.

Table 8: 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)
$\widehat{m}_{full}(X)$	.0293***	.0255***	.0295***	.0277***	.0264***	.0209***	.0292***	.0294***	.0243***	.0279***
	(.00332)	(.00342)	(.00334)	(.00335)	(.00336)	(.00317)	(.00332)	(.00333)	(.00358)	(.00335)
$\widehat{m}_{simple}$(personal features)		.0682***								
		(.015)								
$\widehat{m}_{simple}$(age)			-.0115							
			(.0237)							
$\widehat{m}_{simple}$(male)				.0703***						
				(.0211)						
$\widehat{m}_{simple}$(graduate degree)					.175***					
					(.032)					
$\widehat{m}_{simple}$(grande ecole)						.199***				
						(.0233)				
$\widehat{m}_{simple}$(French nationality)							.0259			
							(.0829)			
$\widehat{m}_{simple}$(relatives)								-.0275		
								(.0627)		
$\widehat{m}_{simple}$(optimism)									.032***	
									(.00867)	
$\widehat{m}_{simple}$(serial entrepreneur)										.0543***
										(.0176)
R^2	.002	.0025	.002	.0023	.0028	.0035	.002	.002	.0024	.0023
Observations	37,853	37,853	37,853	37,853	37,853	37,853	37,853	37,853	37,853	37,853

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)
$\widehat{m}_{full}(X)$	.0293***	.0284***	.0293***	.0294***	.0293***	.0289***	.0293***	.0293***	.0251***
	(.00332)	(.00332)	(.00332)	(.00332)	(.00332)	(.00332)	(.00333)	(.00333)	(.00441)
$\widehat{m}_{simple}$(Paris-based)		.282***							
		(.0645)							
$\widehat{m}_{simple}$(Marseille-based)			.691						
			(6.8)						
$\widehat{m}_{simple}$(Lyon-based)				-.147					
				(.156)					
$\widehat{m}_{simple}$(Bordeaux-based)					.423				
					(.615)				
$\widehat{m}_{simple}$(high tech)						.568***			
						(.154)			
$\widehat{m}_{simple}$(business services)							-.00777		
							(.0457)		
$\widehat{m}_{simple}$(energy)								-.00183	
								(.0132)	
$\widehat{m}_{simple}$(startup traction)									.01
									(.00686)
R^2	.002	.0025	.002	.002	.002	.0024	.002	.002	.0021
Observations	37,853	37,853	37,853	37,853	37,853	37,853	37,853	37,853	37,853

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.

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.

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.

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.

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.*

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

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

Gompers, Paul, and Josh Lerner. 1999. "An analysis of compensation in the US venture capital partnership." *Journal of Financial Economics*, 51(1): 3–44.

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 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 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."

Sterk, Vincent, Petr Sedláček, and Benjamin Pugsley. 2021. "The nature of firm growth." *American Economic Review*, 111(2): 547–79.

Stewart, Bennett. 2019. "EVA, not EBITDA: A new financial paradigm for private equity firms." *Journal of applied corporate finance*, 31(3): 103–115.

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.

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 Additional Tables and Figures

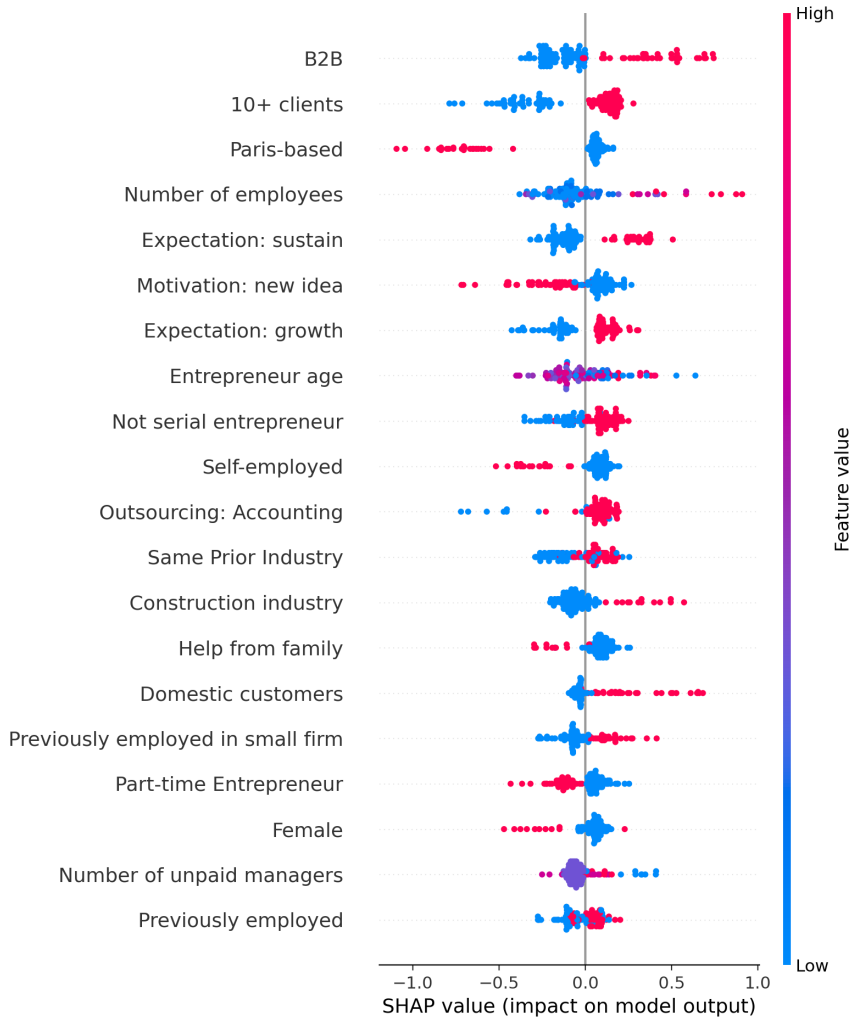

Figure B.1: SHAP Values of Most Important Input Features to Predict Operating Performance When the Model is Trained on VC-backed Firms Only. This figure reports the SHAP values for the top-20 features that are most important in predicting operating performance. Features are ranked in decreasing order 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.

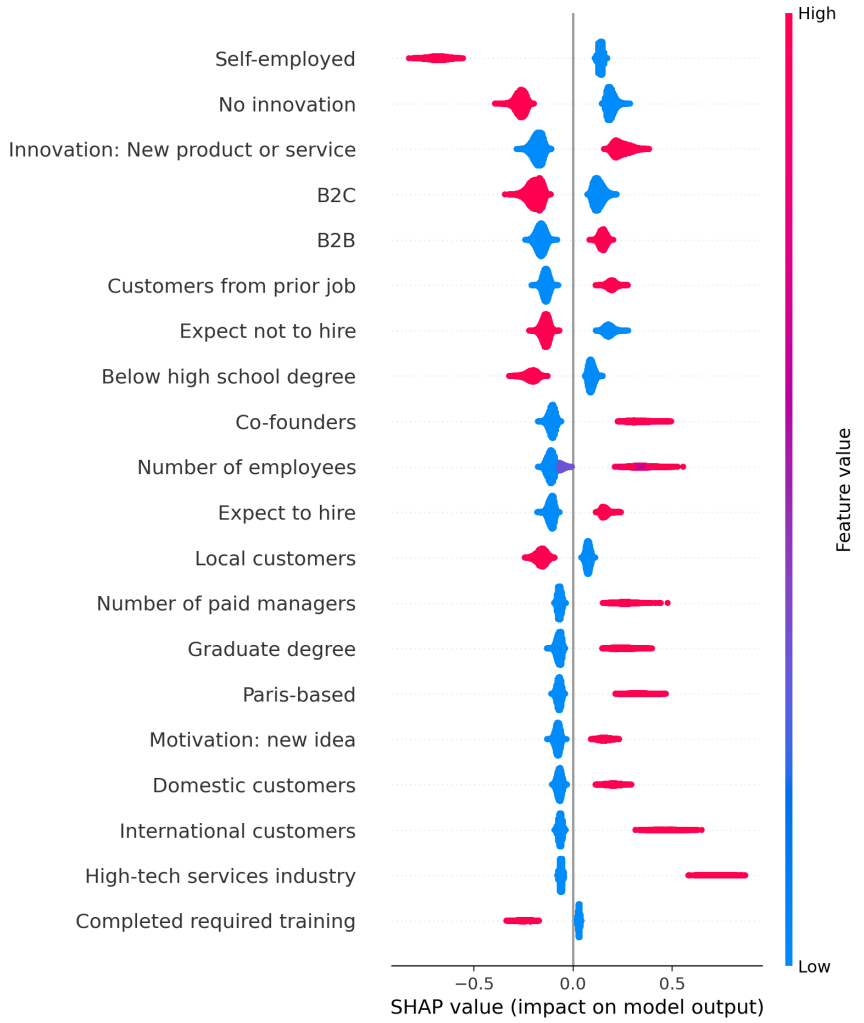

Figure B.2: SHAP Values of Most Important Input Features to Predict Home Runs. This figure reports the SHAP values for the top-20 features that are most important in predicting successful deals. The dummy variable *successful deal* is a proxy for a successful exit by the VC, i.e., a "home run." Whether 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.

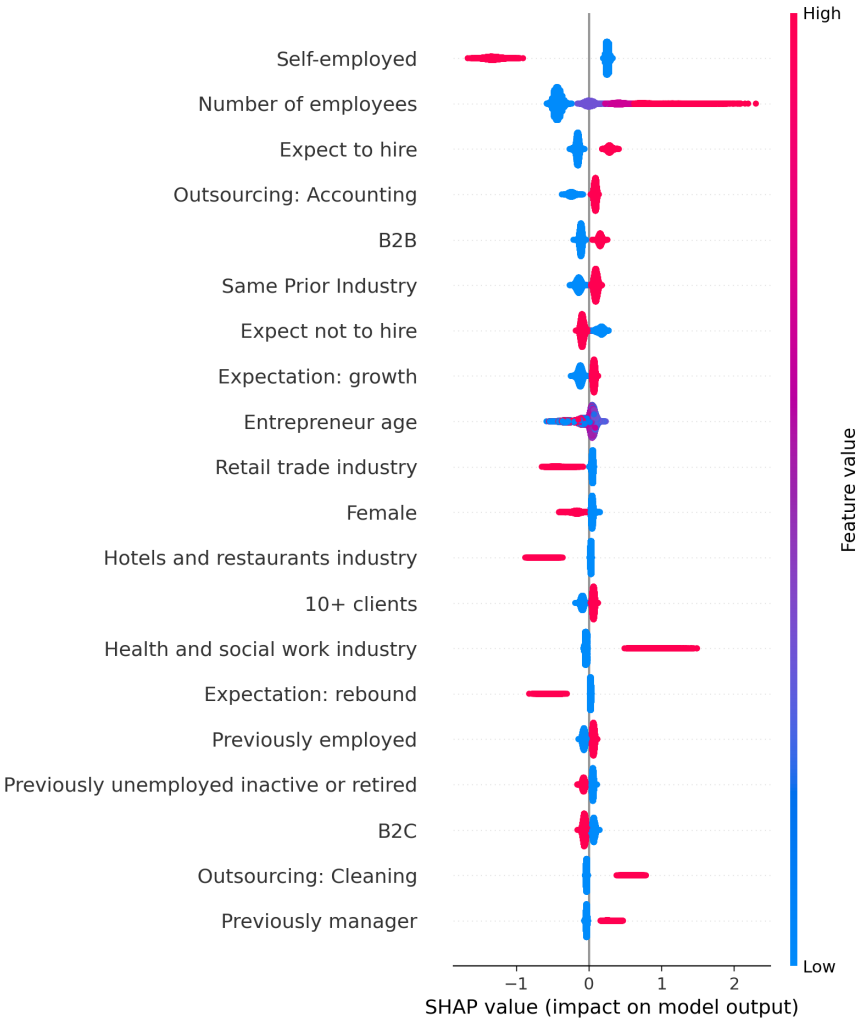

Figure B.3: SHAP Values of Most Important Input Features to Predict Best Performers. This figure reports the SHAP values for the top-20 features that are most important in predicting whether a firm will be in the top 5% of its cohort in terms of operating performance. Features are ranked in decreasing 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.

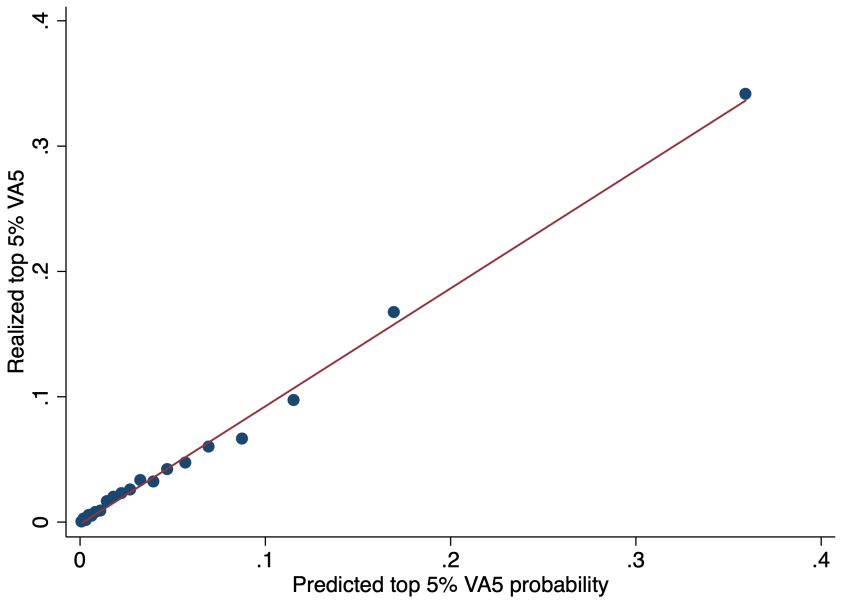

Figure B.4: Predicting Top 5% Operating Performance: All New Firms in Test Set. This figure shows the average realized fraction of firms in the top 5% of operating performance (value added at age 5) on the y-axis across 20 bins of predicted performance (x-axis) among all new firms in the 2010 test set. The predictive model was trained using 10-fold cross validation on the sample of all firms in the 1998, 2002 and 2006 cohorts.

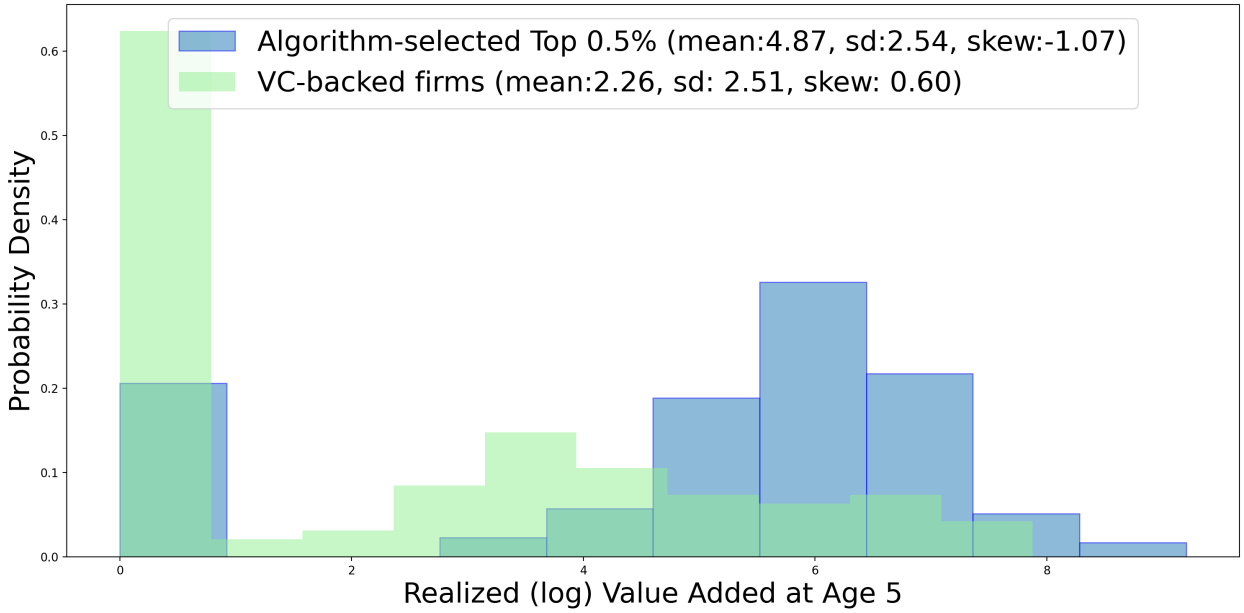

Figure B.5: Realized Performance of Ventures in Test Set When the Algorithm Predicts a Firm's Probability to Be Among the Best Performers. This figure shows the probability density of firm performance for all firms in the 2010 cohort (our test set) as well as the breakdown for VC-backed firms and for algorithm-selected firms using 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. We report the mean, standard deviation and skewness of value added at age 5 (log) for each group.