Venture Capital (Mis)Allocation in the Age of AI

Victor Lyonnet Léa H. Stern*

28th March, 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 covers 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. We estimate the shadow cost of constraints faced by VCs and show that VCs make prediction errors by relying too heavily on certain characteristics such as the entrepreneur gender and education, and not enough on other characteristics such as the firm's industry and location. Algorithmic decision aids show promise to broaden the scope of VCs' investments and founder diversity.

Keywords: venture capital, machine learning, entrepreneurship, algorithmic decision-aid, explainable AI (XAI), capital allocation

JEL Classification: G11, G24, G30, G32, G41, M13, M14

*For helpful comments, we thank Will Cong, Johan Hombert, Camille Hebert, 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, 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 124,204 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 (age, gender, education, etc.) and new firm characteristics (number of employees, customers, suppliers, etc.). 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 on 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. 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-selected and algorithm-selected ventures to uncover sources of errors and we build a 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 sources 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 portfolio firm performance would increase by 85%. Second, VCs pass on some new firms that perform predictably well. If they selected predictably good performers instead of predictably poor performers, VCs' average portfolio firm performance would be an order of magnitude higher. We estimate the shadow cost of 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 (e.g., firms in the same industry, location, or industry-location). For example, the shadow cost of constraining investments to a given commuting zone is 34% of average portfolio performance. Even our most constrained algorithm significantly outperforms VC-backed firms, which suggests that VCs' constraints cannot fully explain the difference in performance between VC-selected and algorithm-selected firms. To better explain this difference, we study the founder characteristics of VC-backed and algorithm-selected firms. Focusing on entrepreneur characteristics that have drawn most attention from the existing literature, we show that VCs tend to select younger entrepreneurs, fewer female founders, founders with lower education levels, and more Paris-based entrepreneurs compared to the algorithmic policy.

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). 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 perform better not only on average, but they are also more likely to become home runs. We also show that 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 accounts for VCs' preference for skewness.

VCs provide various forms of support to the entrepreneurs they select. Therefore, a VC-backed firm performs better on average than an otherwise similar non-VC-backed firm (Puri and Zarutskie, 2012; Bernstein, Giroud and Townsend, 2015). This effect effectively raises the bar for our algorithm to identify outperforming non-VC-backed firms. Gompers et al. (2020) document the importance of active deal generation, showing that only 10% of deals come inbound from company management. We therefore make two assumptions: that VCs could, in principle, have invested in the algorithm-selected firms they did not back, and that these firms would not have performed *worse* had they been backed by VCs. Our analysis suggests that this is the case. First, our analysis is restricted to firms created in industries that receive VC-backing in our data. This restriction alleviates concerns that these firms were 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, we observe whether it received VC from *any* VC. The fact that we do not observe whether a firm matched with one particular VC mitigates concerns related to negotiations and the two-sided matching process of VC investment. Finally, two findings support the idea that VCs can pass on predictable winners due to mis-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 using all new firms).

To understand why VCs do not select the most promising entrepreneurs, we design an algorithmic 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 founder demographics (age, gender, education). However, when it is restricted to these three features, our predictive model of firm performance is much less accurate. We view this result as suggestive evidence that VCs rely on a sparse model to make investment decisions. 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 (2021) and create simple ML models that predict VCs' decisions based on a restricted set of features. We pick the features shown to matter the most for VCs according to the existing literature. This analysis shows that VCs over-weight certain personal characteristics of entrepreneurs and under-weight some firm characteristics when selecting firms. In particular, VCs tend to over-weight gender (Howell and Nanda, 2019; Hebert, 2020) and education (Queiró, 2018), whereas they tend to under-weight new firms' industry and location (Chen et al., 2010). These findings are consistent with the presence of biases and the fact that VCs focus on a narrow set of industries and locations.

We use the SHAP method to better understand how our algorithm maps entrepreneur and firm characteristics into performance predictions. We find a strong overlap in the features that matter the most when predicting firm performance and when predicting VC behavior, 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 VCs and the presence of frictions in VCs' decision making (Hellmann and Puri, 2002; Kaplan, Sensoy and Strömberg, 2009*a*; 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. Consistent with the evidence in Azoulay et al. (2020), we find that VCs overemphasize youth as a trait of successful entrepreneurs. We also find that VCs select fewer female founders compared to our 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 the presence of homophily and network effects (e.g., Hochberg, Ljungqvist and Lu, 2007; Howell and Nanda, 2019; Gompers et al., 2020). They are also consistent with VC performance being driven by their non-local investments (Chen et al., 2010), and could in part explain why entrepreneurs migrate after having founded their firm in a VC hub (Bryan and Guzman, 2021).

A strand of the literature studies which factors affect new firm performance. Kaplan, Sensoy and Strömberg (2009*b*) find that the business is more important than the management team. A few papers focus on changes in new firm quality using performance predictions. Fazio et al. (2016) and Guzman and Stern (2020) study the evolution of new firm quality over time using U.S. business registry data, and Hombert et al. (2016) and Bonelli, Liebersohn and Lyonnet (2021) use French administrative data to measure the effects of unemployment insurance and superstar firms on the quality of new ventures. We extend this literature by identifying which entrepreneur and firm characteristics are predictive of performance and by designing tests to broaden our understanding of VCs' decision making. We study the reasons why VCs' selection of ventures differ from those of the algorithmic policy. 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).

2 Framework

We propose a simple framework of VCs' investment decision that is based on Mullainathan and Obermeyer (2021). In our model, VCs can invest in many new firms that are characterized by a vector of features (X, Z) that are drawn from a fixed distribution. Both X and Z are observed by the VCs, but only X is available to the algorithm. The performance of new firms is a random variable denoted Y. The distribution of new firms' predicted performance based on (X, Z) results in a percentile rank for each new firm denoted as $R(X, Z)$.

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:

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

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

There are two reasons why the VCs' decision might differ from the optimal policy given in (1). First, the VCs could make errors in predicting new firm performance, in which case their prediction model would result in a percentile rank $\tilde{R}(X, Z)$. If $R(X, Z) - \tilde{R}(X, Z)$ is positive, the VCs over-estimate a given firm's probability of success; if it is negative, they under-estimate it. Second, the VCs could have preferences that affect their willingness to invest in a given firm. We denote these preferences as $\Delta(X, Z)$. As shown by Piacentino (2019), uninformed VCs who want to appear informed are biased against backing some firms, which implies that $\Delta(X, Z) > 0$. The VCs' policy writes

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

Given the objective in (2), there are two cases in which the VCs' decision might be inefficient. Either the VCs select firms that are predicted to perform poorly, so that there exists a set $\nu^{\mathbf{L}}$ of *over-backed* firms such that

$$E(I|(X, Z) \in \nu^{\mathbf{L}}) > 0 \text{ and } E(R|(X, Z) \in \nu^{\mathbf{L}}) < t.$$

Or the VCs pass on firms that are predicted to perform well, so that there exists a set $\nu^{\mathbf{H}}$ of *under-backed* firms such that

$$E(I|(X, Z) \in \nu^{\mathbf{H}}) < 1 \text{ and } E(R|(X, Z) \in \nu^{\mathbf{H}}) > t.$$

To establish inefficiencies in VCs' investment decisions, it is enough to observe either a set $\nu^{\mathbf{L}}$ or $\nu^{\mathbf{H}}$ of firms in the data. Our framework also shows how to empirically test for errors in VCs' prediction of new firm performance. If VCs make prediction errors, $\tilde{R}(X, Z) \neq R(X, Z)$ and the objective

[1] 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 .5% of all new firms (Lerner and Nanda, 2020).

in (2) implies that we should observe *both* sets $\nu^{\mathbf{L}}$ and $\nu^{\mathbf{H}}$ of under-backed and over-backed firms in the data. Instead, absent errors in VCs' prediction model, $R(X,Z) = \tilde{R}(X,Z)$ and the VCs' objective becomes

$$I = 1 \text{ iif } R(X,Z) + \Delta(X,Z) > t.$$

In that case, we should only observe a set $\nu^{\mathbf{H}}$ of under-backed firms in the data. Therefore, showing the existence of *both* sets $\nu^{\mathbf{L}}$ and $\nu^{\mathbf{H}}$ in the data is sufficient to show that the VC makes errors in predicting new firm performance.

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. The two main advantages of these data for our study is 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.[2]

The absence of survivorship bias and selection bias is key for our analysis. In particular, our sample contains both VC-backed and non-VC-backed firms. This is in contrast to the existing literature on VC, which is restricted to standard data sources collected from VCs and hence focuses

[2]The French Statistical Office sends questionnaires to approximately 25% of entrepreneurs who started or took over a business in France that year. 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.

on VC-backed firms in isolation.[3,4]

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.[5] 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.[6] We observe firm performance at different ages from total sales and value added reported in the tax files. [7]

New firm creation. We use data on the universe of new firms registered in France in some of our summary statistics. These data come from the firm registry (*SIRENE*) for the period 1998 to 2015.[6] 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 plantinum 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. None of these commercial databases provide the standardized French firm identifiers, hence we build an algorithm to retrieve them from the French Statistical Office's online database accessible through an API.

[3]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).

[4]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.

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

[6]We train our algorithm on the 1998, 2002, 2006, and 2010 entrepreneur cohorts, by comparing its predictions to observed realizations. The firm data 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.

[7]Our sample includes firms that did not survive. We assign the value of their cohort's first percentile to these firms. This value is negative for all cohorts. We do not assign them zeros to avoid making them appear to have performed better than some firms that are still alive (but with negative value added). Our results are not sensitive to the outcome value we choose for firms that did not survive.

4 Algorithmic Policy Design

Choice of objective function. We would like 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 develop an estimator of firm performance $\widehat{m}(\cdot)$ that produces a percentile rank of new firms denoted $M(X)$. This estimator is trained to predict new firms' performance Y five years after their creation. Importantly, the outcome should be a measure of firm success 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). Value added is very similar to EBITDA, with a correlation of .91.[8] 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).[9] Our results are robust to predicting various measures of venture success, and in particular, measures that account for VCs' preference for skewness (i.e., "home run" deals).

Algorithm class and train/test sets. We use Gradient Boosting Trees (*XGBoost*) to generate performance predictions (see Chen and Guestrin, 2016). *XGBoost* is trained on three cohorts of entrepreneurs (1998, 2002 and 2006) representing 69% of our data (86,137 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 (38,067 observations, or 31% of our data). 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 startups

[8] Although private equity firms commonly use EBITDA as the basis for company valuation, we train our algorithm to predict the (log of) value added rather than EBITDA for two reasons. First, value added is measured after taxes, after depreciation and amortization, and after ensuring that 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 of top quality because it serves as the basis for the computation of the value added tax by the French tax authority. Our results are very similar using EBITDA.

[9] 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).

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 465 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.[10] 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 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 founder's age and education are somewhat larger in the test set.[11]

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 (the log of value added at age 5 – $log(va_5)$) to the observed realized performance

[10] To facilitate the interpretation of our results, we exclude firms whose founder reported more than 20 employees in the year of creation. Our results do not depend on this filter.

[11] Bonelli, Liebersohn and Lyonnet (2021) study the time-series evolution of entrepreneurship quality over time, showing that the quality of entrepreneurs has increased over the years. For the purpose of our analysis, long-term changes in entrepreneur characteristics and in the relationship between entrepreneur characteristics and new 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.

of the 38,067 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. We find that the relationship between machine-predicted performance and realized performance is positive, with a coefficient of correlation of .39. Figure 1 illustrates the algorithm's ability to predict the distribution of new firm success reliably. Predictive performance is similar when the algorithm is trained to predict $top5va_5$, a dummy equal to one for firms in the top 5% of their cohort based on value added at age 5.

[Insert Figure 1 here]

Most promising firms in test set. In a typical year, only about .5% of startups receive VC funding. This fraction is similar to that in the US (Puri and Zarutskie, 2012; Lerner and Nanda, 2020). Therefore, we set a threshold t and ask the algorithm to select firms in the top $s = 1 - t\%$ of predicted performance Y. We define the algorithmic policy as:

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

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

[Insert Figure 2 here]

Comparing VC-backed and machine-selected firm performance. We continue our analysis of the algorithm's predictive ability by comparing the performance of VC-selected firms to that of algorithm-selected ones.[12] 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 2.11, whereas it is 1.77 for the entire sample. This performance gap confirms VCs' ability to identify and invest in promising startups.[13] Second, we confirm the *Babe Ruth Effect* in our data: VCs bet on magnitude over frequency, and outcomes follow a power law distribution.[14] On the one hand, VCs are much more likely to invest in startups that die within 5 years than the base rate. On the other hand, conditional on surviving, their portfolio companies do much better than average. Third, the realized average performance of startups identified as most promising by the algorithm is far 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. While data limitations prevent us from performing a complete risk analysis, Table 4 shows a similar standard deviation in the performance of VC-selected and algorithm-selected firms, which is consistent with the idea that the algorithmic investment policy mimics VCs' risk taking. These distributions represent a first indication that the process with which VCs acquire and aggregate signals about a venture's prospects may be inefficient. We reach the same conclusion when the algorithmic policy selects the most promising firms in terms of $top5va_5$. 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) that maximizes the objective of investing in the most promising firms.[15]

5.2 Accounting for unobservables and VC treatment effect

Our results so far suggest that the algorithm is able to pick up signal in the training data that enables the identification of promising new firms in the test set. Two limitations complicate the interpretation of this result. First, although VCs' success critically depends on their ability to

[12]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.
[13]This difference may in part be attributable to VCs' involvement in managing the firm they invest in.
[14]https://www.businessinsider.com/babe-ruth-grand-slams-and-startup-investing-2015-6
[15]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.

generate deal flow (i.e., identifying and connecting with startups, see Gompers et al., 2020), it is possible that unobservables could make VCs not willing, or not able, to invest in the ventures the algorithm identified as particularly promising. Recall, however, that the set of firms the algorithm can select is restricted to firms that operate in industries that have received VC during our sample period. Firms selected by the algorithm are thus suitable candidates for VCs at least with respect to their industry. In addition, concerns related to possibly failed negotiations and the two-sided nature of the matching process are attenuated due to the fact that we observe the union of *all* VCs' decisions to back a given firm, and not the decision from one particular VC firm. Second, we know from the literature that VCs' involvement with their portfolio companies improves average firm outcomes. For example, Bernstein, Giroud and Townsend (2015) find that VCs' involvement increases innovation and the likelihood of a successful exit. Puri and Zarutskie (2012) examine performance differences between VC-backed and non-VC-backed firms and show that although not more profitable at exit, VC-backed firms achieve larger scale and their cumulative probability of failure is lower. This effect effectively raises the bar for our algorithm to identify outperforming non-VC-backed firms.

Two separate analyses help alleviate concerns related to unobservables and treatment effects. First, we evaluate the algorithm's performance in the test set on VC-backed firms only. Figure 4 shows that for the 120 VC-backed ventures in our test set, the algorithm still predicts the distribution of outcomes well. This suggests that the algorithm could prove 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 4 here]

In addition to evaluating the algorithm's performance on the set of VC-backed firms only, 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 received VCs' stamp of approval. Figure 5 shows that an algorithmic policy tasked with selecting the top 15% of firms would improve the allocation of VCs.[16] While this is not our preferred specification (we wish to take advantage of having data on all new firms to learn about VCs' decision making), our findings when training the algorithm on this subsample address

[16]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.

potential concerns about unobservables and VC treatment effect and support the idea that VCs make errors in their investment decisions.

[Insert Figure 5 here]

Chung et al. (2012) show that a fund's current performance not only affects the general partner's compensation through carried interest profit-sharing provisions, it also critically impacts her ability to raise capital for future funds. We find that VCs pass on promising ventures, which is consistent with the prediction that career-concerned VCs would underinvest by restricting their investments to what they expect to be home run deals (Piacentino, 2019).

5.3 Alternative outcome: successful exits

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 deals. Although more the exception than the rule, there are cases when 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). The literature qualifies a VC deal as successful if it has a successful exit (via an acquisition or an IPO, see Fazio et al., 2016; Guzman and Stern, 2020, for instance). To identify 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, and Preqin and create a dummy variable, *successful_deals*, which we set equal to one if the company is either acquired, goes public, or is identified as raising funds through later stage rounds. 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 in our data. We thus ask whether an algorithm could aid VCs hit those "home run" deals.[17]

These successful exits are extremely rare. In our test set (the 2010 cohort of entrepreneurs), we identify 57 companies with such successful exits. There are over 38,000 observations in the test set. 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 the probability of a successful exit for each observation in the test set. We implement an algorithmic policy which selects the firms assigned the highest probabilities by our algorithm, $M(X_i) = P[\widehat{m}(X) \leq \widehat{m}(X_i)] \geq$

[17] 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.

t. In order to compare the performance of the algorithmic investment policy to the performance 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 are "successful deals", therefore VCs' decisions have a precision of 3.3% and a recall of 7.1%.[18] This is the benchmark we use to evaluate the performance of our algorithm.

We test whether the algorithm can pick up sufficient signal in the training data to enable it to flag new companies more likely to experience a successful exit. Given the sparsity of successful exits, the bar is extremely high. We report in Table 2 that the algorithmic investment policy we implement identifies *ex ante* 13 firms that subsequently experience a successful exit.[19] Therefore, our prediction model of home runs has a precision of 6.9% and a recall of 23.2%, both of which are higher than VCs'.

[Insert Table 2 here]

While there are obvious data limitations in this exercise, we view this finding as highly encouraging, especially in light of 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, 2015).

For completeness, we further verify that our results are robust to other measures of firm success. We train predictive models using various outcome measures. Importantly, we show that firms selected by a model that predicts one measure of success also do at least as well as VCs' selections not only based on that specific measure, but also on all other success measures, including measures that account for skewness preferences. Results are reported in Table 3. For example, when the algorithmic policy (s=0.5%) is trained to predict $log(va_5)$ ($log(va_7)$), it selects four (five) firms classified as home runs, and VCs select four.

[Insert Table 3 here]

5.4 Estimating potential performance gains

Contracting the set of VC-backed firms. We would like to assess VCs' deviation from the objective to invest in the most promising firms and evaluate the potential performance gains from

[18] Precision is calculated by dividing true positives by true positives plus 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.

[19] This predictive model of "successful deals" has an area under the curve (AUC) of .84.

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 as the algorithm excludes ventures with low predicted performance. We drop firms one at a time starting with the one with the lowest percentile rank $M(X_i)$. Figure 6 reports the results of this first counterfactual exercise which contracts the set of VC-backed firms. The origin represents the status quo: the full set of VC-backed firms in the test set and their average performance at age 5. 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 6 here]

The results in Figure 6 show that the average performance of portfolio firms would almost double 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. Therefore, these gains do not capture VCs' returns directly; instead, they measure gains in terms of operational performance, which is highly correlated with VCs' returns. Despite our data limitations, our findings reveal that VCs invest in firms that were predictably poor choices (the set $\nu^{\mathbf{L}}$ of "over-backed" firms is not an empty set).

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

[Insert Figure 7 here]

This Centaur model sequentially drops VC-backed firms with low predicted performance and replaces them with firms predicted to become high performers. 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.

Figure 7 shows that as the Centaur model replaces firms in $\nu^{\mathbf{L}}$ with firms in $\nu^{\mathbf{H}}$ (as shown on 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 produces 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 average portfolio firm performance by 60% and 34%, respectively. That even when subject to the most restrictive set of constraints (same industry and location), the Centaur model generates a portfolio performance much higher than that of VCs suggests that these constraints alone do not explain why VCs' and the algorithmic policy's selections differ. We explore other explanations in Section 6.

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 5.4 indicate that constraints in dealflow generation can partly explain why VCs select different firms from the ones selected by our algorithm.

In this section, we explore another (non-mutually exclusive) explanation based on the observation that VCs likely rely on heuristics to help them make investment decisions. In particular, 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, this could help explain our results.

6.1 VC-selected vs. algorithm-Selected firms

One way to shed light on the process by which VCs gather signals of venture's potential is to examine how various demographic measures of entrepreneurs differ across VC-selected entrepreneurs and algorithm-selected ones. Figure 8 reports the probability densities of founders' ages, gender, education level, and geographic location of VC-backed and algorithm-selected ventures.

[Insert Figure 8 here]

Age. Panel A shows that the distribution of founders' age for the algorithm-selected firms tends to be right-skewed relative to VC-backed firms. This result is in line with findings in Azoulay et al. (2020) on venture capitalists overemphasizing 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 startups, 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 female founders. A plausible interpretation of this finding is that heuristics used by VCs to predict venture prospects lead them to overlook promising female entrepreneurs.

Education. In Panel C, we show that VCs are more likely to invest in ventures of entrepreneurs with lower education levels when compared to the algorithm.

Geography. Finally, Panel D explores the role of proximity in VCs' investment decisions. Howell and Nanda (2019) document the importance of networking effects in the industry. In our test set, only 9% of new firms are located in the Paris region. Yet, approximately one in five VC-backed firms is located in Paris, which is a key investors cluster. In contrast, the algorithm would fund new ventures from Paris at a rate below the base rate.

Taken together, the results in Figure 8 illustrate the discrepancies in demographic features for VC-backed and algorithm-selected founders. These results are consistent with the idea that VCs may be passing up promising ventures whose founders are female and/or outside their immediate network. To gain a better understanding of how VC-backed firms differ from machine-selected firms, beyond founders' demographics, we report summary statistics for a subset of features as well as t-tests for difference in means for VC-selected and machine-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%.

Founder experience, motivation, and other input features. While 16% of entrepreneurs in the test set reported that the motivation for starting their company was a "new idea", 39% (7%) of VC (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-selected founders have same prior activity experience, while 84% 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.

[Insert Table 4 here]

6.2 The Shadow Cost of Gender Preferences

Our findings suggest that VCs pass up 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). We therefore use our Centaur model approach to provide an estimate of the shadow cost of gender preferences. The setup is the same as in Section 5.4 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 9 reports the results graphically.

[Insert Figure 9 here]

We use the difference in realized average portfolio firm performance as a measure of the shadow cost of the gender constraint. When the Centaur model gives full weight to the algorithm, the realized average performance of the firms it selects drops by 8% (7%) when the gender constraint is added to the industry (industry and location) constraint. While this approach is subject to the same interpretational limitations and caveats as those described in Section 5.4, this estimate of the shadow cost of gender preferences contributes to broadening our understanding of the extent to which frictions affect VC allocation efficiency.

6.3 Model Interpretability

We would like to make our model more transparent for two main reasons. First, we wish to gain an understanding of which features matter in generating the predictions in order to improve interpretability. Second, we are interested in contrasting the prediction-relevant features of the model to VCs' decision-making practices.

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 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. An input feature's SHAP values can be aggregated across observations to facilitate the model's global interpretability by yielding a ranking of features that contribute the most to the predictions. We stress that SHAP values do not allow any causal interpretation. However, they offer a helpful basis for expanding our qualitative understanding of how our algorithmic models generate predictions. Erel et al. (2021) use the SHAP method to understand better their model's predictions of corporate directors' performance.

Figure 10 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 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 10 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 was working in the same industry as the new venture, and the founder's age. The analysis of SHAP values shows that a large number of founders were working in the same activity prior to creating the venture (stacked red observations *same activity as before*), 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. Most founders are middle aged (represented in 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 their age has a positive impact on the prediction.

In Figure 11, we report the SHAP summary plot when the algorithm is trained only using VC-backed firms, as in Section 5.2. 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 founders are predicted to perform *worse* than other VC-backed founders. In addition, among the set of VC-backed founders, those who answered that their motivation to start the business was a new idea on average are predicted to perform worse than those who did not list "new idea" as a motivation. Those with entrepreneur relatives are predicted to perform better.

[Insert Figure 11 here]

Finally, in Figure 12, we show the SHAP summary plot for our predictive model of home runs, i.e., deals associated with a successful exit such as a new round of VC financing, an acquisition, or an IPO. We document in Section 5.3 that our model shows promise to identify these low probability events. There is a strong overlap in the most relevant features to predict home runs and those that predict operating performance, lending further support to the idea that there is a strong overlap between the model that predicts home runs and the one that predicts operating performance. Features related to innovation are key in moving the predictions. The model suggests that B2B businesses are more likely to hit home runs. When they are VC-backed, B2B firms are predicted to perform better; yet, we show in the next section that B2B businesses are less likely to be VC-backed.

[Insert Figure 12 here]

Venture capitalists are well-incentivized, sophisticated investors. It is therefore surprising that across various performance measures, an algorithm could outperform their selection. In Section 5.4, we explored a possible explanation for this puzzle: deal sourcing involves a costly search process.

A complementary interpretation is that heuristics-based investment decisions lead VCs to select some predictably poor performers and pass on other predictably good performers. Kanze et al. (2018) show that VCs, regardless of their gender, ask different kinds of questions to male and female entrepreneurs when they pitch their business idea. They tend to ask questions about potential for gains (losses) to men (women), leading to significant differences in funding outcomes. Calder-Wang and Gompers (2021) find that homophily-driven biases in VCs' funding decisions make them pass up funding female founders. The findings in this section support the view that the reason why VCs' selections differ from the algorithmic policy is not exclusively due to the costly search process.

6.4 Predicting VCs' decisions

To better understand why VCs' decisions differ from the algorithmic investment policy, we develop a separate estimator, $\widehat{h}(\cdot)$, that predicts VCs' decisions. This classification algorithm predicts whether a firm is VC-backed. It is trained on a random split of 70% of the observations in the 1998, 2002, and 2010 cohorts, and tested out-of-sample on the remaining 30% of observations.[20]

Model performance Our predictive model predicts VCs' investment decisions reasonably well. Figure 13 shows that our model has an area under the curve (AUC) of .79. 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 79%.

[Insert Figure 13 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 we restrict the model predicting venture performance to these three features, the algorithmic policy's performance decreases dramatically. The model's much lower

[20] 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: only a small number of firms 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.

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 to these three demographic features when making investment decisions.

Model interpretability In Figure 14, we report the SHAP summary plot for our predictive model of VCs' investment decisions, i.e., which types of entrepreneurs are more likely to be VC-backed. 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 services, as well as those who state that the motivation for starting their business was a new idea. 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 11 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.

[Insert Figure 14 here]

Signal beyond venture performance To further our understanding of VCs' firm selection, we follow the approach in Jens and Mullainathan (2021) and test in a regression framework whether there is signal to predict VCs' decisions, beyond predicted performance. We first regress VCs' actual decisions on our algorithmic predictions of VCs' decisions:

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

Table 5 shows that our model of VCs' decisions is correlated with VCs' actual decisions (column 1).[21] We then regress VCs' actual decisions on venture performance predictions, first using a model that predicts $logva_5$, and then one that predicts home runs:

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

If VCs did not care about portfolio firms' operating performance, we would expect performance

[21] 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 (VC-backed status is not available for the 2006 cohort).

predictions to not load significantly. This is not the case: Column 2 shows that predicted operating performance correlates with VC-backed status. Column 3 shows that home run predictions also correlate with VCs' decisions. Next, we test whether our predictions' of VCs' decisions remain significant once we control for performance predictions by estimating:

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

Columns 4 through 6 show that there remains significant predictability in VCs' behavior beyond our algorithmic predictions of venture performance: the coefficient on our predictions of VCs' decisions is 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, suggesting that VCs either care about other objectives than performance or home runs, or that they fail to predict performance accurately. In column 6, we include both predictive models of venture success (operating performance and home runs) in addition to the algorithmic predictions of VC behavior. Home run predictions lose significance, but operating performance predictions do not. Importantly we draw the same conclusion that there remains strong predictability in VC behavior beyond what we would expect if VCs were only interested in future venture success.

[Insert Table 5 here]

Biases in VCs' decisions How well do VCs use the features that enter their firm selection model? As noted by Lerner and Nanda (2020), while we know that "early-stage investors rely heavily on signals of entrepreneur quality, we know very little about whether the emphasis on these signals is efficient." To tackle this question, we start with Figure 15, which shows the correlation of each entrepreneur characteristic with both firm operational performance (x-axis) and VCs' investment decision (y-axis).[22] We find a strongly positive relationship (R^2: 0.1132), which suggests that VCs correctly weight most entrepreneur characteristics. Nevertheless, we observe some outliers in Figure 15. This finding, coupled with Figure 13 showing that a large part of the signal that predicts VCs' investment decisions lies in three demographic features, suggests that VCs put a lot of weight on some specific entrepreneur characteristics.

[Insert Figure 15 here]

[22]We exclude industry and location features to account for the fact that VCs might be constrained to select firms within a specific industry or geographic area.

To further assess whether the strong emphasis on certain signals is efficient, we follow the approach in Mullainathan and Obermeyer (2021) and create new simple models $\widehat{m}_{simple}(\cdot)$ to predict operational performance. In these simple models, the only departure from the estimator $\widehat{m}(\cdot)$ is that we restrict the set of input features to variables VCs have been shown to be sensitive to according to the literature. We then regress VCs' decisions on our full model predicting operational performance as well as our simple models:

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

We interpret the sign of the coefficient β_2 as in Mullainathan and Obermeyer (2021). If the coefficient β_2 on a simple model's prediction is positive, this means that the features used in this model matter for VCs' decisions over and beyond their effect on firms' performance. In this case, VCs *over-weight* the features used in the simple model. Instead, if the coefficient β_2 is negative, the features used in the simple model matter less for VCs' decisions than they do to predict firm performance. In that case, VCs *under-weight* the features used in the simple model. We report results when $\widehat{m}(\cdot)$ and $\widehat{m}_{simple}(\cdot)$ predict operating performance, but the results are similar when the models predict home runs.

Since potential investors are highly responsive to information about the founding team (Bernstein, Korteweg and Laws, 2017; Gompers et al., 2020), our first simple model uses the personal characteristics of the entrepreneur as input features: age, gender, education, nationality, and whether there are entrepreneurs among her relatives. Column 1 of Table 6 shows the results of regressing our VC-backed variable on $\widehat{m}(X_i)$, the full prediction model. Column 2 adds our first simple model based on personal characteristics. $\widehat{\beta}_2$ is positive and significant, which means that $\widehat{m}_{simple}(\cdot)$ is *additionally* predictive of VCs' decisions, so that VCs over-weight personal characteristics of entrepreneurs in their decisions.

[Insert Table 6 here]

In Columns 3 to 7 we test other simple models focusing on each personal characteristic in isolation. We use entrepreneur age in Column 3 and do not find evidence that VCs systematically under- or over-weight age in their decisions to back a firm. Similarly, in Columns 6 and 7 we do not find that VCs are biased towards the entrepreneur's nationality or her family's entrepreneurial background. However, Columns 4 and 5 show that VCs tend to over-weight gender (e.g., Howell and Nanda, 2019; Hebert, 2020) and education (e.g., Queiró, 2018), respectively, in their decision to back a firm.

Column 8 shows that VCs tend to under-weight new firms' industry and location in their decision to back a new firm. This finding is consistent with evidence that VCs typically invest in specific industries and locally (Chen et al., 2010). Finally, Column 9 focuses on proxies for the venture's traction (Bernstein, Korteweg and Laws, 2017): the total number of workers, the number of clients, and the clients' location. We do not find evidence that VCs over-weight new firms' traction in their decisions.

Finally, we compare the characteristics of ventures that have low predicted performance but high chances of being VC-backed to those that have high predicted performance but low chances of being VC-backed. First, we sort firms into quintiles according to their predicted performance and their predicted chance of being VC-backed. Second, we keep firms that are in the first and fifth quintiles of these distributions and end up with two groups of firms: one group containing those firms that have both, the lowest predicted performance *and* the highest chances of being VC-backed, and one group containing those firms that have both, the highest predicted performance *and* the lowest chances of being VC-backed. Third, we run a t-test of the difference in entrepreneur characteristics between these two groups. Table 7 contains the results.

[Insert Table 7 here]

Consistent with our earlier results, the results in Table 7 suggest that VCs' investment decisions differ from the algorithmic policy because they put too much weight on certain entrepreneur characteristics that do not predict success well. In particular, we find that founders who are more likely to be VC-backed even though our algorithm predicts them to perform poorly tend to be Paris-based, older, male, serial entrepreneurs, with lower levels of education. Those results do not imply that these characteristics are not predictive of firm performance; rather, they provide information on which entrepreneurs are more likely to be casting errors.

7 Conclusion

The goal of this paper is to broaden our understanding of 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. 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. Compared to the algorithmic policy, VCs tend to select younger entrepreneurs, fewer female founders, founders with lower education levels, and more Paris-based entrepreneurs. We design an algorithmic model that predicts for each new firm whether it is VC-backed. Consistent with the observed differences between VC-selected and algorithm-selected firms, we find that VCs over-weight some characteristics, such as the entrepreneur gender and education, and under-weight others, such as the firm's industry and location.

Our findings imply that an algorithmic policy could assist VCs in identifying untapped markets, addressing some important questions raised in Lerner and Nanda (2020) about the limitations of the venture capital model. One critique of the venture model is that it is *narrow* in terms of the industries, technologies, and entrepreneurs that are VC-backed, and in the diversity of its decision makers. Our results show a glimpse of what a counterfactual economy would look like in a world in which algorithms would augment VCs' decisions. Algorithms would increase the scope of VCs' investments and the diversity of entrepreneurs who receive VC-backing. Funds would flow to geographical regions that are less connected to financial hubs.

As technology enablers, VCs are uniquely positioned to take up algorithmic decision aids. Their financial incentives should help overcome any residual algorithm aversion. We caution that any implementation of an algorithmic decision aid warrants careful analysis and audit of the algorithm (Ludwig and Mullainathan, 2021; Acemoglu, 2021). Building fair algorithms requires understanding the importance of data incompleteness, how human decisions may respond to algorithmic predictions, and other general equilibrium effects. Our algorithmic policy was built with predictive accuracy as its unique objective. Thus, it minimizes prediction error but not unfairness (Kearns and Roth, 2019). Despite this unique objective, we find that the algorithmic policy would move the allocation of venture capital closer to the (error, unfairness) Pareto frontier compared to VCs' allocation, for instance in the gender fairness dimension.

Figures and Tables

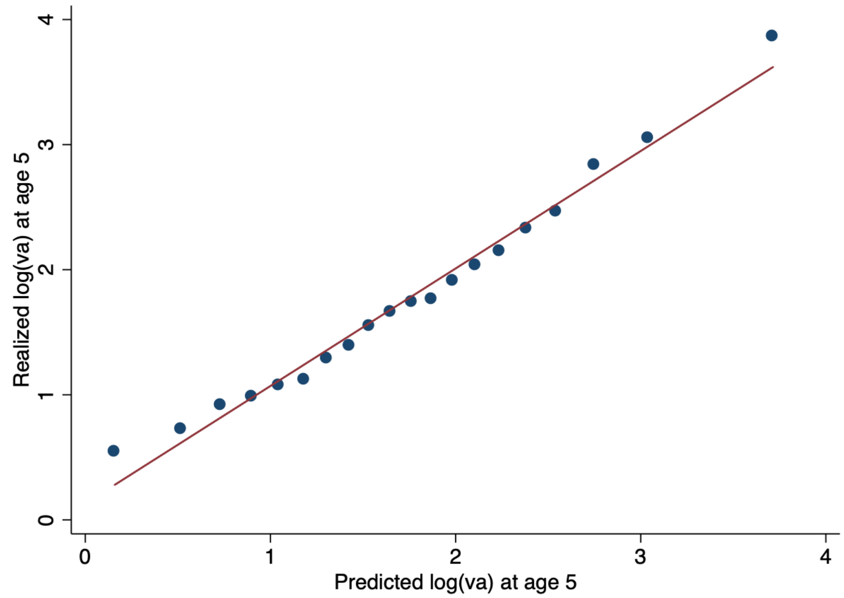

Figure 1: Algorithm performance: all new firms in test set. This figure shows the average observed performance (y-axis) across the 10 deciles of predicted performance (x-axis) across 20 bins of predicted performance (x-axis) among all new firms in the 2010 test set. The performance measure is level of 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.

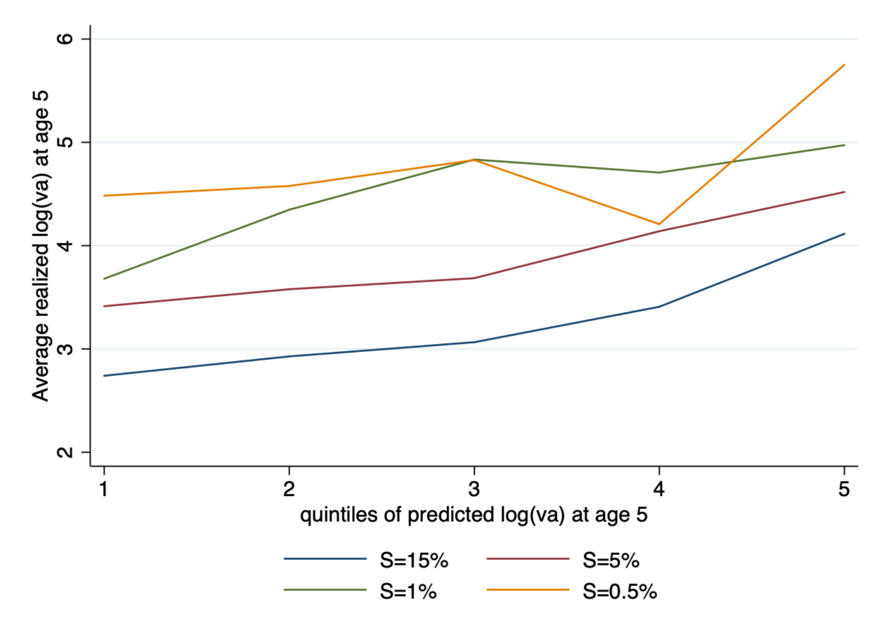

Figure 2: Algorithm performance: algorithm-selected new firms in test set for various selectivity thresholds. This figure shows the average observed performance (y-axis) across the 10 deciles of predicted performance (x-axis) for various selectivity thresholds among all new firms in the 2010 test set. The performance measure is level of 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.

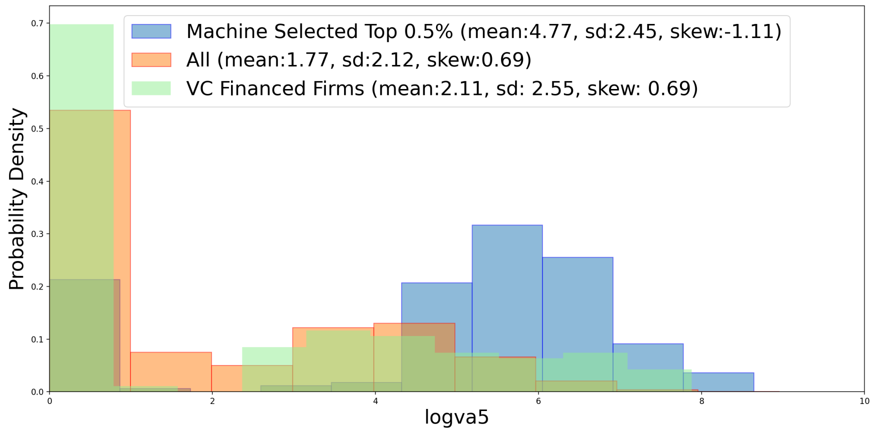

Figure 3: Realized performance of ventures in test set when the algorithm predicts value added at age 5 (log) This figure shows the probability density of firm performance for all firms in the 2010 cohort (our test set) as well as the breakdown for VC-backed firms and for algorithm-selected firms using 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.

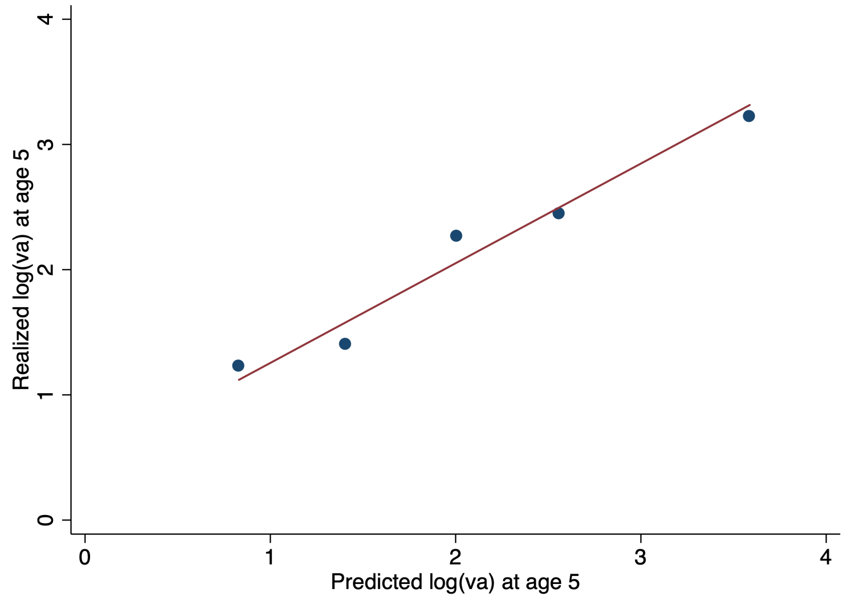

Figure 4: Algorithm performance: VC-backed firms in test set. This figure shows the average observed performance (y-axis) across 5 bins of predicted performance (x-axis) for the VC-backed firms in our 2010 cohort (our test set). The predictive model is trained on the sample of all new firms in the 1998, 2002 and 2006 cohorts using 10-fold cross validation.

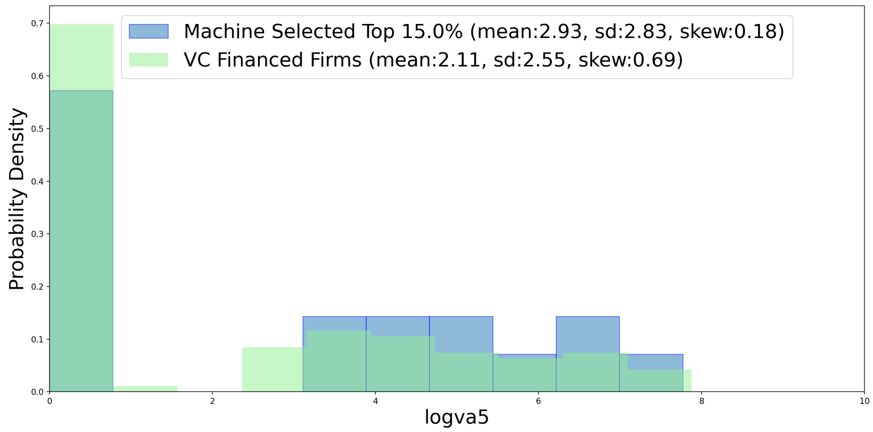

Figure 5: Realized performance of ventures in test set when the algorithm is trained on VC-backed firms only. This figure shows the probability density of log value added at age 5 for the VC-backed firms in the 2010 cohort. We train a predictive model using VC-backed firms only and set the algorithmic financing policy threshold at S=15%. The predictive model is trained on the sample of VC-backed firms in the 1998 and 2002 cohorts using 10-fold cross validation. We report summary statistics for VC-selected firms and algorithm-selected firms.

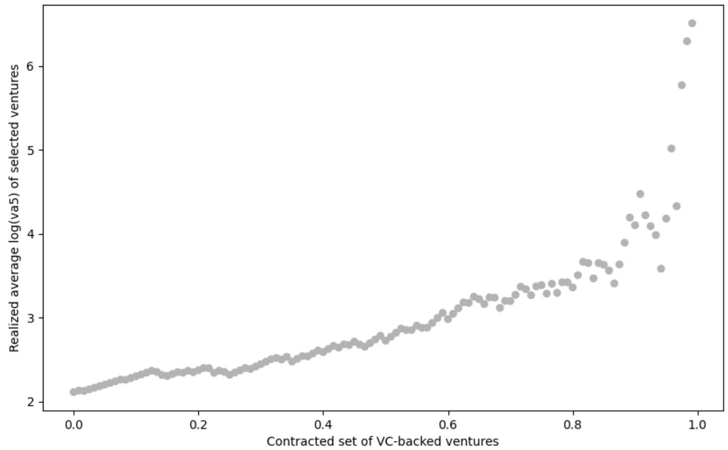

Figure 6: Potential performance gains: contracting the set of VC-backed firms. This figure shows the average portfolio firm' value added at age 5 (in log) if VCs were to drop one by one the firms with the lowest algorithm-predicted performance. The origin represents the status quo: it includes the full set of VC-backed firms in the test set and their observed average performance at age 5. The number of 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.)

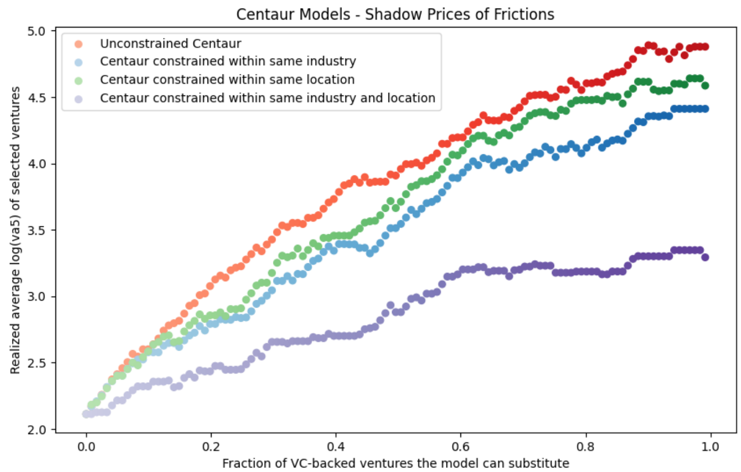

Figure 7: Potential performance gains: Centaur models. This figure shows the average realized performance at age 5 for several Centaur models that replace VC-backed firms that are predicted to become poor performers with firms that are predicted to become high performers by the algorithm. The origin 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 orange line shows the performance of the unconstrained Centaur model. Each line 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 location *and* industry (in purple).

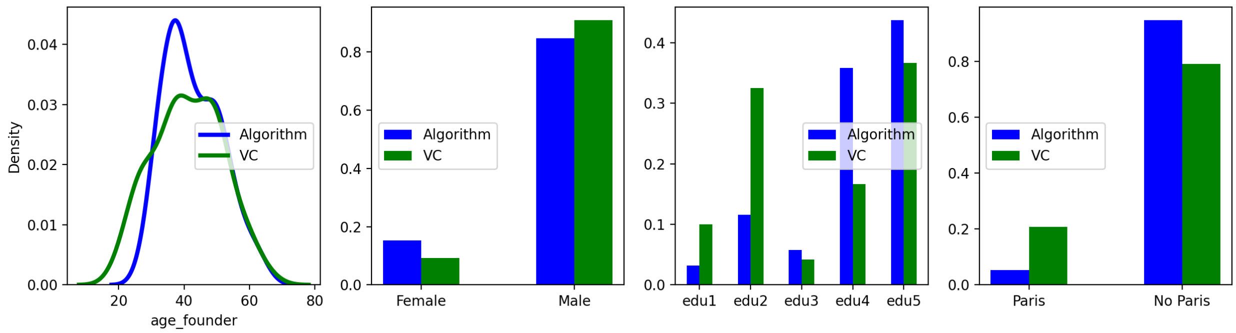

Figure 8: Entrepreneur demographics for VC-backed and algorithm-selected ventures. This figure shows the probability densities of founders' ages, gender, education level and geographic location in the 2010 cohort (our test set) for VC-backed and algorithm-selected firms at the S=0.5% threshold. The predictive model is trained on the sample of all new firms in the 1998, 2002 and 2006 cohorts using 10-fold cross validation.

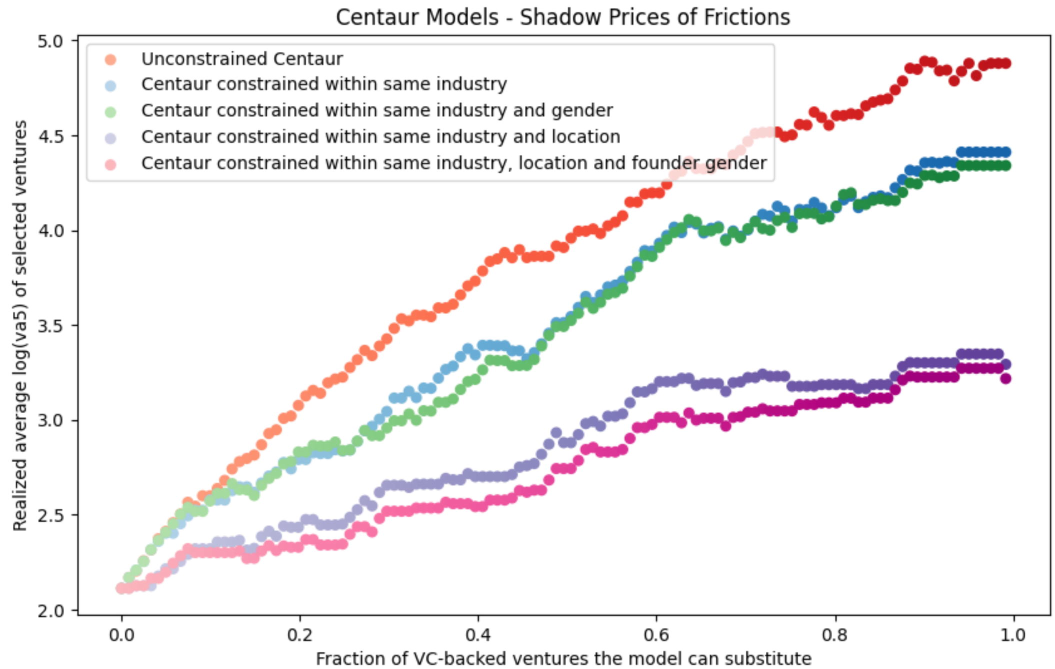

Figure 9: Centaur models with gender constraints. This figure shows the average realized performance at age 5 for several Centaur models that replace VC-backed firms that are predicted to become poor performers with firms that are predicted to become high performers by the algorithm. The origin represents 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 orange line shows the performance of the unconstrained Centaur model. Each line 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).

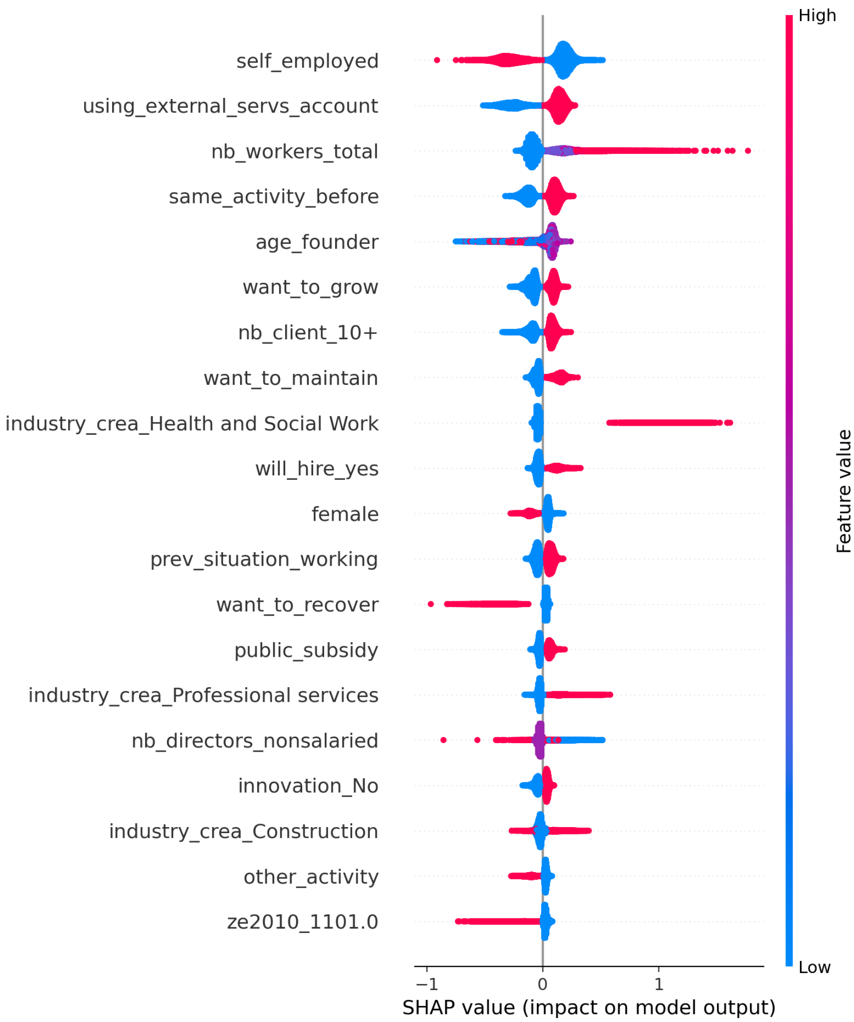

Figure 10: SHAP values of most important input features. This figure reports the SHAP values for the top-20 features that are most important in predicting operating performance. Features are ranked in decreasing order of importance. For each feature, each point represents one observation and its location 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.

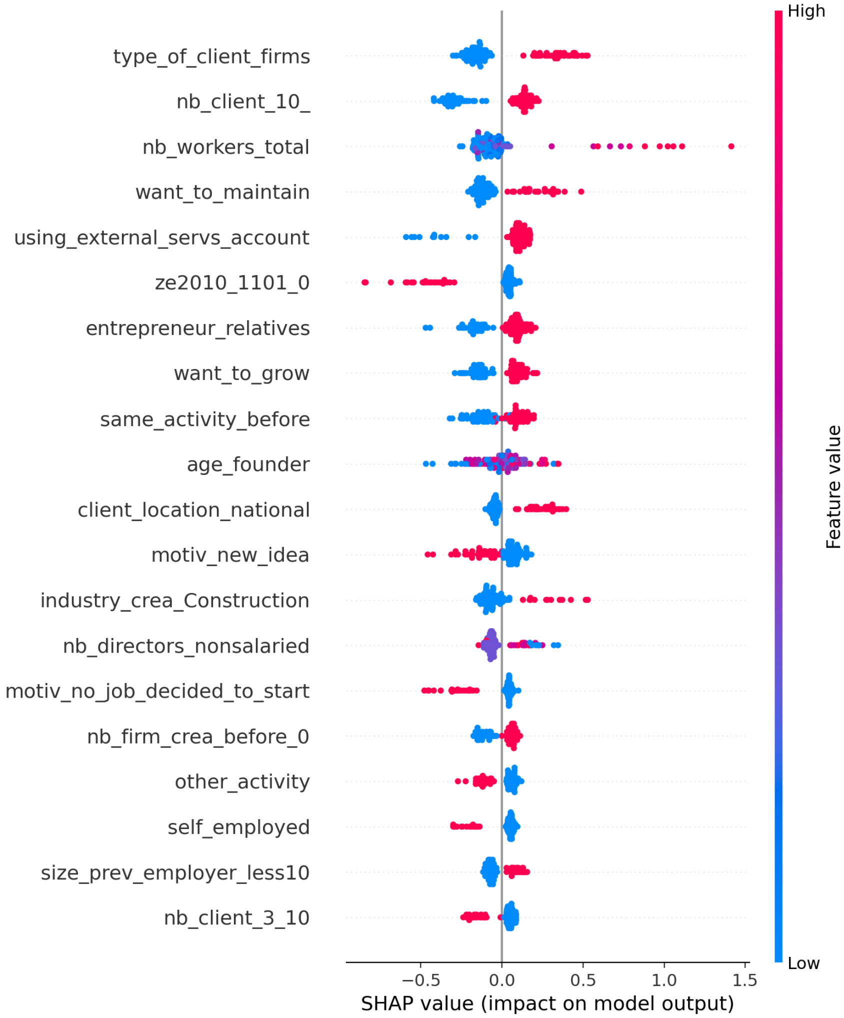

Figure 11: SHAP values of most important input features when the model is trained on VC-backed firms only. This figure reports the SHAP values for the top-20 features that are most important in predicting operating performance. Features are ranked in decreasing order of importance. For each feature, 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.

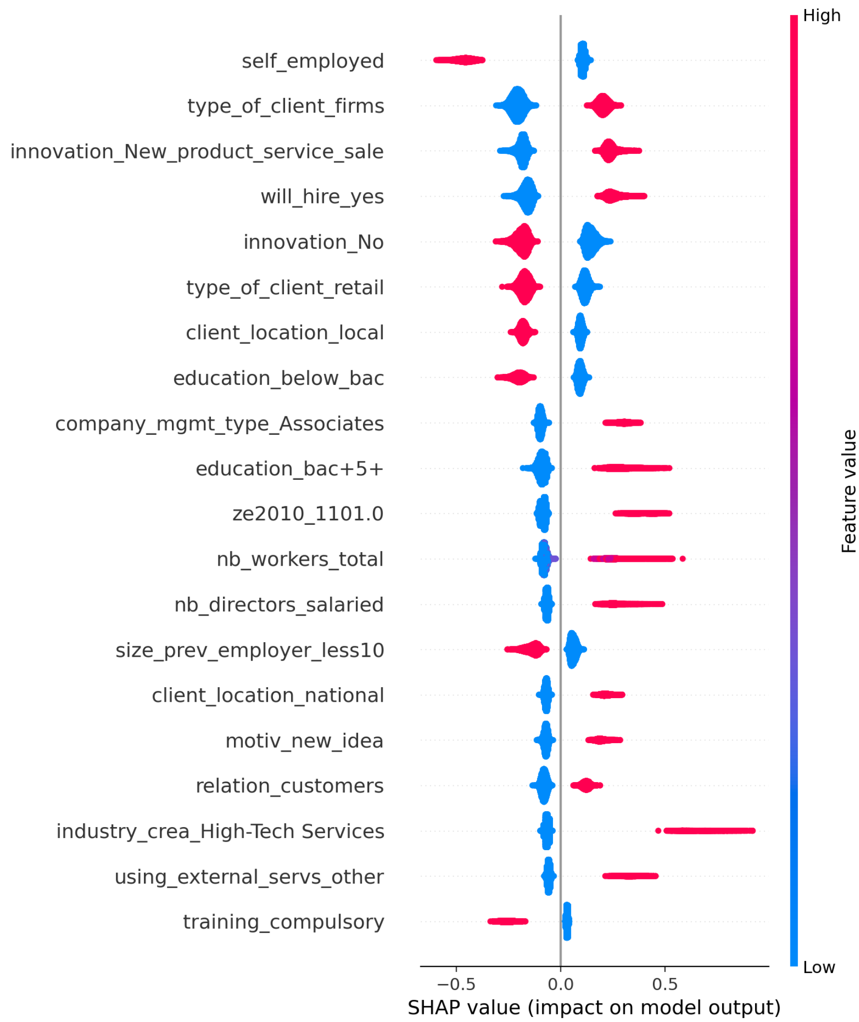

Figure 12: SHAP values of most important input features when predicting home runs. This figure reports the SHAP values for the top-20 features that are most important in predicting successful deals. The dummy variable *successful deal* is a proxy for a successful exit by the VC, i.e., a "home run." Whether 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.

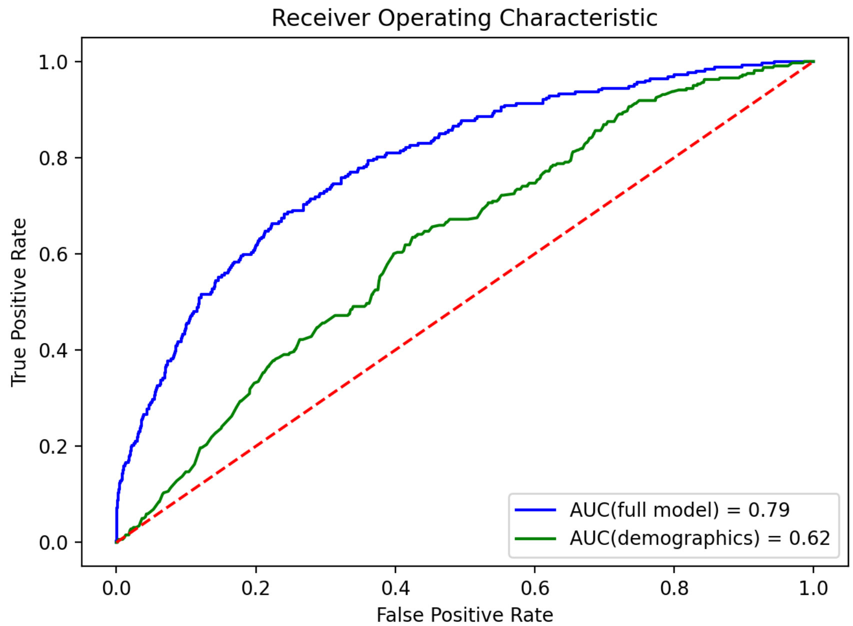

Figure 13: Area under the curve (AUC) of a predictive model of VCs' decisions. This figure presents the ROC and AUC of a predictive model of VCs' decisions. The AUC of .79 for the full model implies that for two randomly picked ventures, one VC-backed and one not, the odds that our model assigns a higher probability of being VC-backed to the one that is indeed VC-backed is 79%. We also report the ROC and AUC of a model that only includes founder's demographic features (age, gender and education).

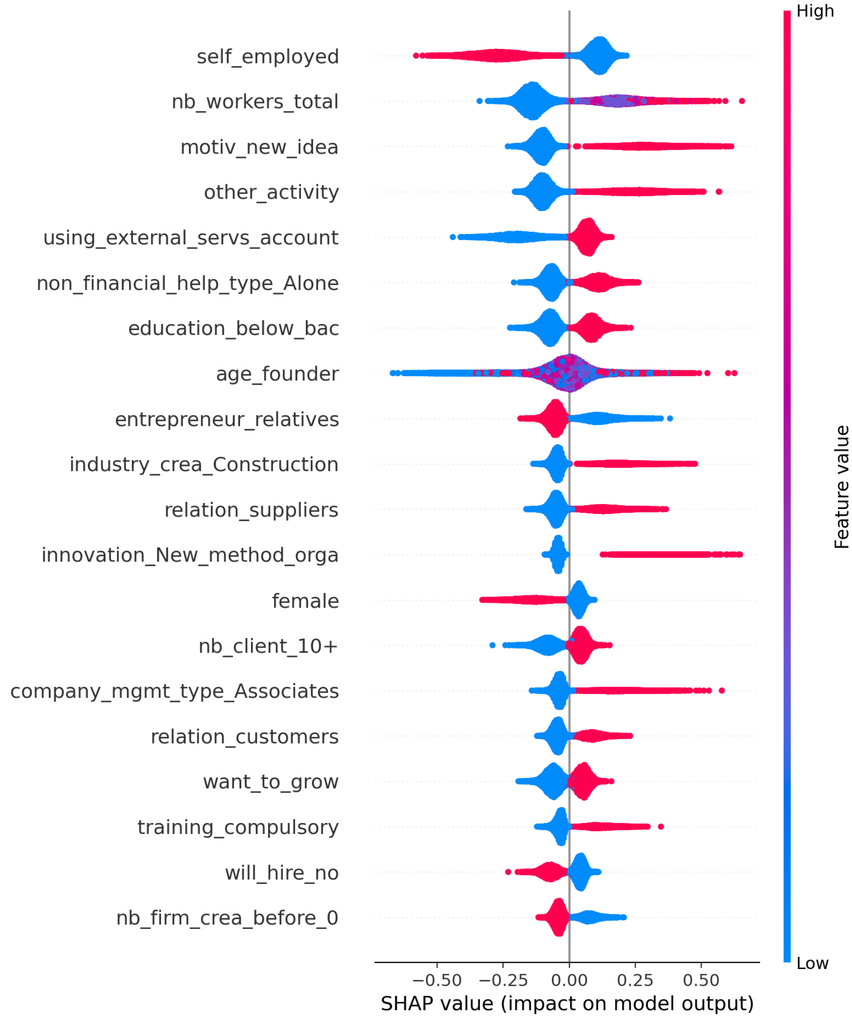

Figure 14: SHAP values of most important input features when predicting VC financing. This figure reports the SHAP values for the top-20 features that are most important in predicting whether a firm will receive VC financing. The predictive model is trained on a random sample of all new firms in the 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.

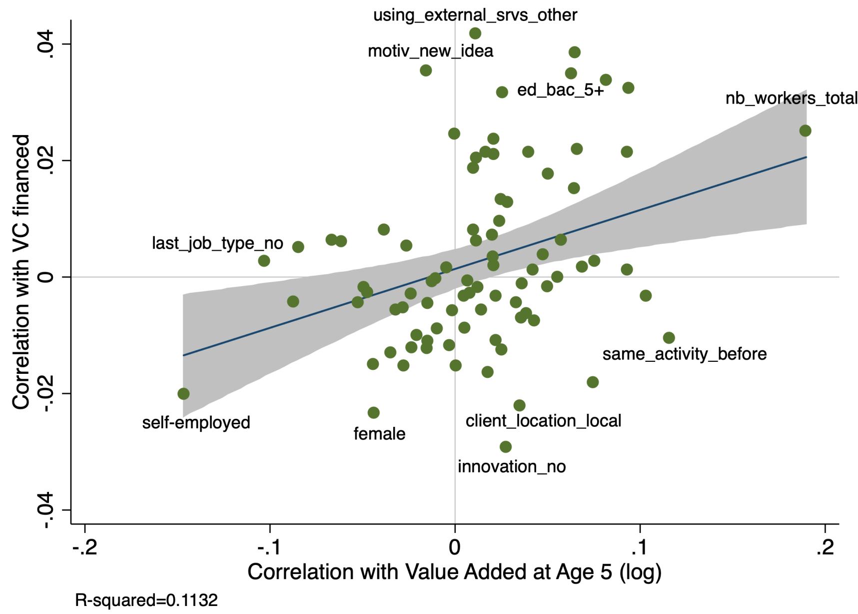

Figure 15: Correlation with VC-backed and performance. This figure shows the correlation coefficients of all input features, excluding location and industry ones, with VC-backed (y-axis) and operating performance (x-axis). We also plot the best linear fit from a univariate regression of VC-backed on log value added at age 5, along with a 95% confidence interval. Some outlier points of interest are labeled.

	Training Set			Test Set		
variable	Mean	SD	N	Mean	SD	N
Value Added in 5 years (log)	1.64	2.09	85658	1.77	2.12	37853
Alive at Age 5	0.62	0.48	85658	0.66	0.48	37853
Founder's Age	37.78	10.01	85658	39.74	10.66	37853
Founder's Nationality (FR)	0.86	0.34	85658	0.92	0.27	37853
Female	0.28	0.45	85658	0.28	0.45	37853
Education (1-5 scale)	2.70	1.25	85658	2.93	1.30	37853
Paris-based	0.10	0.30	85658	0.08	0.28	37853
Same Prior Activity	0.54	0.50	85658	0.61	0.49	37853
Motivation: New Idea	0.18	0.38	85658	0.16	0.37	37853
Motivation: Opportunity	0.32	0.47	85658	0.44	0.50	37853
Motivation: Peer	0.11	0.31	85658	0.09	0.28	37853
External Accounting	0.64	0.48	85658	0.74	0.44	37853
External Management	0.06	0.23	85658	0.06	0.24	37853
External Logistics	0.06	0.24	85658	0.09	0.28	37853
Growth-minded	0.53	0.50	85658	0.42	0.49	37853
Total Workers	1.59	1.52	85658	1.61	1.56	37853
Innovation	0.39	0.49	85658	0.47	0.50	37853
Workers in Previous Employer (<10)	0.40	0.49	85658	0.43	0.49	37853
Created Firm Before	0.27	0.45	85658	0.29	0.46	37853
Will Hire	0.24	0.43	85658	0.26	0.44	37853
Number of Industries	-	-	38	-	-	41
Number of Regions	-	-	321	-	-	321

Table 1: Summary Statistics. This table reports summary statistics for the outcome measure (Value Added in 5 years) and a subset of features in our training and test sets. 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).

	All Successful Deals			Acquisition Events Only		
cohort	All	VC-backed	Machine Selected S=0.5% (189 firms)	All	VC-backed	Machine Selected S=0.5% (189 firms)
2010	57	4	13	37	3	7

Table 2: Summary Statistics: Successful Deals. This table presents the total number of successful deals and acquisition events for firms in each cohort of the *SINE* survey. 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 S=0.5% threshold. The data come from Crunchbase, Capital IQ, CB Insights, Preqin, Venture Xpert, SDC and Zephyr.

		Outcome algorithm evaluated on					
		Top 5% VA_5	Top 5% VA_7	$VA_5(log)$	$VA_7(log)$	$Ebitda_5/capital_0(log)$	Successful Deals
Outcome Y	Mean for all firms in test set	0.05	0.05	1.88	1.61	1.01	57
algorithm trained on	Mean for VC-backed firms	0.14	0.13	2.26	1.94	1.06	4
Top 5% VA_5		0.60	0.56	4.97	4.60	3.09	4
Top 5% VA_7		0.56	0.55	4.90	4.64	3.09	4
$VA_5(log)$		0.57	0.57	5.14	4.90	3.16	4
$VA_7(log)$		0.52	0.51	5.01	4.83	3.01	4
$\text{Ebitda}_5/\text{capital}_0(log)$		0.45	0.45	4.73	4.50	3.01	3
Successful Deals		0.27	0.25	3.07	2.62	1.75	13

Table 3: Performance of Algorithmic Financing Policy Measured with Various Measures of Firm Success. 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. The first row shows the mean of each performance measure for the 2010 cohort (test set). The second row does so for the subset of VC-backed firms only also for the 2010 cohort (test set). In the third row, we train the algorithm to predict whether the firm will be in the top 5% of its cohort in terms of value added at age 5. We report average outcomes across the other measures of firm success for algorithm-selected firms at the S=0.5% threshold. The remaining rows repeat this exercise for models trained on different outcome measures Y.

	Test Set										
variable	Human Selected			Machine Selected (S=0.5%)			Machine Selected (S=1%)			Diff (S=0.5%)	Diff (S=1%)
	Mean	SD	N	Mean	SD	N	Mean	SD	N	T-Test	T-Test
Value Added in 5 years (log)	2.11	2.56	120	4.77	2.46	190	4.51	2.43	379	-2.66***	-2.39***
Alive at Age 5	0.69	0.46	120	0.87	0.33	190	0.87	0.33	379	-0.18***	-0.18***
Founder's Age	41.26	10.58	120	42.81	8.77	190	42.01	8.93	379	-1.55	-0.75
Founder's Nationality (FR)	0.94	0.24	120	0.99	0.07	190	0.98	0.12	379	-0.05**	-0.04*
Female	0.09	0.29	120	0.15	0.36	190	0.16	0.36	379	-0.06	-0.06**
Education (1-5 scale)	3.38	1.50	120	4.05	1.12	190	3.78	1.25	379	-0.68***	-0.40***
Paris-based	0.21	0.41	120	0.05	0.22	190	0.05	0.21	379	0.16***	0.16***
Same Prior Activity	0.52	0.50	120	0.84	0.37	190	0.87	0.34	379	-0.33***	-0.35***
Motivation: New Idea	0.39	0.49	120	0.07	0.26	190	0.09	0.29	379	0.32***	0.30***
Motivation: Opportunity	0.38	0.49	120	0.54	0.50	190	0.54	0.50	379	-0.17***	-0.16***
Motivation: Peer	0.06	0.24	120	0.11	0.31	190	0.10	0.30	379	-0.05	-0.04
External Accounting	0.91	0.29	120	0.83	0.38	190	0.87	0.34	379	0.08**	0.04
External Management	0.09	0.29	120	0.25	0.44	190	0.24	0.43	379	-0.16***	-0.15***
External Logistics	0.15	0.36	120	0.28	0.45	190	0.24	0.43	379	-0.13***	-0.09**
Growth-minded	0.58	0.50	120	0.55	0.50	190	0.53	0.50	379	0.02	0.05
Total Workers	2.30	2.82	120	6.79	5.20	190	5.61	4.52	379	-4.49***	-3.31***
Innovation	0.73	0.44	120	0.46	0.50	190	0.45	0.50	379	0.28***	0.29***
Workers in Previous Employer (<10)	0.29	0.46	120	0.31	0.46	190	0.37	0.48	379	-0.02	-0.08*
Created Firm Before	0.42	0.50	120	0.39	0.49	190	0.37	0.48	379	0.03	0.04
Will Hire	0.51	0.50	120	0.56	0.50	190	0.55	0.50	379	-0.05	-0.05
Number of Industries	-	-	29	-	-	27	-	-	31	-	-
Number of Regions	-	-	68	-	-	89	-	-	147	-	-

Table 4: Summary Statistics: Human vs Machine Selected Start-ups. 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 human and machine selected firms. 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)
$\widehat{h}(X_i)$	0.202***			0.200***	0.200***	0.199***
	(0.00830)			(0.00831)	(0.00839)	(0.00840)
$\widehat{m}(X_i)_{log(va_5)}$		0.00360***		0.00279***		0.00272***
		(0.000634)		(0.000628)		(0.000630)
$\widehat{m}(X_i)_{homerun}$			0.0430***		0.0140*	0.0113
			(0.00836)		(0.00836)	(0.00838)
Observations	26,776	26,776	26,776	26,776	26,776	26,776
R-squared	0.022	0.001	0.001	0.022	0.022	0.022

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. $\widehat{h}(X_i)$ predicts whether a firm is VC-backed; $\widehat{m}(X_i)_{log(va_5)}$ predicts firm operating performance; $\widehat{m}(X_i)_{homerun}$ predicts home runs. 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 first and third predictive models at the S=0.5% threshold.

	VC-backed								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\widehat{m}(X)$ (full model)	0.00109***	0.000749**	0.00112***	0.000888***	0.000937***	0.00107***	0.00110***	0.00136***	0.000985***
	(0.000331)	(0.000348)	(0.000336)	(0.000334)	(0.000336)	(0.000334)	(0.000333)	(0.000365)	(0.000374)
$\widehat{m}_{simple}(X)$ (personal features)		0.00233***							
		(0.000747)							
$\widehat{m}_{simple}(X)$ (age)			-0.000658						
			(0.00135)						
$\widehat{m}_{simple}(X)$ (gender)				0.00729***					
				(0.00178)					
$\widehat{m}_{simple}(X)$ (education)					0.00561***				
					(0.00212)				
$\widehat{m}_{simple}(X)$ (nationality)						0.000803			
						(0.00199)			
$\widehat{m}_{simple}(X)$ (relatives)							-0.000715		
							(0.00453)		
$\widehat{m}_{simple}(X)$ (industry-location)								-0.00121*	
								(0.000702)	
$\widehat{m}_{simple}(X)$ (startup traction)									0.000414
									(0.000681)
Observations	37,853	37,853	37,853	37,853	37,853	37,853	37,853	37,853	37,853
R-squared	0.000	0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.000

Standard errors in parentheses

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

Table 6: Full vs. simple models. We train simple estimators $\widehat{m}_{simple}(X)$ to predict new firms' operating performance five years after creation. 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 our predictions using $\widehat{m}(X)$ (full model) and on predictions from the simple models $\widehat{m}_{simple}(X)$, which take as inputs only a subset of features. Estimator $\widehat{m}_{simple}(X)$ (personal features) is trained taking as inputs the founding entrepreneur's age, gender, education, nationality, and whether there are entrepreneurs among her relatives. Estimator $\widehat{m}_{simple}(X)$ (startup traction) is trained taking as inputs the total number of workers, the number of clients, and the client's location.

Founder's Age	4.463***
	(5.16)
Female	-0.273***
	(-7.53)
Education (1-5 scale)	-1.619***
	(-18.26)
Created Firm Before	0.312***
	(9.36)
Paris-based	0.178***
	(6.67)
Innovation	0.198***
	(5.31)
Growth-minded	0.0687*
	(1.79)
Relatives	-0.162***
	(-4.32)

t statistics in parentheses

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

Table 7: Over-weighted entrepreneur characteristics. We sort firms into quintiles according to their predicted performance and their predicted chance 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 characteristics between these two groups.

Bibliography

Acemoglu, Daron. 2021. "Harms of AI." National Bureau of Economic Research.

Acharya, Viral V, Oliver F Gottschalg, Moritz Hahn, and Conor Kehoe. 2013. "Corporate governance and value creation: Evidence from private equity." *The 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." *The Journal of Finance*, 72(2): 509–538.

Bernstein, Shai, Xavier Giroud, and Richard Townsend. 2015. "The Impact of Venture Capital Monitoring." *The Journal of Finance*, 71: n/a–n/a.

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

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.*

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." *The Review of Financial Studies*, 24(12): 4037–4090.

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

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

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

Cong, Lin William, and Yizhou Xiao. 2021. "Persistent Blessings of Luck: Theory and an Application to Venture Capital." *The 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." *The 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, Kansas City, MO: February.*

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." *The 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." *The 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." *The Journal of Finance*, 62(1): 251–301.

Hombert, Johan, Antoinette Schoar, David Sraer, and David Thesmar. 2016. "Does Unemployment Insurance Change the Selection into Entrepreneurship?" In *Measuring Entrepreneurial Businesses: Current Knowledge and Challenges*. 351–369. University of Chicago Press.

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." *Available at SSRN*.

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

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.” *The Journal of Finance*, 59(5): 2177–2210.

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

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

Kearns, Michael, and Aaron Roth. 2019. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design.* USA:Oxford University Press, Inc.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. “Human decisions and machine predictions.” *The quarterly journal of economics*, 133(1): 237–293.

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. “Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System.” National Bureau of Economic Research.

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

Mullainathan, Sendhil, and Ziad Obermeyer. 2021. “Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care.” *The Quarterly Journal of Economics*. qjab046.

Piacentino, Giorgia. 2019. “Venture Capital and Capital Allocation.” *The Journal of Finance*, 74(3): 1261–1314.

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. 2018. “Entrepreneurial human capital and firm dynamics.” *Available at SSRN 3280925*.

Raina, Sahil. 2019. “VCs, founders, and the performance gender gap.” *Finance Down Under 2017 Building on the Best from the Cellars of Finance*.

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