Predicting Entrepreneurial Success is Hard:

Evidence from a Business Plan Competition in Nigeria[#]

David McKenzie[1]

Dario Sansone[2]

Abstract

We compare the absolute and relative performance of three approaches to predicting outcomes for entrants in a business plan competition in Nigeria: Business plan scores from judges, simple ad-hoc prediction models used by researchers, and machine learning approaches. We find that i) business plan scores from judges are uncorrelated with business survival, employment, sales, or profits three years later; ii) a few key characteristics of entrepreneurs such as gender, age, ability, and business sector do have some predictive power for future outcomes; iii) modern machine learning methods do not offer noticeable improvements; iv) the overall predictive power of all approaches is very low, highlighting the fundamental difficulty of picking competition winners.

Keywords: entrepreneurship, machine learning, business plans, Nigeria

JEL: C53, L26, M13

[#] We are grateful to the editor Jeremy Magruder, two anonymous referees, Eric Auerbach, Keisuke Hirano, William Maloney, Givi Melkadze, and Allison Stashko for their helpful comments. Participants at the CSAE, AI and Development, and WADES conferences and seminar participants at Georgetown University, The World Bank and UIUC provided useful feedback. Funding from the Strategic Research Program (SRP) of the World Bank is gratefully acknowledged. An earlier version of this paper has circulated under the title "Man vs. Machine in Predicting Successful Entrepreneurs: Evidence from a Business Plan Competition in Nigeria".

[1] Development Research Group, The World Bank Group. E-mail: dmckenzie@worldbank.org

[2] Georgetown University. E-mail: ds1289@georgetown.edu

1. Introduction

Millions of small businesses are started every year in developing countries. However, more than a quarter of these die within their first year (McKenzie and Paffhausen, 2018), while only a small subset of firms grows rapidly, creating disproportionate value in terms of employment and incomes (Olafsen and Cook, 2016). The ability to identify ex ante which firms will succeed is of key interest to investors seeking to maximize returns, determining the extent to which capital can be allocated to the highest return projects. Being able to identify these high growth potential firms is also important for governments seeking to target programs to these firms (OECD, 2010), and is of interest to researchers seeking to characterize what makes a successful entrepreneur. Moreover, if the characteristics that are predictive of high growth are malleable (and additional studies show a causal link between such characteristics and firm outcomes), this research can also spur policy efforts to attempt to change these attributes in individuals lacking them.

Business plan competitions are increasingly used in both developed and developing countries to attempt to spur high growth entrepreneurship. These competitions typically attract contestants with growth prospects that are much higher than the average firm in the economy. But then the key question is whether one can predict which firms from among these applicants have better growth prospects. We use data from applicants to Nigeria's YouWiN! program, the world's largest business plan competition (Kopf, 2015), to answer this question. We compare the absolute and relative performance of three different methods for identifying which firms will be most successful: the standard approach of relying on expert judges, simple ad hoc prediction models used by researchers that employ a small number of variables collected in a baseline survey, and modern machine learning (ML) methods that consider 566 possible predictors and non-linear ways of combining them.

Business plan competitions typically rely on expert judges to score proposals, with higher scores given to those businesses judges view as having higher likelihoods of success. This approach has the advantage of using context-specific knowledge and enabling holistic judgement of business plans. However, using judges to score proposals can be costly and time-consuming, and there are also concerns that human evaluations can sometimes lead to discriminatory decisions (Blanchflower et al., 2003), or be driven by overconfidence (Zacharakis and Shepherd, 2001). One alternative may then be to attempt to predict which businesses will succeed based on simple regression models that include survey variables like gender, age, and ability suggested by theory and researcher judgement (Fafchamps and Woodruff, 2017; McKenzie, 2015). However, the selection of which variables to include in these models is ad hoc and, since it reflects a researcher's opinion, may omit important determinants of future success. Machine learning techniques offer an alternative approach that is less reliant on human judgement, and are designed specifically for out-of-sample prediction (Kleinberg et al., 2015). They offer the potential for better prediction performance by using high-dimensional data reduction methods to identify patterns that might be too subtle to be detected by human observations (Luca et al., 2016). This could then offer a cheaper and more effective alternative to identifying which businesses will succeed.

Using data on more than 1,050 winners and more than 1,050 non-winners, we find that the business plan scores from judges do not significantly predict survival, or employment, sales and profits outcomes three years after entry. This is not due to the scores all being similar to one another, nor to the outcomes being similar for all firms. For example, among the non-winners, a firm at the 90th percentile has 13 times the employment and 2.4 times the profits of a firm at the 10th percentile. Rather, conditional on getting through the first phase of the competition, the scores of judges do not predict which firms will be more successful.

We then use several ad hoc prediction models. One set of these models are simple univariate prediction models using variables like gender, age, education, and ability that the literature has suggested might be related to firm performance. A second set of models comes from a simple logit model based on this literature, used by McKenzie (2015), and from a similar model used by Fafchamps and Woodruff (2017). We do find that some observable characteristics of entrepreneurs help predict future success, with high-ability males in their thirties doing better. Nevertheless, the overall accuracy of these models is still low.

These results are then contrasted with out-of-sample predictions from three modern ML approaches - LASSO, Support Vector Machines, and Boosted Regression – that use data reduction techniques and highly nonlinear functions to make predictions starting with 566 possible predictors. These methods may end up choosing interactions of variables and survey responses that would be unlikely to be used by human experts. However, we find that the overall accuracy of these models is similar to that of the simpler ad hoc models, and thus also low. Although we find that machine learning is often unable to beat simple predictors or simple models from economists when it comes to predicting average performance, we do find some role for machine learning and simple models in identifying the top tail of high-growth firms. Nevertheless, even our best models are typically only able to ex ante identify only 20 percent of the firms that will end up in the top decile of the outcome distribution.

We then examine different potential explanations for why the judges and machine-learning algorithms do not perform better. We discuss whether judges were trying to predict a different outcome, whether the use of anonymized plans restricted their ability to statistically discriminate, whether they lacked population-specific knowledge, and whether they are better at telling apart the bottom tail. None of these appear to be the main reasons for their poor performance. We then examine whether the machine-learning performance is hampered by insufficient sample, by a lack of independent variation in the inputs, or by poor model choice. Again, these do not seem to be the main factors behind the poor performance.

Taken together, these results point to the fundamental difficulty of identifying in advance which entrepreneurs will be more successful from among a sample that has already made it through a first phase of a business plan competition. This is a feature of the fundamental riskiness inherent in entrepreneurship (Hall and Woodward, 2010), and it is consistent with the inability of

professional investors to know in advance whether a technology or venture will succeed, conditional on it passing some initial screening (Kerr et al., 2014; Hall and Woodward, 2010).

This paper builds on two existing literatures. The first is a literature on predicting which firms will succeed. Most of the existing literature looks at start-ups and venture-backed firms in developed countries, and while some of these studies find that judges' scores have some predictive power (e.g. Astebro and Elhedhli, 2006; Scott et al., 2015), they also point to the immense difficulty of identifying who will be more successful (Kerr et al., 2014; Nanda, 2016). Data analysts have also started to predict which new tech firms will be successful by looking at which sub-sectors have attracted the highest investments by venture capitalists and which companies have raised funds from the top-performing investment funds in the world (Ricadela, 2009; Reddy, 2017). There has been much less study of this issue in developing countries (Nikolova et al., 2011). An important exception is Fafchamps and Woodruff (2017), who compare the performance of judges' evaluations to survey-based measures, and find that both have some (independent) predictive power among business plan contestants in Ghana. We build on their work with a much larger sample, and through the incorporation of machine learning.

Secondly, we contribute to a growing literature that uses machine learning in economics (Mullainathan and Spiess, 2017). Kleinberg et al. (2015) note that this method may be particularly useful for a number of policy-relevant issues that are prediction problems, rather than causal inference questions. Economists have started applying these methods to a number of different settings such as predicting denial in home mortgage applications (Varian, 2014), quality of police and teachers (Chalfin et al., 2016), poverty from satellite data (Jean et al., 2016), conflicts in developing countries (Celiku and Kraay, 2017), if defendants will commit crimes when released on bail (Kleinberg et al., 2017), and high school and college dropout (Aulck et al., 2016; Sansone, 2018). Our work shows the limits of this approach when it comes to predicting entrepreneurial success, and that machine learning need not necessarily improve on the performance of simpler models that have been more commonly used by economists.

2. Context and Data

Business plan competitions are a common approach for attempting to attract and select entrepreneurs with promising business ideas, and have been increasingly used in developing countries as part of government efforts to support entrepreneurship. The typical competition attracts applicants with the promise of funding, training or mentoring, and/or publicity for the winners. Applications are then usually subjected to an initial screening, with the most promising plans then being scored and prizes awarded to the top-scoring entries.

2.1 The YouWiN! Competition

Our context is the largest of these competitions, the *Youth Enterprise with Innovation in Nigeria* (YouWiN!) competition. This competition was open to all Nigerians aged 40 or younger, who could apply to create a new firm or expand an existing one. We work with the first year of the

program, which had a closing date for initial applications of November 25, 2011. The competition attracted 23,844 applications. A first stage basic scoring selected the top 6,000 applicants, who were offered a 4-day business plan training course. Those who attended the course were then asked to submit a detailed business plan, with 4,517 proposals received. These proposals were then scored by judges - in a process described below - with the highest scores used to select 2,400 semi-finalists. 1,200 among those were then chosen as winners and received grants averaging almost US$50,000 each. 480 of these winners were chosen on the basis of having the top scores nationwide or within their region, while 720 winners were randomly chosen from the remaining semi-finalists.[3]

McKenzie (2017) uses this random allocation to measure the impact of winning this competition, and provides more details on this competition and selection process. His analysis found that being randomly selected as a competition winner resulted in greater firm entry from those with new ideas, higher survival of those with existing businesses, and higher sales, profits, and employment in their firms. It shows that the competition was able to attract firms that, on average, had high growth prospects. Our goal in this paper is to use this setting to see whether the judge scores, simple heuristic models, or machine learning approaches can identify ex ante which firms from among these applicants will be most likely to survive and grow.

2.2 Data

Our starting sample consists of 2,506 firms that are comprised of 475 non-experimental winners (the national and regional winners), 729 experimental winners, 1,103 semi-finalists that are in the control group, and 199 firms that submitted business plans but did not score high enough for the semi-finals.[4] By considering non-winners in addition to winners, we are able to test how well human experts and machine learning approaches do in forecasting firm growth for both a sample of firms that receive financial support, and for a sample of firms that are not receiving grant support. The majority of this sample consists of individuals seeking to create new firms (63%), with 37% coming from existing firms looking to expand. For each of these firms we have data from their initial online application, from the submission of their business plans, the score provided by judges on their business plans, and short-term follow-up survey data on some ability and personality measures that are assumed to be stable over time. We discuss each in turn, and provide a full description of the different variables in Appendix A.1.

The application form required individuals to provide personal background such as gender, age, education, and location, along with written answers to questions like "why do you want to be an

[3] Selection of winners was stratified by whether the proposal was for an existing or new firm, and by region. 300 National winners (225 existing, 75 new) were selected as the top overall, then 30 regional winners were chosen from each of the six regions from a shortlist of 45 existing firms and 15 new firms per region. The 720 experimental winners were chosen as 120 per region, again oversampling existing firm applicants. See McKenzie (2017) for additional details.

[4] McKenzie (2017) dropped 5 non-experimental winners who were disqualified by the competition, and includes 9 experimental winners who were disqualified. We drop these 14 firms, along with 9 control group firms that were non-randomly chosen as winners. In addition, McKenzie (2017) attempted to survey 624 firms that were not selected for the business plan training. Since these firms did not submit business plans, they do not have business plan scores and are thus excluded from our analysis.

entrepreneur?" and "how did you get your business idea?". We construct several simple measures of the quality of these written responses, such as the length of answer required, whether the application is written in all capital letters, and whether they claim to face no competition. We also use administrative data on the timing of when applications were submitted relative to the deadline.

The business plan submission contained more detailed information, including a short baseline survey instrument. This collected additional information about the background of the entrepreneur, including marital status, family composition, languages spoken, whether they had received specific types of business training, employment history, time spent abroad, and ownership of different household assets. It elicited risk preferences, asked about motivations for wanting to operate a business, and measured self-confidence in the ability to carry out different entrepreneurial tasks. More detailed information about the business they wanted to operate or their existing business they intended to expand included the name of the business, business sector, formal status, access to finance, taxes paid, challenges facing the business, and projected outcomes for turnover and employment. Unfortunately, we do not have access to the written text of the business plans, and so cannot employ textual analysis of the plans themselves.

The initial application and business plan submission were all done online. While they contain a rich set of data on these applicants, there were some measures which needed to be collected face-to-face with explanations, and so were left for follow-up surveys. Two ability measures were collected in the follow-up surveys: digit-span recall, and a Raven progressive matrices test of abstract reasoning. In addition, the grit measure of Duckworth et al. (2007) was also collected during these short-term follow-ups. We assume that these measures are stable over one to two years, and therefore include them as potential characteristics that could be used to predict future outcomes had they been collected at the time of application. To be consistent with the set of predictors included in Fafchamps and Woodruff, 2017, we also include two measures of current and expected life satisfaction, as well as additional measures of self-confidence.

2.3 Measuring Success

McKenzie (2017) tracked these firms in three annual follow-up surveys, and then in a five-year follow-up during a period of recession in Nigeria. We use data from the three-year follow-up survey, fielded between September 2014 and February 2015, to measure business success in this paper. This survey had a high response rate, with data on operating status and employment available for 2,128 of the 2,492 firms in our sample of interest (85.4%).[5] We believe this is a more useful test than the five-year follow-up, which, in addition to a lower response rate, occurs in an atypical period in which Nigeria suffered its worst economic performance in thirty years, resulting

[5] Appendix 8 of McKenzie (2017) examines the characteristics of this attrition. It shows that attrition rates were 6% higher for non-winners than winners, but that this did not change the balance on baseline observables (including business scores). Our analysis is done separately for winners and non-winners, and we cannot reject that round 3 attrition is orthogonal to the business plan score for any of our sample groups. Coupled with the relatively low absolute level of attrition, this suggests that attrition is unlikely to drive our results, and we thus focus on the sample of firms interviewed in this three-year follow-up.

in many firms struggling. Moreover, three years is the time horizon over which business plans were prepared (see next section).

We focus on four metrics of success: business operation and survival, total employment, sales, and profits. Appendix A.1 describes how these are measured. 93.6% of the competition winners were operating businesses three years after applying, compared to 58.4% of the control group and other non-winners. Figure 1 then plots the distributions of employment, sales, and profits outcomes for the winners and non-winners. We see substantial dispersion in the outcomes of these firms three years after they applied for the program. For example, among winning firms in operation, a firm at the 90th percentile has 20 employees, five times the level of a firm at the 10th percentile of employment; and a firm at the 90th percentile of profits earns 500,000 naira per month (USD $2,730), around four times the level of a firm at the median, and 100 times that of a firm at the 10th percentile. Likewise, while average levels of these outcomes are lower for non-winners, there is also a long right tail for this group, with some firms doing substantially better than others. Given how much variability there is in these outcomes, it is therefore of interest to examine whether the more successful of these firms could have been identified in advance.

3. Human Approaches to Predicting Business Success

The standard approach to choosing businesses to support is to rely on the human judgement of experts. We consider prediction by two types of experts. The first group of experts is composed by the judges in the business plan competition, while the second group is composed by economists who generate their prediction models based on their expert judgements of what variables should predict business success.

3.1 Using Judges to Evaluate Business Plans

The business plans prepared by applicants were lengthy documents, averaging 13,000 words of narrative coupled with spreadsheets of financial projections. They contained a detailed description of the proposed business, market and industry analysis, production plans, operations and logistic details, customer analysis, financial analysis, competitive environment, and other details, along with financing plans, discussion of what the grant would be used for, discussion of the risks, threats, and opportunities, and sales and growth projections for the next three years.

Scoring for the competition was overseen by an independent project unit supported by DFID and the World Bank, with significant effort taken to ensure that the scoring was impartial and rigorous. Scoring was done by the Enterprise Development Center (EDC), a sister unit of the Lagos Business School within the Pan-Atlantic University in Lagos; and by the Nigerian office of PriceWaterhouseCoopers (PwC). Each selected 10 judges, who were experts in business and entrepreneurship and knowledgeable about local conditions. These judges included business owners and alumni of Lagos Business School, as well as experts involved in working directly with successful businesses in Nigeria. These judges were given training, which included marking

sample plans and receiving feedback. Markers had also access to a forum where they could seek advice from their peers and resolve issues.

In order to ensure impartiality, and to be able to score such a large number of applications, applicants did not make in-person pitches to judges, but were scored only on the basis of their business plan. These plans were anonymized. Judges were given a set template for scoring, designed by the two leads from EDC and PwC, which assigned marks on 10 different criteria covering the management skills and background of the owner, the business idea and market, financial sustainability and viability, job creation potential, and a capstone score to allow for their overall assessment of the chance of business success (Appendix A.1). Importantly, these judges were asked to assess which firms were most likely to succeed, not which firms would have the largest treatment effects from receiving a grant from the competition. In particular, judges were told that they were looking for submissions that "stand out in their innovativeness, feasibility and job creation potential". A typical plan took 30 to 45 minutes to mark. A random sample of 204 of the business plans were then re-scored by an independent team from Plymouth Business School and any significant disparities reviewed. The mean score in this sample was within 0.4 points out of 100 of the mean from the original judges, with this difference not statistically significant, and their quality assurance concluded that "the marking criteria deployed by LBS/EDC in this case did offer a broad mechanism to differentiate leading business plans from less convincing business plans".

Figure 2 plots the distribution of business plan scores according to their competition outcome. Note that the distribution for non-experimental winners overlaps with that of the experimental winners and losers because of the preference given towards existing firm applicants, and because of differences across regions in the scoring threshold needed to be a regional winner. We see that that distributions of scores are wide for both the winners and the non-winners in our sample. The winning scores range from 30 to 91 out of 100, with a mean of 56 and standard deviation of 12. The scores of the non-winners in our sample range from 2 to 73 out of 100, with a mean of 51 and a standard deviation of 11. This shows that the judges did view firms as substantially different from one another in terms of their potential for business success. Moreover, we have seen that the firms also had wide dispersion in subsequent outcomes. This then raises the question of whether the judges were able to identify in advance which firms were more likely to succeed.

3.2 Do Judges' Scores Predict Business Success?

Since the winning firms received substantial business grants, which were shown in McKenzie (2017) to have a causal impact on firm outcomes, we separate all our analysis by whether or not the firm was a competition winner. We are then interested in seeing whether the winners scored more highly by the judges did better than winners with lower scores, and likewise whether the non-winners with higher business plan scores did better than those with lower business plan scores.

Figure 3 plots the relationship between our four key outcomes (business operation and survival, sales, profits, employment) and the judges' scores for the non-winners. We see that the likelihood

of operating a firm three years later is actually slightly lower for those in the upper quintiles of the business plan score. We then fit both unconditional lowess (which codes employment, profits, and sales, as zero for firms not operating) and conditional lowess (which only considers firms in operation) and plot these lines along with the scatterplots for sales, profits and employment. It is apparent that the business plan scores explain very little of the variation in these outcomes, with the fitted lines showing, if anything, poorer results for those with higher scores.

The first column of Table 1 (survival), Table 2 (employment), Table 3 (profits), and Table A1 (sales) then tests whether the relationship between business plan scores and firm outcomes is statistically significant for these non-winners by estimating a logit (survival) or least squares regression (other outcomes) of the outcome on the business plan score and an indicator for whether the application was for an existing firm compared to a new firm. The regressions are unconditional, coding outcomes as zero for non-operating firms. For all four outcomes, we see the business score has a small, negative, and not statistically significant association. The R^2 are below 0.04, with even this small amount of variation explained coming largely from the dummy for existing firm status. Judges' scores therefore do not predict which of these non-winners are more likely to succeed.

Figure 4 then plots the associations between business outcomes and the judges' scores for the winners. We again see the business plan scores explain very little of the variation in these outcomes, with the fitted lines showing a fairly flat relationship for profits and sales, and a slightly upward sloping relationship for employment. Column 4 of Tables 1-3 and A1 then shows the fitted logit and regression relationship, after controlling for whether the application was for an existing firm. The amount of grants the winners received varied among firms, with an average of US$48,991 and standard deviation of $17,546. There is a significant positive relationship between business score and amount received, and more capital has a direct impact on firm outcomes. Therefore, column 5 of Tables 1-3 and A1 also controls for the logarithm of the grant awarded. We see the judges' score has a small and statistically insignificant impact on survival, profits, and sales, with or without this control for grant amount. In contrast, we see in Table 2 that the judges' scores are significantly correlated with employment three years later in these firms, although this relationship weakens and is no longer statistically significant after controlling for the grant amount. Moreover, the share of variation in employment accounted for by these judges' scores is still very low, with an adjusted R^2 of 0.02 in column 4 of Table 2.

This analysis shows that the overall scores of judges are uncorrelated with business success for non-winners, and at most weakly correlated with future employment outcomes, but not other outcomes, for winners. As an additional robustness check, Table A2 shows that this result continues to hold when we condition on judge or region fixed effects.

The overall score aggregates multiple scores on 10 subcomponents, and so it is of interest to examine whether this aggregation masks predictive ability of some subcomponent of the score. In Appendix Table A3 we test whether the ten subcomponents of the business plan score are jointly associated with future outcomes, as well as examining individual subcomponents. For the non-

winners, we cannot reject that the ten sub-scores are jointly unrelated to survival, sales, or profits, but we can do so for employment (p=0.039). We also find that the score out of 10 for "ability to manage" is positively and significantly related to all four outcomes - with p-values of 0.003, 0.001, 0.008 and 0.017 respectively - and continues to be so even after applying a Bonferroni or step-down correction for multiple testing. This effect mainly appears to be coming through the extensive margin of predicting whether or not the individual will have a firm in operation. The employment sub-score is also weakly predictive of employment (p=0.055).

For the winners, this "ability to manage" score is not significant for any outcome, and we cannot reject that the ten sub-scores are jointly unrelated to survival, sales, or profits. The employment score does significantly predict employment among the winners (p=0.001), and we can reject that the ten sub-scores are jointly unrelated to the employment outcome (p=0.050). As a reference, a one standard deviation (6.0) increase in this employment sub-score is associated with the winning firm having 1.1 more workers three years later. Notably, the capstone score - which is intended to reflect the judge's overall assessment of promise not picked up in the other scores - has no significant association with future business success for either winners or non-winners, and it is even negatively associated with future profits for winners.

3.3. Using Human Expert Judgements to Predict Business Success

An alternative human approach to predicting business success is for economists to use their expert judgement to choose characteristics measured in the baseline survey (or that are measured later but assumed to be time-invariant) that they think may predict outcomes, and then to fit a simple logit or OLS regression model as a function of these characteristics.

This approach was carried out in Appendix 18 in the working paper of McKenzie (2015), but dropped for space reasons in the published version (McKenzie, 2017). It uses principal components and aggregate index measures for data reduction: a household wealth measure is constructed as the first principal component of 20 durable assets; an ability measure is constructed as the first principal component of the Raven test (itself an aggregate of responses on 12 questions) and digit-span recall scores; and a grit measure is an average of responses on 12 questions proposed by Duckworth et al. (2007). In addition to this, gender, age, attitude towards risk, a dummy for having graduate education, whether the applicant has worked abroad, indicators for the six geographic regions, and indicators for the three most common industries (agricultural business, manufacturing business, IT business) were selected as variables that the literature and an understanding of the business setting suggested might help predict future performance. No model selection criteria were used to select these variables: they were just proposed as an ad hoc judgement of what might matter for predicting business outcomes.

A similar approach is used by Fafchamps and Woodruff (2017) for a business plan competition they ran in Ghana. They propose a set of core measures that they think are likely to be viewed as potentially important: the ability of the owner, past borrowing activity, and management practices, along with controls for gender, age, sector, and location. They then add attitudinal questions based

on the reason for going into business, and on optimism and control over life. Again, no model selection criteria were used to select these variables.

We examine the extent to which these proposed variables coming from baseline survey measures are associated with future business outcomes in the remaining columns of Tables 1-3 and A1. Columns 2 and 6 use the variables selected in McKenzie (2015), while Columns 3 and 7 attempt to match as closely as possible the variables used in Fafchamps and Woodruff (2017). Appendix A.1 describes in detail how these variables were constructed. These regressions allow us to examine whether the proposed models are useful for predicting business outcomes within sample – in the next section we will examine out-of-sample performance to allow comparison with machine learning methods. We also include the business plan score in these regressions so that we can test whether these survey measures add value beyond that contained in the judges' scores, and conversely, whether the judges' scores now contain information once we have conditioned on other features of the applicants.

Consider predicting outcomes for the non-winners. We can strongly reject that the McKenzie (2015) set of regressors, and the Fafchamps and Woodruff (2017) set of regressors, are not associated with each of our four business outcomes ($p<0.001$). This greater performance relative to the judges comes largely from a few characteristics of the owner. First, consistent with a broad body of literature (e.g. Boden and Nucci 2000; Bruhn 2009; Robb and Watson 2012; McKenzie and Paffhausen 2018), female applicants who do not win are less likely to start or continue businesses than men, and operate firms that are smaller in size and less profitable.

Second, while the business plan competition was designed for youth, it maintained an expansive definition of youth which ranged from age 18 to 40, with a median age of 30 among our non-winners. We see that older applicants were more likely to be running firms three years later among the non-winners, and to run firms of larger size, sales, and profitability. This is consistent with a return to work and life experience, and with evidence from other developing countries that businesses have the lowest failure rates when the owners are middle-aged (McKenzie and Paffhausen, 2018). In contrast to pitching situations where judges see the applicants, the anonymized business plans used for scoring meant that judges did not observe gender or age, and therefore did not consider it in their assessments (and indeed, concerns about sex- or age-discrimination would occur if they did directly consider it).

Third, higher ability of the owner is associated with better performance in all four outcomes for the non-winners. Both the McKenzie (2015) measure, which is the first principal component of Raven test and digit span, and the Fafchamps and Woodruff (2017) measure, which is the first principal component of Raven, digit span, and education, are significant predictors. Examining the sub-components, the Raven test score is the most important contributor here. We then also see some role for business sector in predicting some outcomes: e.g., IT firms tend to hire fewer workers. Our measure of optimism is also positively related with survival, profits, and sales.

These models based on few key variables selected by economists do better than the judges in predicting outcomes for the non-winners. The business plan scores remain insignificant, even conditioning on the variables in these models. However, these models still only explain a small fraction of the variation in our key outcomes: the adjusted R^2 are 0.04-0.05 for employment, 0.10-0.13 for sales, and 0.08-0.12 for profits (the unadjusted R^2 are also low, always below 0.15).[6] Moreover, much of the predictive ability comes from the extensive margin of whether firms are operating or not. Appendix Table A4 shows that these R^2 are even lower when we consider only the ability of these predictors to explain the variation in outcomes among firms in operation.

Since almost all of the winners are operating firms after three years (93.5%), there is not much variation in this outcome. We cannot reject that the McKenzie (2015) or Fafchamps and Woodruff (2017) sets of variables are jointly orthogonal to this outcome. We do find some predictive power for employment, sales, and profits, but the R^2 are only 0.07-0.08, 0.02 and 0.02 respectively (the unadjusted R^2 are always below 0.10). Several of the attributes that helped predict outcomes for non-winners are much less predictive of outcomes for winners: gender, while negative, is only significant for profitability and sales; ability is never significant, and has negative coefficients in all but one specification; and being older predicts higher employment, but not higher sales or profits. Sector continues to matter, with IT firms hiring fewer workers, and retail firms having higher profits and sales. We do see that the business plan score of judges continues to be a significant predictor of employment at the 5 percent level, even after controlling for the variables in these models.

It is also notable from Tables 1-3 and A1 that some characteristics that were expected to help distinguish among business owners did not have significant predictive power. A first example is education, which is negatively associated with employment among winners, and otherwise not significant. One possible reason for not seeing the expected positive return to education (e.g. Michelacci and Schivardi 2016; Queiro 2016) is that the earlier stages of the competition (including the need to apply online) appear to have selected heavily on education. Only 5.5% of Nigerian youth have university education, but 52% of the applicants, and 64% of those invited to business plan training had this level. Likewise, applicants who got through to the business plan submission stage are wealthier, and more likely to have overseas work experience, than the average Nigerian youth, yet conditional on getting to the business plan submission stage, we generally do not find these variables to be significant predictors of business outcomes.[7] The literature has also suggested risk preferences determine selection into entrepreneurship and the success of these businesses (e.g. van Praag and Cramer 2001; Hvide and Panos 2014; Skriabikova et al. 2014), yet

[6] By way of comparison, Fafchamps and Woodruff (2017) report an R^2 of 0.14 for profits, 0.23 for sales, and 0.46 for employment. However, their firms had been in operation for an average of 9 years, and do not include start-ups. Moreover, they are able to include baseline sales and profits among their predictors, which exhibit some persistence and which we do not have for start-up firms.

[7] Wealthier individuals have higher profits and sales as winners (at the 5 percent significance level), while we find overseas work experience has a negative association with profits and survival for winners (at the 10 percent level), while these variables are not significant for other outcomes for winners, or for any outcome for non-winners.

we see no significant impact of our risk preference measure. Nor do we see any predictive power for the grit measure of Duckworth et al. (2007).

4. Machine Learning Approaches to Predicting Business Success

The human prediction approaches rely on a small set of variables, and on methods like principal components and simple aggregation for data reduction. Yet, in practice, the application and baseline data contain a much richer set of possible predictors. Many of these come from just considering a lot more variables, but the number of predictors becomes even larger once we consider how responses to certain questions should be coded. We discuss a few specific examples, and then the total number of potential predictors we consider.

A first example comes from questions with Likert scales. For example, applicants were asked a set of self-efficacy questions about their confidence in carrying out nine different business tasks, such as "find and hire good employees to expand your business". They had six possible responses to their confidence on this task: not at all confident, somewhat confident, quite confident, very confident, don't know, or not applicable. Rather than trying to code these as binary variables and then aggregate, an alternative is to code each task and response as a separate dummy variable (e.g. being not at all confident that they can estimate accurately the costs of a new project). This yields 54 potential predictors.

A second example comes from the Raven test. Applicants were given 12 different puzzles to solve, and had a choice from among 8 options each time, with one of these options correct. The standard approach, used in the models estimated in the previous section, is to aggregate these into a single score out of 12. Yet some questions may be potentially more useful in predicting business success than others, and different types of wrong answers may reflect different abstract reasoning deficiencies. For example, question 6 had 39% choose the correct answer, five other answers were chosen by 9% to 11% of individuals each, and the remaining two answers were all chosen by between 5% and 7% each. Coding each question and answer as a separate dummy variable then yields 96 potential predictors.

A third example comes from the business sector that the applicant proposes for their new business venture, or has for their existing business. They were asked to select from 32 different business sector categories. In the ad hoc model of McKenzie (2015), four of these were aggregated together to form a manufacturing dummy, and two other sectors were also chosen, while the rest were lumped together as the base comparison group. In contrast, we consider all 32 categories as possible predictors here.

However, the increase in the number of variables considered does not just come from dis-aggregation of variables into more categories, but also from considering a wide range of additional variables collected during the initial application and business plan submission. As noted in Section 2.2 and detailed in Appendix 1, these include measures of answer quality, family composition, languages spoken, how the business was named (Belenzon et al., 2017; Guzman and Stern, 2017),

business and employment background of the owner, registration and tax status of existing firms, and the owner's projections of firm growth. These are all variables which may plausibly be correlated with business success.

Following Mullainathan and Spiess (2017), we do not drop redundant variables, e.g. aggregate indexes such as the grit or ability, since they could be useful to obtain better predictions with less complexity. The net result of this is that there are 566 possible predictors that we can consider. This figure includes the initial application score and the ten different subcomponent scores from the judges, in addition to the survey data. In order to make a fair comparison, we initially do not include the business plan scores as inputs in the machine learning algorithms. Moreover, if we also start considering interaction terms between some of these predictors, the number of variables to consider can easily exceed the number of firms for which we have data. It is then not possible to include all of these variables in a standard OLS regression, or logit model. Machine learning provides tools for using these high-dimensional data.

4.1 The Machine Learning Algorithms

We use three different machine learning approaches. We briefly summarize each here, and provide more details in Appendix A.3. A more comprehensive review of the tools available to practitioners is provided by Hastie et al. (2009), Ng (2016), as well as Mullainathan and Spiess (2017).

The first method is **LASSO** (Least Absolute Shrinkage and Selection Operator). This method adds a penalization term to the OLS objective function:

$$\hat{\beta}(\lambda) = \underset{\beta \epsilon \mathbb{R}^k}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{k} |\beta_j|$$

Where k is the number of potential predictors (potentially larger than the number of observations, n), and λ is the penalization term. In practice, this selects variables with high predictive power, and shrinks their coefficients, while constraining all other variables to have coefficients of zero. As explained in Section 4.2, we choose the penalization term λ via cross-validation.

The second method is **Support Vector Machines (SVM)** for our business operation outcome, and the extension of **Support Vector Regression (SVR)** for our continuous outcomes of employment, sales, and profits (Guenther and Schonlau, 2016). SVM aims to classify the data into two groups (in our case, into firms that will be operating and those that will not) by means of a hyperplane in a high-dimensional space. It can be written as a modified penalized logistic regression that solves the following objective function:

$$\hat{\beta}(C) = \underset{\beta \epsilon \mathbb{R}^k}{\operatorname{argmin}} C_1 \left[\sum_{i=1}^{n} y_i \, max\{0, 1 - K_i'\beta\} + (1 - y_i) \, max\{0, K_i'\beta - 1\} \right] + \|\beta\|_2$$

There are four terms here that differ from that in the standard logistic regression, and which allow SVM greater flexibility in covering a wide range of functional forms and dealing with high-

dimensional data. In addition to the penalization factors, C_1 and $\|\beta\|_2$; there is a kernel K which controls the smoothness of the fitted curve, and a "soft margin" that comes from the $max(0,.)$ component, which allows for some observations to be misclassified close to the separating margin. We consider a Gaussian kernel, and use cross-validation to choose the penalization term and the kernel smoothing parameter.

SVM has been found to achieve better performance than a number of other machine learning algorithms in some applications (Maroco et al., 2011). Nevertheless, given the similarity in terms of objective function, in other cases performances have been found to be similar to those obtained with logistic regressions (Verplancke et al., 2008). One downside of SVM and SVR are that the parameters are difficult to interpret, and so we will not be able to say which variables matter most for determining the predictions made by this method.

The final method we consider is **Boosting** (also called Boosted Regression), using the gradient boosting algorithm of Friedman et al. (2000). Gradient boosting is an ensemble method, which builds a classifier out of a large number of smaller classifiers. Like random forests, it works by combining many regression trees together. Key differences are that boosting does this sequentially, where in each iteration, observations that were misclassified by the previous classifier are given larger weights; and that boosting typically uses shorter trees, including tree stumps (a single split of the data), aggregating over more of them than random forests.[8] We consider trees with up to 8 splits, thus allowing up to 8-way interactions between variables in a very flexible way.

Performance is improved by using bagging, which uses only a random subset of the training data set to build the tree in each iteration. To avoid over-fitting, i.e. the risk of estimating a model which fits a training sample very well, but may fail to properly work with new samples, it is also possible to introduce a shrinkage parameter. This reduces the contribution of each additional tree, thus decreasing the impact of an over-fitted tree. Boosting has been found to have superior performances than a number of other machine learning algorithms in many simulations (Bauer et al. 1999; Friedman et al. 2000) and was used by Chalfin et al. (2016) in their work on predicting police hiring.

4.2 Estimation by Machine Learning

We follow the recommendation of Mullainathan and Spiess (2017) and split the data into two sub-samples (separately for winners and non-winners). A training sample (80% of the data) is used to calibrate and estimate the algorithm under each of the three methods. Out-of-sample performance is reported using the hold-out sample (the remaining 20% of the data).

The algorithms are calibrated using 5-fold cross-validation.

1. Divide the 80% training sample in 5 folds.
    a. Select a possible numeric value for each parameter (e.g. 0.08 for λ in LASSO).

[8] Gradient boosting appears to perform slightly better than random forests in dimensions of 4,000 or fewer predictors, but requires more tuning (Zygmunt, 2016).

i. Train the algorithm on 4 of the 5 folds using the selected parameter values.
ii. Predict the outcome variable (e.g. profits) for the firms in the remaining fold and compute the relevant statistics (e.g. mean squared error).
iii. Repeat the above procedure five times, one for each fold.
iv. Compute the average performance in the 5 folds.

b. Repeat for each possible combination of the parameter values.
c. Select the combination of the parameter values that minimizes a given loss function (e.g. minimizing the mean squared error). These in-sample statistics are reported in Appendix A.4.

2. Train the algorithm using the 80% training sample with the selected parameter values.
3. Predict the outcome variable for the firms in the 20% hold-out sample and compute the relevant statistics (e.g. out-of-sample mean squared error). These are the out-of-sample statistics reported in Tables 4-6 and A10.
4. Compute 95% confidence intervals for these statistics using bootstrapping. It is important to mention that these are not the standard confidence intervals computed in regression models. As emphasized in Mullainathan and Spiess (2017), "these uncertainty estimates represent only variation of the hold-out sample for this fixed set of prediction functions, and not the variation of the functions themselves".

For the algorithms that do not require any tuning, such as OLS and Logit, we simply follow steps 2-4. Since almost all winners were operating firms, we only predict operation and survival for non-winners, giving seven outcomes to predict.

4.3 Assessing Accuracy

We follow standard practice in using the mean squared error (MSE) as the main criteria to compare the accuracy of different models for our continuous outcomes (employment, sales, and profits). For the sake of completeness, we have also reported the square of the Pearson correlation coefficient, i.e. the correlation between the actual and fitted dependent variable. This is equivalent to the R^2 in linear least squares regressions.

In contrast, there are several different measures used in the literature when considering the goodness of fit for binary outcomes, such as business operation and survival in our case. The starting point is usually the matrix comparing predicted outcomes with actual ones:

		Predicted values	
		0	1
Actual values	0	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)

An overall measure of the goodness of fit is the accuracy rate, which is the proportion of predictions that are correct out of all observations:

$$Accuracy = \frac{TP + TN}{Total\ number\ of\ observations}$$

The recall rate is also commonly used in the case of imbalanced data, i.e. when the number of positive values is much smaller or larger than the number of negative values:

$$Recall\ (or\ Sensitivity) = \frac{TP}{TP + FN}$$

Both LASSO and Boosted regression predict probabilities for binary outcomes (in contrast, SVM classifies observations as 1s or 0s, and does not produce probabilities). We follow convention of predicting an outcome of one (survival) when the predicted probability is greater than or equal to 0.5, and zero otherwise. This is in line with the Bayes classifier (Hastie et al., 2009): the accuracy rate is maximized by assigning each observation to the most likely class, given its predicted probabilities.

Subrahmanian and Kumar (2017) and Sansone (2018) note that the choice of criteria should depend on context and whether there are strong reasons to prefer the ability to detect true positives versus to avoid false negatives. In the context of predicting firm operation, we do not believe there are strong reasons to prefer one versus the other, and so use accuracy as our measure of performance. In Section 5.3, we consider the ability of these methods to predict the upper tail of performance on profitability, in which case we also report the recall rate.

To compare the prediction success of our machine learning approaches to those of the judges and of the models of economists shown in Section 3, we also use the same 80% training samples to fit these models, and then the 20% hold-out sample to get predictions, MSE, and accuracy rates. We fit the models shown in Tables 1-3, which predict outcomes as a function of the judges score, the McKenzie (2015) predictors, or the Fafchamps and Woodruff (2017) predictors. Unless otherwise noticed, we do not include the judge score when fitting the McKenzie (2015) or the Fafchamps and Woodruff (2017) models. For robustness, we consider two other models based on the judges scores: using just the first-round application score, or using all ten sub-scores (as in Appendix Table A3). We show the performance of the Fafchamps and Woodruff (2017) predictors with and without the business plan score included.

Finally, we also compare to the performance of seven models each based on just a single predictor (gender, age, necessity firm, grit, digit-span, Raven test, registration time for the competition[9]). It is worth emphasizing that these seven variables have not been selected at random, but based on previous research citing them as important predictors of future success. Therefore, these univariate models can be seen as simpler version of the models developed by McKenzie (2015) or Fafchamps and Woodruff (2017).

[9] Banerjee and Duflo (2014) suggest enrolling late is a marker for being disorganized and less able to complete tasks on time, which we hypothesize could be bad for business growth.

4.4 Results from Machine Learning and Human Predictions

Table 4 examines the out-of-sample performance of the different models in predicting business operation and survival for the non-winners three years after they applied to the competition. As a benchmark, the first row reports the out-of-sample accuracy rate (58.8%) of a model which contains only a constant. The next three models consider the performance of the scores from judges – the initial application score (model 2), the business plan score (model 3), or the ten sub-scores from the business plan (model 4). All three are no more accurate than just using the constant model, re-iterating the poor performance we saw in Table 1.

Models 5 through 11 just use a single predictor, and most of these models perform better than the judges. The best performance here comes from the Raven test, which is able to correctly predict whether or not a business will be operating in 67.1% of the cases. Models 12 through 16 consider the ad hoc models of the human experts described in Section 3.3. Both the McKenzie (2015) and the Fafchamps and Woodruff (2017) models have relatively high performances, with 63.9% and 67.1% accuracy respectively. We see that the accuracy of the Fafchamps and Woodruff (2017) does not improve when we also add the business plan score, and is similar whether we use probit model or a linear probability model instead of a logit function. The Fafchamps and Woodruff (2017) model provides a 14% improvement in accuracy compared to only using the business plan score of the judges, and its lower bound in the 95% confidence interval is above the mean accuracy obtained using the judges' scores. However, the Fafchamps and Woodruff (2017) model does no better than just using the Raven test score alone.

Finally, models 17 through 19 are the machine learning models. Their out-of-sample accuracy rates range from 60.2% for LASSO to 66.2% for SVM. In-sample 5-fold average accuracy rates are close to these out-of-sample accuracy rates, thus reducing any concern about over-fitting (Table A11). The ranges used to tune the ML parameters are appropriate for the sample used: selected values of the model parameters (such as λ in LASSO) are in the interior of the intervals used during cross-validation (Tables A6-A9). Figure A2 further shows that the cross-validation procedure identifies the best λ in LASSO, and that the choice of such parameter deeply affect the in-sample performances of the algorithm. Therefore, even if properly calibrated, these high-dimensional modeling approaches do no better than the ad hoc models of human experts, or than just using the Raven test score alone, in terms of predicting business survival among the non-winners. Moreover, none among these 19 models achieves outstanding performances: the upper bounds in the 95% confidence intervals do not exceed a 73.6% accuracy rate.

Tables 5-6 and A10 then shows the out-of-sample MSE and square of the Pearson correlation coefficient for employment, profits, and sales, for both the non-winners and winners. Consider the non-winners. For both employment (Table 5) and sales (Table A10), Boosting, LASSO and the Fafchamps and Woodruff (2017) model have the lowest MSE, followed by SVR. The Raven test alone is also a powerful predictor. The business plan scores of judges are among the worse. Nevertheless, the differences are small across models, with only a 9% reduction in MSE for

employment obtained when moving from using the judge's business plan scores to the lowest MSE model (that of Boosting). SVR and the Fafchamps and Woodruff (2017) models have also the lowest MSE for profits. The squared correlations are below 11% for employment, while they are 13% or below for sales and profits.

Among the winners, we do not see large differences in performance across the different methods, especially for sales and profits. The MSE of the judges, human expert models, and machine learning models are within 7% of one another for these two outcomes for winners, and the squared correlations are always below 4%. We see more differences in performance for the employment outcome. The best performance comes from LASSO, with a squared correlation of 15%. For this outcome, the judge scores have superior performances than the other parsimonious models.

As with business survival, the Online Appendix includes several additional results to confirm that the ML algorithms are properly calibrated. Tables A6-A9 show that the selected model parameters in LASSO, SVM and Boosting are in the interior of the intervals used during cross-validation. Figures A3-A4 provide further examples of how cross-validation selects the optimal value of λ in LASSO, and that different values lead to large differences in MSE. In-sample performances are reported in Tables A12-A14.

4.5 Does machine learning use different predictors from human experts?

The machine learning models potentially select different predictors in each fold, and for each outcome. Nevertheless, it is of interest to see whether the machine learning algorithm ends up putting high weight on variables that would not typically be considered by experts when predicting business success. SVM and SVR use kernels over many variables and so are not easily interpretable in this regard.

Friedman (2001) and Schonlau (2005) note that for boosting regression, one can compute the influence of a variable. This depends on the number of times a variable is chosen across all iterations (trees) and its overall contribution to the log-likelihood function. Such values are then standardized to sum up to 100. We therefore look at the variables which have been selected at least once in the 6 status-outcome estimations (non-winners/winners, employment/sales/profits) reported in Tables 5-6 and A10. It is interesting to note that, among the 566 predictors considered, 182 have been picked by the algorithm to construct a tree. However, around 95 of them have been selected only once, 87% of them have been selected 3 times or less, and only 24 variables have been used at least 4 times.

Table A20 lists these 24 predictors along with the number of boosted regression they have been used in, and the influence for each status-outcome. The most striking result is that most of these regressors are similar to the ones selected by heuristic models. Indeed, the list includes demographic characteristics, wealth and income indicators, measures of self-confidence, optimism and happiness, number of employees at baseline, Raven test, grit and overall ability. The most frequently selected variable is the ability measure used in Fafchamps and Woodruff (2017),

followed by registration time. Given these results, it is perhaps not surprising that the machine learning algorithms do not produce substantial improvements over the other models considered: the inputs selected by the models end up being very similar, and most of the boosted regressions select only 1 or 2 splits, implying low levels of interactions among these variables. In other words, the simple models of human experts already contain many of the most powerful predictors, and there seems not to be important nonlinearities in this context.

Nevertheless, the list also includes some additional variables which may be useful to improve predictions. Taxes paid in the past and future expectations about the amount of money needed to expand are often selected. Experience abroad seems to be relevant as well. Family links and household composition can also play a role: the list includes the number of brothers and sisters. The length of the first answer is often chosen. Finally, one specific answers on the Grit scale is chosen.

The same exercises can be replicated for LASSO. Among the 566 possible predictors, only 51 variables were selected at least once by the LASSO algorithms reported in Tables 5-6 and A10, but 19 of them were selected two or three times. Table A21 lists these predictors. There are some overlaps with the variables selected in Table A20, e.g. age, gender, and some indicators of self-confidence, optimism, ability and family composition. Quite interestingly, one incorrect answer from the Raven test is selected, thus suggesting that there might be important information in not just whether individuals get the right answer, but also in which wrong answer they choose.

It is worth emphasizing that, as discussed in Mullainathan and Spiess (2017), that different algorithms and different samples may lead to different variable selections. Indeed, if some variables are highly correlated, then they be used as substitutes when predicting business success. The final set of selected variables depends on the specific finite sample used to train an algorithm. Nevertheless, it is remarkable that several variables are selected by both LASSO and Boosting. This supports the conclusion regarding their high predictive power. Moreover, the aim of this section is to identify top predictors, and to test whether this set of variables overlaps with those selected by economists. We are not making any causality claim.

We investigate two further questions regarding the comparison of machine learning to other approaches. The first is whether the judges reading of the business plans and subsequent scoring do provide valuable predictive information that can not otherwise be picked up by the machines. Table A16 shows that, except when predicting survival, business plan scores and sub-scores are not selected at all by LASSO when added to the set of possible predictors. This suggests that any information in the business plan scores is either predictable by other variables used in the model, or else of little predictive value. The second question is whether any differences in performance between the ad hoc models of economists and the machine learning variables reflect differences in the variable inputs, or the ability of machine learning methods to combine these with more flexible functional forms. Table A17 compares the performance of the Fafchamps and Woodruff (2017) model to our full SVM and Boosting models, and to SVM and Boosting models that only

start with the same input variables as Fafchamps and Woodruff, with the differences then coming from functional form flexibility these models provide. We see the performances are very similar across models, but where small differences arise (e.g. sales and employment for winners), they appear to stem more from the machine learning models using different inputs than from their use of more flexible functional forms.

5. Why can't we predict better, and how important is this economically?

Comparing all of these sets of results leads to several conclusions. First and foremost, predicting outcomes for these businesses is difficult, with the out-of-sample performance of all models quite low, especially for the profits and sales of winners. Second, machine learning does not consistently lead to improved performance compared to human predictions in this context. In fact, simple models from human experts slightly outperform at least one out of three machine methods for every outcome considered. Confidence intervals are also wide and they often overlap across different models. However, some of the machine learning predictions do yield on average more accurate predictions than using the judge scores, and are often among the best performing models. Third, a simple measure of cognitive skills, the Raven test, performs at least as well on average as using judge scores, and has often similar performances to those of human expert models, or machine learning. Fourth, the scores from the initial applications almost always provide a lower MSE than the business plan scores, raising doubts about the cost-effectiveness of this added time and cost.

These results raise questions of why the business plan scores from judges were not more powerful in predicting success, and of why the machine learning did not perform better. We examine each in turn, and then consider whether the methods do better at the very top tail.

5.1 Why didn't the human experts do better?

While there are several plausible explanations as to why the judges' scores are not more predictive of business success, our data enables us to rule out a number of these.

A first potential explanation is that perhaps the judges scores do not reflect which firms the judges thought would succeed most for one of two possible reasons. A first reason could be if individual judges differed in how important they viewed the different components of the overall score in determining business success, but could not reflect this because they had no control over the weights used. However, as Table A3 shows, most of the sub-scores also do not predict performance, and, in particular, the capstone score, which best reflects their overall assessment of whether there is something special about the firm, is also not a good predictor of performance. A second reason could be if judges had instead wanted to choose firms on the basis of which would benefit most from the grants (who had the largest treatment effects), instead of which firms would be most successful. However, Appendix 18 of McKenzie (2017) finds that, if anything, new firm applicants scored more highly by judges had lower treatment effects over the first year, and by the

time of the three-year horizon that we use here, there is no significant treatment heterogeneity according to the business plan score.

A second possible explanation for the poor performance of the judges could stem from them only observing anonymized plans, and so not being able to statistically discriminate on the basis of characteristics like gender and age that our analysis has found to have predictive power. To examine this possibility, in Table A15 we drop age and gender from the economist models and Lasso models, and find that these models still do better than the judges in predicting out-of-sample performance, even without this statistical discrimination.[10]

A third possibility is that, unlike the economist models and machine learning algorithms, the judges did not have access to the training dataset with outcomes for a subset of applicants. However, the judges did have their past experience of the Nigerian business environment and working with similar firms, and the lessons of their fellow judges discussed in judge training and codified in the scoring rubric. Moreover, even if they had observed outcomes, it may have been difficult for them to understand what factors mattered most for business success. As an example, McKenzie (2018) shows that even after having gone through the YouWin program, firm owners were unable to tell how much winning the program had mattered to their firm outcomes.

Fourth, perhaps the problem is that it is much easier for judges to distinguish a bad business plan from a good one, than it is for them to distinguish among top business plans. Our data only comes from those who go through to the stage of submitting a business plan, who are the top-25 percent of applicants. But even among this group, the judges determined some of the business plans to be particularly poor, giving scores as low as 3 out of 100. If judges found it easier to distinguish the bottom firms from the rest, we should expect to see non-linearities in the relationship between score and outcomes, with at least a strong positive relationship among lower-scored proposals. Figure 3 shows this is not the case.

5.2 Why didn't machine learning do better?

Likewise, there are a number of potential explanations as to why the machine learning did not have better predictive performance, and we can use our data to explore the plausibility of each of these. These concerns relate to the sample size, inputs into the models, and implementation of the machine learning methods.

A first potential explanation could be that the machine learning needs a larger sample size to perform well. Although we cannot see whether increasing the sample would improve performance, we can examine how sensitive our results are to decreasing the training sample size used. We start with our full 80 percent training sample (recall this is just over 800 observations), and then randomly remove a fraction of this training sample to see how stable the out-of-sample accuracy rate is. The results are shown in Figure A5. Prediction accuracy for non-winners is substantially

[10] All models continue to perform poorly when predicting business success among winning firms, especially for profits and sales. It would also be an interesting exercise in future research to test whether judges would change how they scored proposals if age and gender information were added to the plans, as it is unclear whether they would statistically discriminate.

worse when using only 30 percent or less of the training sample (samples of under 250 observations), continues to improve as we add more data, and is then reasonably stable when using only 80 or 90 percent of our training sample (samples of 600 or 700 instead of 800): the out-of-sample MSE is 1.55 when using 80 percent of the training sample, 1.54 when using 90 percent, and it remains 1.54 when using the full training sample to calibrate our LASSO algorithm. Similar conclusions can be reached when repeating the same exercise for winners (Figure A6): if anything, the out-of-sample MSE increases when moving from using 90 percent of the training sample to 100 percent (from 0.837 to 0.847). This suggests that small increases in sample size would not improve substantially the model performance, but of course we cannot rule out that performance would be better if data on millions of firms were available. Nevertheless, we note that our sample of over 2,000 firms is substantially larger than is possible for most business plan competitions, or than the information sets of most venture capitalists looking to make investments.

A second set of concerns are based on the inputs into the models. A first issue might be that the inputs into the model may lack predictive power, if respondents strategically mis-report items on their application in order to enhance their perceived chances of winning the competition. To the extent that this is occurring, we view it as a feature, not a bug, since we are using data collected under competition conditions to predict which competition applicants will succeed.[11] Nevertheless, we do not have strong reasons to expect strategic reporting to greatly affect our machine-learning results, in part because it is unclear for many of our variables in which direction a strategic applicant should report (e.g. is it better to report having overseas work experience or not? is it better to say you are risk-seeking or not? Is it better to say your household has a freezer or not?).

Another issue concerning the inputs is the possibility that the large number of additional input variables we consider for machine learning are highly correlated with the simple set of variables used in our ad hoc economist models, and so do not contain useful independent information. To investigate this possibility, we regress each machine learning input on the McKenzie (2015) or Fafchamps and Woodruff (2017) variable sets, using an OLS for non-binary variables and a Probit for binary variables. Figure A1 shows kernel densities of the resulting R^2 (or pseudo-R^2 for binary variables), for winners and non-winners separately. While there is a mass close to $R^2=1$ (there is some overlapping between the two sets), most of the R^2 are below 0.2, showing that there is indeed substantial independent information beyond the set of variables used in the ad hoc models.

Nevertheless, the two sets of variables are strongly related. We have conducted a canonical analysis to further investigate the relationship between the variables used in the economist models and the ones available by the ML algorithms. As expected, since the ML set of inputs includes sub-scores used to construct indicators included in the economist models (e.g. ability), the first canonical correlations are very high (Table A5). A substantial fraction of the higher-order canonical correlations is instead statistically insignificant and less correlated. Therefore, one can

[11] In contrast, if we used data from a standalone firm panel survey to predict firm growth, there would be a concern that any relationships found in the data might disappear if the data were collected under competition conditions.

conclude that the two sets are closely connected, but that the set of variables used in the ML algorithms might have included additional powerful predictors.

A third, and final set of explanations for why the machine learning did not perform better are related to the machine learning implementation. Section 4.4 already noted that the tuning parameters were chosen by cross-validation and are in the interiors of the ranges considered, so that poor fit is unlikely to be due to tuning. There are a wide variety of different machine learning methods, which raises the possibility of whether using other methods would perform better. Different machine learning algorithms may capture different features of the data. For instance, SVR allows flexible functional forms, while Boosting can include multiple high-order interactions among variables. Furthermore, human experts and algorithms may provide complementary predictions. Therefore, combining these methods may potentially lead to superior performances. Nevertheless, ensemble models merging predictions from LASSO, SVR and Boosting does not lead to substantial improvements (Table A18). Similar results are also obtained when constructing ensembles merging predictions from the machine learning algorithms with the Raven univariate regression, the business plan scores, or the Fafchamps and Woodruff (2017) model. In addition to this, Table A19 shows that using an alternative commonly-used machine learning algorithm, Elastic Net, leads to out-of-sample performances similar to those in Tables 4-6 and A10.

5.3 Predicting High Performance

Figure 1 showed substantial variation in outcomes for both winners and non-winners, with the performance of the very best firms much better than those of the typical firm. For example, the average number of workers among the non-winning firms in the top 10% of employment is 16 (median 13), while the average in the bottom 90% is 2 (median 1, i.e. no employees) and the average profits in the top decile is almost three times that in the rest of the firms. Even if the human experts and machine predictions have difficulty distinguishing between firms across the whole distribution, the question remains as to whether they fare better when trying to identify in advance these very top performers.

To investigate this, we use the predictions obtained in the previous sections and reported in Tables 5 and 6. For each model, we sort firms based on the predicted third-year employment or profit levels. We generated a dummy variable equal to one if the firm is in the top 10%, zero otherwise. Using the actual employment and profit levels, we also generate for each outcome a dummy variable equal to one if the firm is indeed in the top 10%, zero otherwise. The accuracy rate is then obtained by computing the proportion of firms in the hold-out sample correctly predicted to be in the top decile or the bottom 90%.[12] Since we are particularly interested in being able to detect

[12] We have followed this approach of setting a fixed number of businesses to predict as successful instead of the alternative of predicting one (success) if the predicted probability is above 0.5 (or the mean of the outcome variable, 0.1 in our case) because we deem this more realistic and similar to how a venture capitalist would act when having to decide on which firms to invest.

which firms will excel, we also compute the recall rate – i.e. the proportion of the top tail of firms that were correctly predicted.[13]

Panels A and B of Table 7 reports the results for the top tail in terms of employment and profits respectively. We see that the using the highest business plan scores given by judges only correctly identifies 13 to 16 percent of the firms that are in this top tail. The Raven test, ad hoc models, and machine learning models show slightly better performance, typically identifying around 20 percent of the firms that end up in the top tail, with this reaching as high as one-third for the LASSO in identifying tail employment, and 25 percent for SVM and the Raven score in identifying the top tail of profits. However, note that the sample size for these calculations is relatively small – since we are trying to predict which 21 firms will be best in a hold-out sample of around 210 firms in each case. The result is that accuracy confidence intervals for all the different methods overlap.

From an investor viewpoint, it not only matters whether or not they can identify the exact firms that end up in the top 10%, but how bad their mistakes are if they choose incorrectly. In panel C of Table 7, we therefore consider the case of an investor who is picking the top 10% of firms. We assume this investor makes an investment in each of these firms, and then would get a percentage of profits, so that their goal is to maximize total profits of the firms they choose. For each model, we then aggregate up the profits of the firms chosen in panel B, to report the total monthly profits that the investor would achieve if choosing firms randomly, or according to each selection method. For the non-winners, using the business plans scores from judges would yield returns that are twice as high as random choice, and using any of the Raven test score, Fafchamps and Woodruff model, and machine learning models would yield three to four times the profits. For the sample of competition winners, using the business plan scores would actually yield a portfolio earning lower profits than choosing firms by pure random chance, using the economist models or machine learning models would give approximately twice the level of profits, and the Raven test score yields four times the profits. These results provide suggestive evidence that an investor could use these modeling approaches to gain higher returns, but the bootstrapped confidence intervals are wide and overlap each other.

6. Conclusions and Discussion

Our results from a large business plan competition show that successful entrepreneurial performance is very difficult to predict. Expert judges who know the local context, models from several economists with vast experience studying entrepreneurship in developing countries, and multiple machine learning algorithms all struggle to predict which entrants to the competition will succeed most from among those who reached the stage of submitting business plans. We find judges' scores are uncorrelated with firm outcomes three years later. This does not mean business

[13] An advantage of our machine learning approach is the flexibility to consider different objective functions. For instance, since each start-up may have different requests in terms of initial capital necessary to found or expand the business, some investors may actually be interested in maximizing the recall rate subject to a fixed amount of money invested in the businesses predicted to be successful. The calibration procedure can be easily adapted to take this objective into account (see also Sansone, 2018).

performance is completely unpredictable, as several basic demographic characteristics of the entrepreneurs (age and gender) and measures of their ability (Raven test score) do increase prediction accuracy: males in their 30s who score highly on an ability test are more likely to succeed. However, the overall predictive power of simple models of economists which use these variables is still low, and machine learning algorithms that consider many more potential inputs do not typically outperform these simple models.[14] Even when machine learning algorithms do perform better than the other models considered, the improvements are not substantial and the confidence intervals are large and overlapping.

These results suggest several implications for economists studying entrepreneurship, and for governments and investors seeking to identify high-return entrepreneurs. First, although entrepreneurship is difficult to predict, economists do appear to be capturing the most relevant characteristics for predicting performance, and adding local expert judgement or machine learning does not offer systematic performance gains. Second, for policymakers running competitions, expensive and time-consuming scoring by judges may be able to be replaced by less costly simple predictive models. Third, despite large heterogeneity in outcomes, the fundamental riskiness inherent in entrepreneurship will make it difficult for investors to determine who will grow most, potentially restricting the flow of capital to high-return entrepreneurs.

Several caveats are worth noting. First, our analysis is all conditional on individuals applying for a business plan competition and getting through the first round to the point of submitting a business plan. This process likely selects out many individuals with very low growth prospects, and it may be easier to predict who definitely will not succeed than to predict who might. For example, de Mel et al., (2010) find that measures of background, ability, and attitudes can help distinguish subsistence self-employed from those likely to hire workers. Nevertheless, the problem facing investors and judges of business plan competitions is to decide among those who have self-selected into participating and passed through initial screening, and we show that this is hard to do, despite substantial variation in business performance among the sample.

Second, the poor performance of machine learning in our context may even be an upper bound on how well one could expect such an approach to work in other business plan competitions. We have data from the world's largest business plan competition, and use a training set of outcomes for these participants to predict out of sample fit in a test sample that comes from the same country, population, and time period. We might expect fit to be even worse if models were trained on one competition and then applied to predicting success in another. Policy makers and investors would typically have to make investment decisions based on the past performance of a more limited number of firms. However, our ad hoc economist models of McKenzie (2015) and Fafchamps and Woodruff (2017) did draw upon what the authors had learned from reviewing data from several

[14] This lack of improvement from machine learning is in line with results in several other domains. For example, Beattie et al. (2016) found a simple average of seven non-academic variables performed as well as sophisticated machine learning techniques when analyzing the importance of personality traits in college student performance; while a simple autoregressive model has similar performances to Google's Flu trends algorithm (Goel et al., 2010).

sources and studying findings from similar contexts. The machine learning algorithms did not have access to these previous datasets, thus limiting by construction their ability to learn, so it is possible that combining data from multiple business plan competitions in different countries could further improve performance.

Third, although we have used three of the most popular machine learning algorithms, it is possible that more advanced algorithms or extensive grid-searchers could further improve performance. However, such algorithms may still be computationally infeasible, even for big companies (Johnston, 2012), or extremely difficult to code, which is the reason behind the extremely high prizes – often reaching $1 million (Netflix, 2009) - offered in machine learning competitions. Often these algorithms have been designed to be applied to large data sets with millions of observations, and the sample sizes of firm samples may limit the gains to be had from such an approach.

Finally, our focus has been on predicting which firms will have the best outcomes, which is an object of most interest to investors and lenders. In contrast, such predictions are usually inappropriate for a policy maker aiming to optimally allocate resources (Athey, 2017), who would instead be most interested in allocating funds to firms that would benefit most from the grants, not to those firms which would be successful anyway.[15] But even in such cases, identifying a pool of firms that are likely to survive and grow may be a useful first step in screening applications, and policy makers may also care about offering support to the next flagship company.

[15] See Hussam et al. (2016) for a recent application of using machine learning to predict which small entrepreneurs will have the highest returns to grants.

References

Astebro, T., Elhedhli, S., 2006. The effectiveness of simple decision heuristics: Forecasting commercial success for early-stage ventures. Manage. Sci. 52, 395–409.

Athey, S., 2017. Beyond prediction: Using big data for policy problems. Sci. Mag. 355, 483–485.

Aulck, L., Velagapudi, N., Blumenstock, J., West, J., 2016. Predicting Student Dropout in Higher Education. ArXivWorking Pap., arXiv 1606.06364.

Banerjee, A. V., Duflo, E., 2014. (Dis)Organization and success in an economics MOOC. Am. Econ. Rev. 104, 514–518.

Bauer, E., Kohavi, R., Chan, P., Stolfo, S., Wolpert, D., 1999. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Mach. Learn. 36, 105–139.

Beattie, G., Laliberte, J.-W.P., Oreopoulos, P., 2018. Thrivers and Divers: Using Non-Academic Measures To Predict College Success and Failure. Econ. Educ. Rev.

Belenzon, S., Chatterji, A.K., Daley, B., 2017. Eponymous entrepreneurs. Am. Econ. Rev. 107, 1638–1655.

Blanchflower, D.G., Levine, P.B., Zimmerman, D.J., 2003. Discrimination in the Small-Business Credit Market. Rev. Econ. Stat. 85, 930–943.

Boden, R.J., Nucci, A.R., 2000. On the survival prospects of men's and women's new business ventures. J. Bus. Ventur. 15, 347–362.

Bruhn, M., 2009. Female-Owned Firms in Latin America: Characteristics, Performance, and Obstacles to Growth. World Bank Policy Res. Work. Pap. 5122.

Celiku, B., Kraay, A., 2017. Predicting Conflict. World Bank Policy Res. Work. Pap. 8075.

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., Mullainathan, S., 2016. Productivity and Selection of Human Capital with Machine Learning. Am. Econ. Rev. Pap. Proc. 106, 124–127.

de Mel, S., McKenzie, D., Woodruff, C., 2010. Who are the Microenterprise Owners?: Evidence from Sri Lanka on Tokman v. de Soto, in: Lerner, J., Schoar, A. (Eds.), International Differences in Entrepreneurship. University of Chicago Press, pp. 63–87.

Duckworth, A.L., Peterson, C., Matthews, M.D., Kelly, D.R., 2007. Grit: Perseverance and passion for long-term goals. J. Pers. Soc. Psychol. 92, 1087–1101.

Fafchamps, M., Woodruff, C., 2017. Identifying Gazelles: Expert Panels vs. Surveys as a Means to Identify Firms with Rapid Growth Potential. World Bank Econ. Rev. 31, 670–686.

Friedman, J., 2001. Greedy Function Approximation: A Gradient Boosting Machine. Ann. Stat. 29, 1189–1232.

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive Logistic Regression: A Statistical View of Boosting. Ann. Stat. 28, 337–407.

Goel, S., Hofman, J.M., Lahaie, S., Pennock, D.M., Watts, D.J., 2010. Predicting consumer behavior with Web search. Proc. Natl. Acad. Sci. 107, 17486–17490.

Guenther, N., Schonlau, M., 2016. Support Vector Machines. Stata J. 16, 917–937.

Guzman, J., Stern, S., 2017. Nowcasting and placecasting entrepreneurial quality and performance, in: Haltiwanger, J., Hurst, E., Miranda, J., Schoar, A. (Eds.), Measuring Entrepreneurial Businesses: Current Knowledge and Challenges. University of Chicago Press, pp. 63–109.

Hall, R.E., Woodward, S.E., 2010. The burden of the nondiversifiable risk of entrepreneurship. Am. Econ. Rev. 100, 1163–1194.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed, Springer Series in Statistics. Springer.

Hussam, R., Rigol, N., Roth, B., 2016. Targeting High Ability Entrepreneurs Using Community

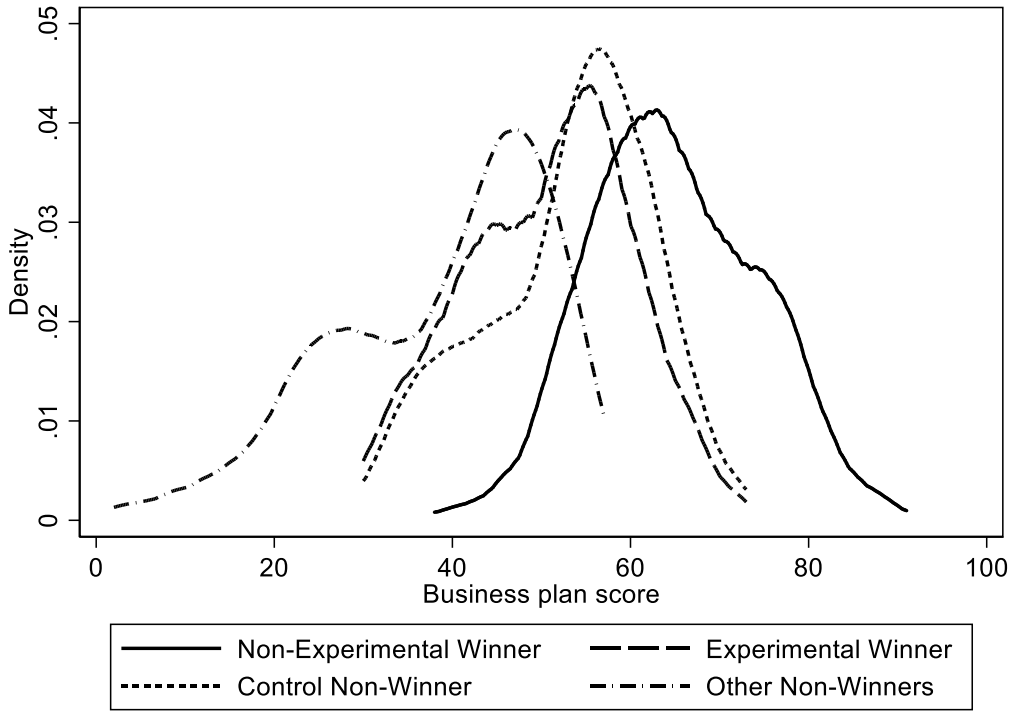
Information: Mechanism Design In The Field. Work. Pap.
Hvide, H.K., Panos, G.A., 2014. Risk tolerance and entrepreneurship. J. financ. econ. 111, 200–223.
Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., Ermon, S., 2016. Combining satellite imagery and machine learning to predict poverty. Sci. Mag. 353, 790–4.
Johnston, C., 2012. Netflix Never Used Its $1 Million Algorithm Due To Engineering Costs. Wired.
Kerr, W.R., Lerner, J., Schoar, A., 2014. The consequences of entrepreneurial finance: Evidence from angel financings. Rev. Financ. Stud. 27, 20–55.
Kleinberg, J., Lakkaraj, H., Leskovec, J., Ludwig, J., Mullainathan, S., 2017. Human Decisions and Machine Predictions. Q. J. Econ. 133, 237–293.
Kleinberg, J., Ludwig, J., Mullainathan, S., Obermeyer, Z., 2015. Prediction Policy Problems. Am. Econ. Rev. Pap. Proc. 105, 491–495.
Kopf, D., 2015. What happened when Nigeria created the World's Largest Business Plan Competition [WWW Document]. Priceonomics.
Luca, M., Kleinberg, J., Mullainathan, S., 2016. Algorithms Need Managers, Too. Harv. Bus. Rev. 104, 96–101.
Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., de Mendonça, A., 2011. Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Res. Notes 4, 299.
McKenzie, D., 2018. Can business owners form accurate counterfactuals? Eliciting treatment and control beliefs about their outcomes in the alternative treatment status. J. Bus. Econ. Stat. 36, 714–722.
McKenzie, D., 2017. Identifying and Spurring High-Growth Entrepreneurship: Experimental Evidence from a Business Plan Competition. Am. Econ. Rev. 107, 2278–2307.
McKenzie, D., 2015. Identifying and Spurring High-Growth Entrepreneurship: Experimental Evidence from a Business Plan Competition. World Bank Policy Res. Work. Pap. 7391.
McKenzie, D., Paffhausen, A.L., 2018. Small firm death in developing countries. Rev. Econ. Stat.
Michelacci, C., Schivardi, F., 2016. Are They All Like Bill, Mark, and Steve? The Education Premium for Entrepreneurs. EIEF Work. Pap. 16.
Mullainathan, S., Spiess, J., 2017. Machine Learning: An Applied Econometric Approach. J. Econ. Perspect. 31, 87–106.
Nanda, R., 2016. Financing high-potential entrepreneurship. IZA World Labor 1–10.
Netflix, 2009. Netflix Prize [WWW Document]. URL http://www.netflixprize.com/ (accessed 1.1.17).
Ng, A., 2016. Machine Learning [WWW Document]. Coursera. URL https://www.coursera.org/learn/machine-learning (accessed 1.1.16).
Nikolova, E., Ricka, F., Simroth, D., 2011. Entrepreneurship in the transition region: an analysis based on the Life in Transition Syrvey, in: Crisis and Transition: The People's Perspective. EBRD, pp. 76–96.
OECD, 2010. High-Growth Enterprises : What Governments Can Do To Make A Difference. OECD Stud. SMEs Entrep.
Olafsen, E., Cook, P.A., 2016. Growth Entrepreneurship in Developing Countries: A Preliminary Literature Review. The World Bank Group, Washington, D.C.
Queiro, F., 2016. The Effect of Manager Education on Firm Growth. Q. J. Econ. 118, 1169–1208.

Reddy, S., 2017. A computer was asked to predict which start-ups would be successful. The results were astonishing. World Econ. Forum.
Ricadela, A., 2009. Fifty Best Tech Startups. Bloom. Businessweek.
Robb, A.M., Watson, J., 2012. Gender differences in firm performance: Evidence from new ventures in the United States. J. Bus. Ventur. 27, 544–558.
Sansone, D., 2018. Beyond Early Warning Indicators: High School Dropout and Machine Learning. Oxf. Bull. Econ. Stat.
Schonlau, M., 2005. Boosted regression (boosting): An introductory tutorial and a Stata plugin. Stata J. 5, 330–354.
Scott, E., Shu, P., Lubynsky, R., 2015. Are "better" ideas more likely to succeed? An empirical analysis of startup evaluation. Harvard Bus. Sch. Work. Pap. 16.
Skriabikova, O.J., Dohmen, T., Kriechel, B., 2014. New evidence on the relationship between risk attitudes and self-employment. Labour Econ. 30, 176–184.
Subrahmanian, V.S., Kumar, S., 2017. Predicting human behavior: The next frontiers. Sci. Mag. 355, 489.
van Praag, C.M., Cramer, J.S., 2001. The roots of entrepreneurship and labour demand: Individual ability and low risk aversion. Economica 68, 45–62.
Varian, H.R., 2014. Big Data: New Tricks for Econometrics. J. Econ. Perspect. 28, 3–28.
Verplancke, T., Van Looy, S., Benoit, D., Vansteelandt, S., Depuydt, P., De Turck, F., Decruyenaere, J., 2008. Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies. BMC Med. Inform. Decis. Mak. 8, 56.
Zacharakis, A.L., Shepherd, D.A., 2001. The nature of information and overconfidence on venture capitalists' decision making. J. Bus. Ventur. 16, 311–332.
Zygmunt, Z., 2016. What is better: gradient-boosted trees, or a random forest? [WWW Document]. FastML. URL http://fastml.com/what-is-better-gradient-boosted-trees-or-random-forest/ (accessed 1.1.17).

Figure 1: Substantial Dispersion in Firm Outcomes Three Years After Applying

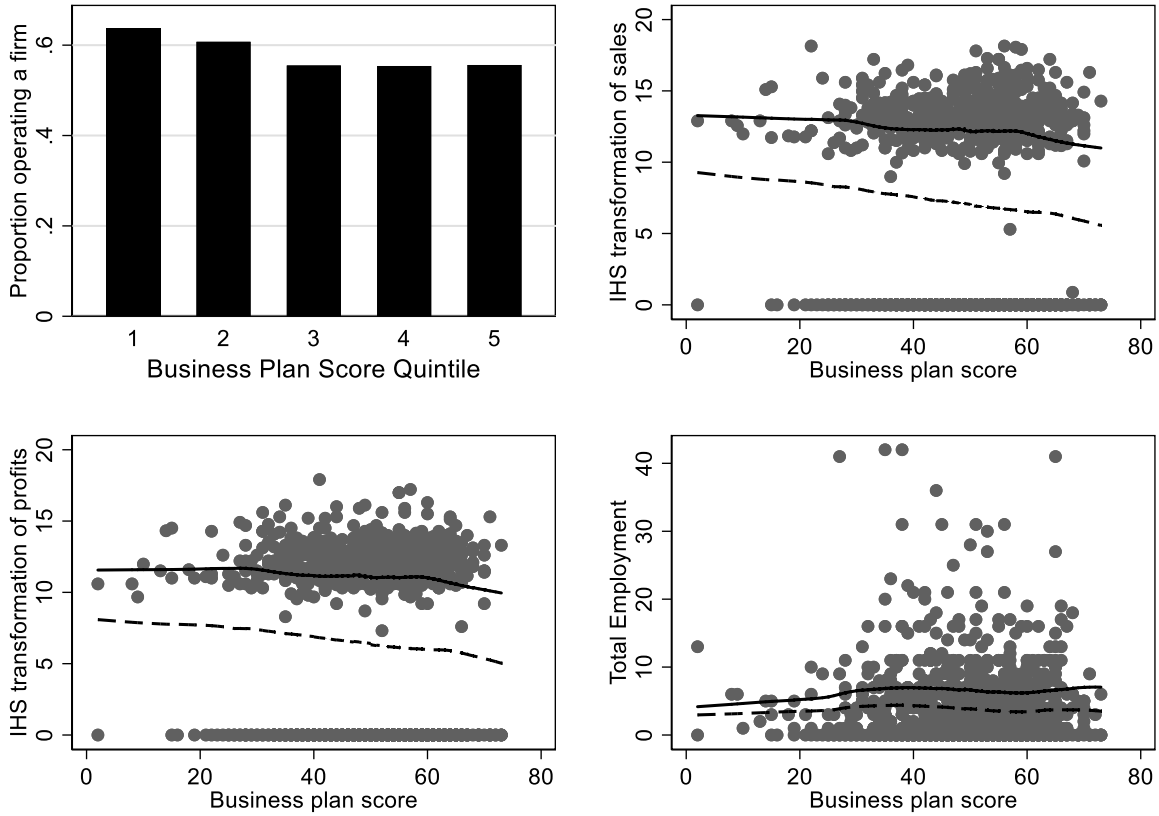
Notes: Figure shows kernel densities of firm outcomes for firms surveyed in third follow-up, and are conditional on the firm operating at that time. Winners include both non-experimental and experimental winners, and non-winners include control group semi-finalists as well as a sample of firms submitting business plans which had first round application scores just above the cut-off for getting selected for business plan training. Total employment truncated at 60 (9 outliers) for readability. Profits and Sales are measured in monthly Naira, and have been transformed using the inverse hyperbolic sine transformation.

Figure 2: Distribution of Business Plan Scores by Competition Status



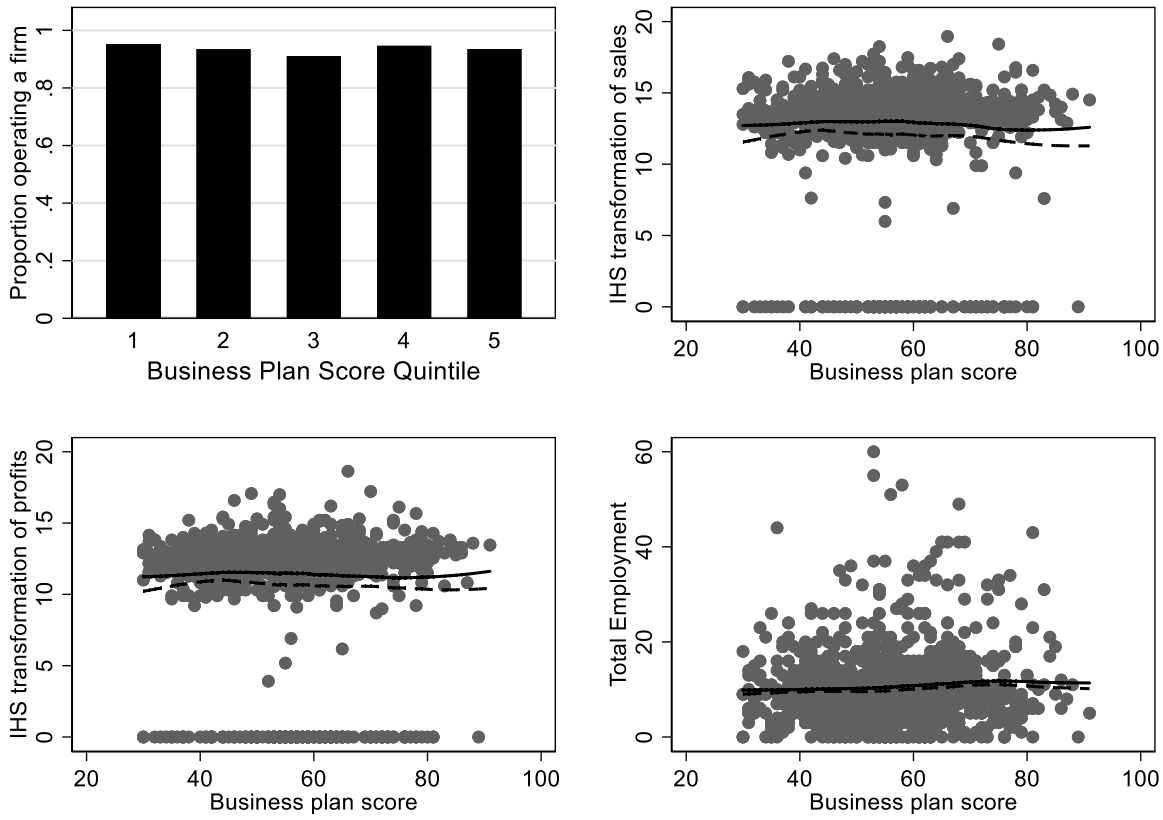
Notes: kernel densities of business plan scores shown. Non-experimental winners were chosen on the basis of having the top scores nationwide, or as being the most promising within their geographic region, with fixed quotas by whether the application was for a new or existing business. Experimental winners and control non-winners were chosen randomly from within the remaining semi-finalists, with this randomization done within strata defined by region and existing or new business status. Other non-winners are the scores for our remaining sample of non-winners, selected on the basis of only just making the first round cut to receive business plan training.

Figure 3: The Relationship between Business Plan Scores of Judges and Business Outcomes Three Years after applying for Non-Winners in the Business Plan Competition



Notes: Top left panel shows business operation three years after applying by quintile of business plan score. Remaining panels show scatterplots of the sales, profits, and employment three years after applying against the business plan score. The dashed line plotted on each figure is a lowess line of the unconditional relationship, which codes outcomes as zero for applicants who are not operating firms. The solid line on each figure is a lowess line of the relationship conditional on the business being in operation. Employment truncated at 60 workers for readability.

Figure 4: The Relationship between Business Plan Scores of Judges and Business Outcomes Three Years after applying for Winners in the Business Plan Competition



Notes: Top left panel shows business operation three years after applying by quintile of business plan score. Remaining panels show scatterplots of the sales, profits, and employment three years after applying against the business plan score. The dashed line plotted on each figure is a lowess line of the unconditional relationship, which codes outcomes as zero for applicants who are not operating firms. The solid line on each figure is a lowess line of the relationship conditional on the business being in operation. Employment truncated at 60 workers for readability, profits truncated from below at zero.

Table 1: Can Human Experts Predict Business Survival After 3 Years?

	Non-Winners			Winners			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Judges	McKenzie	FaWo	Judges	Judges	McKenzie	FaWo
Business Plan score	-0.0002	-0.0009	-0.0009	-0.0000	-0.0003	-0.0006	-0.0006
	(0.0014)	(0.0015)	(0.0015)	(0.0006)	(0.0006)	(0.0008)	(0.0008)
Female		-0.1505***	-0.1272***			-0.0018	0.0064
		(0.0385)	(0.0384)			(0.0212)	(0.0218)
Age		0.0110***	0.0085***			0.0036*	0.0040**
		(0.0032)	(0.0032)			(0.0021)	(0.0020)
Graduate Education		-0.0211				0.0262	
		(0.0632)				(0.0362)	
Worked abroad		0.0233				-0.0447*	
		(0.0589)				(0.0243)	
Lottery choice		-0.0194				-0.0171	
		(0.0292)				(0.0160)	
Wealth		-0.0028				0.0030	
		(0.0057)				(0.0027)	
Grit		-0.0369				-0.0031	
		(0.0285)				(0.0154)	
Agriculture		-0.0490				-0.0156	
		(0.0352)				(0.0196)	
IT		-0.0546				-0.0264	
		(0.0509)				(0.0244)	
Ability		0.0509***	0.0333**			-0.0086	-0.0101
		(0.0169)	(0.0162)			(0.0087)	(0.0081)
Manufacture		0.0018	0.0296			-0.0259	-0.0182
		(0.0395)	(0.0355)			(0.0203)	(0.0183)
Retail			0.0673				0.0210
			(0.0793)				(0.0446)
Self-confidence			0.0083				0.0062
			(0.0091)				(0.0050)
Motivation			0.0193*				-0.0054
			(0.0107)				(0.0048)
Happiness			-0.0031				-0.0090
			(0.0120)				(0.0072)
Optimism			0.0332**				0.0072
			(0.0136)				(0.0082)
Credit			0.0172				-0.0144**
			(0.0174)				(0.0073)
Existing	0.2125***	0.1688***	0.0677	0.0353**	0.0347**	0.0168	0.0063
	(0.0387)	(0.0411)	(0.0469)	(0.0157)	(0.0158)	(0.0194)	(0.0210)
Log(Total award paid)					0.0248*	0.0219	0.0183
					(0.0145)	(0.0140)	(0.0154)
Regional Indicators	No	Yes	Yes	No	No	Yes	Yes
Sample Size	1,077	1,077	1,077	1,051	1,051	1,047	1,048
P-Value		0.000	0.000			0.530	0.105
Pseudo-R^2	0.024	0.076	0.098	0.011	0.020	0.054	0.064
Accuracy	58.4%	65.6%	66.9%	93.6%	93.6%	93.6%	93.6%

Robust Standard Errors in parentheses. *, **, *** denote significance at the 10, 5, and 1 percent levels respectively. See Appendix A.1 for variable definitions. Dependent variable is having a business in operation 3 years after applying. Marginal effects from logit shown. P-value is for testing that set of controls added by McKenzie (2015) or Fafchamps and Woodruff (2017) (abbreviated as FaWo) are jointly zero.

Table 2: Can Human Experts Predict Employment After 3 Years?

	Non-Winners			Winners			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Judges	McKenzie	FaWo	Judges	Judges	McKenzie	FaWo
Business Plan score	-0.0023	0.0008	0.0046	0.0881***	0.0494	0.0819**	0.0703**
	(0.0168)	(0.0182)	(0.0180)	(0.0312)	(0.0308)	(0.0358)	(0.0336)
Female		-1.0647**	-0.7637			-0.6078	-0.0574
		(0.4792)	(0.4865)			(0.8249)	(0.7792)
Age		0.1069***	0.0775*			0.1705**	0.1541**
		(0.0413)	(0.0398)			(0.0823)	(0.0775)
Graduate Education		0.9593				-2.4624***	
		(0.9798)				(0.9219)	
Worked abroad		0.1514				-1.0069	
		(0.8827)				(1.0370)	
Lottery choice		0.5630				-0.4718	
		(0.3516)				(0.6622)	
Wealth		0.1012				0.1315	
		(0.0685)				(0.1338)	
Grit		-0.4658				-0.4396	
		(0.3613)				(0.6060)	
Agriculture		-0.5865				-1.0522	
		(0.4563)				(0.7892)	
IT		-1.3288***				-2.7039***	
		(0.4937)				(0.8412)	
Ability		0.4502**	0.2993			0.0807	-0.0850
		(0.2175)	(0.1836)			(0.3156)	(0.3168)
Manufacture		-0.4825	-0.1086			-0.3916	0.0146
		(0.4566)	(0.4158)			(0.9174)	(0.8037)
Retail			0.6546				-0.6681
			(0.9276)				(0.8879)
Self-confidence			0.0726				0.0990
			(0.1164)				(0.1854)
Motivation			0.0815				-0.2382
			(0.1140)				(0.2048)
Happiness			0.1961				0.4688*
			(0.1411)				(0.2705)
Optimism			0.1695				-0.1639
			(0.1647)				(0.2880)
Credit			-0.0378				1.0346*
			(0.1571)				(0.5763)
Existing	1.6260***	1.4832***	0.3739	2.5040***	2.3173***	2.9320***	1.7554**
	(0.4510)	(0.4699)	(0.5704)	(0.6355)	(0.6226)	(0.7556)	(0.8518)
Log(Total award paid)					3.8655***	3.8208***	3.8180***
					(0.4328)	(0.4444)	(0.4416)
Regional Indicators	No	Yes	Yes	No	No	Yes	Yes
Sample Size	1,077	1,077	1,077	1,051	1,051	1,051	1,051
P-value		0.000	0.000			0.000	0.013
Adjusted R^2	0.012	0.038	0.049	0.020	0.053	0.072	0.079
MSE	34.06	33.18	32.79	103.33	99.77	97.76	97.01

Robust Standard Errors in parentheses. *, **, *** denote significance at the 10, 5, and 1 percent levels respectively. See Appendix A.1 for variable definitions. Dependent variable is number of employees 3 years after applying, coded as zero for applicants without an operating firm. P-value is for testing that set of controls added by McKenzie (2015) or Fafchamps and Woodruff (2017) (abbreviated as FaWo) are jointly zero.

Table 3: Can Human Experts Predict Profits After 3 Years?

	Non-Winners			Winners			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Judges	McKenzie	Fafchamps	Judges	Judges	McKenzie	Fafchamps
Business Plan score	-0.0045	-0.0061	-0.0066	-0.0030	-0.0075	-0.0093	-0.0064
	(0.0180)	(0.0188)	(0.0183)	(0.0126)	(0.0130)	(0.0143)	(0.0139)
Female		-1.8901***	-1.5593***			-1.0142**	-0.8647*
		(0.4955)	(0.4890)			(0.4454)	(0.4465)
Age		0.1253***	0.0830**			0.0460	0.0363
		(0.0417)	(0.0410)			(0.0358)	(0.0351)
Graduate Education		-0.9393				0.0187	
		(0.8041)				(0.6114)	
Worked abroad		0.7856				-1.3221**	
		(0.7839)				(0.6147)	
Lottery choice		-0.0687				-0.2843	
		(0.3756)				(0.2985)	
Wealth		-0.0222				0.1071**	
		(0.0742)				(0.0461)	
Grit		-0.4456				-0.1065	
		(0.3614)				(0.2752)	
Agriculture		-0.4053				-0.2147	
		(0.4597)				(0.3716)	
IT		-0.8472				-0.4611	
		(0.6560)				(0.4675)	
Ability		0.6695***	0.4487**			-0.1976	-0.2346
		(0.2150)	(0.2014)			(0.1712)	(0.1570)
Manufacture		0.2671	0.5600			0.2026	0.4204
		(0.4931)	(0.4426)			(0.3869)	(0.3590)
Retail			0.9146				1.5207***
			(0.9333)				(0.5244)
Self-confidence			0.0781				0.0782
			(0.1148)				(0.0921)
Motivation			0.0647				-0.1942*
			(0.1391)				(0.1035)
Happiness			0.1016				-0.0817
			(0.1569)				(0.1253)
Optimism			0.4209**				-0.0094
			(0.1759)				(0.1376)
Credit			0.2742*				-0.0890
			(0.1466)				(0.1985)
Existing	2.7393***	2.4020***	0.8730	1.0225***	1.0015***	0.7810**	0.5783
	(0.4736)	(0.5072)	(0.5724)	(0.2961)	(0.2983)	(0.3548)	(0.4190)
Log(Total award paid)					0.4434	0.3996	0.4044
					(0.2968)	(0.3009)	(0.3025)
Regional Indicators	No	Yes	Yes	No	No	Yes	Yes
Sample Size	1,058	1,058	1,058	1,036	1,036	1,036	1,036
P-value		0.000	0.000			0.000	0.000
Adjusted R2	0.034	0.081	0.122	0.010	0.011	0.022	0.021
MSE	37.50	35.67	34.08	21.56	21.55	21.31	21.32

Robust Standard Errors in parentheses. *, **, *** denote significance at the 10, 5, and 1 percent levels respectively. See Appendix A.1 for variable definitions. Dependent variable is inverse hyperbolic sine of profits 3 years after applying, coded as zero for applicants without an operating firm. P-value is for testing that set of controls added by McKenzie (2015) or Fafchamps and Woodruff (2017) (abbreviated as FaWo) are jointly zero.

Table 4: Prediction (Out-of-sample) Accuracy for Business Survival among Non-winners

		Model	Predictors	Accuracy	
				Mean	C.I.
	1	Logit	Constant	58.8%	[52.3%, 65.3%]
Judges	2	Logit	Application Score	58.8%	[52.3%, 65.3%]
	3	Logit	Business Score	58.8%	[52.3%, 65.3%]
	4	Logit	SubScore	56.9%	[50.5%, 63.4%]
Single Predictor	5	Logit	Gender	58.8%	[52.3%, 65.3%]
	6	Logit	Age	63.0%	[56.6%, 69.4%]
	7	Logit	Necessity Firm	55.1%	[48.3%, 61.9%]
	8	Logit	Grit	61.6%	[55.2%, 67.9%]
	9	Logit	Digit-span Recall	60.6%	[54.3%, 67.0%]
	10	Logit	Raven Test	67.1%	[61.0%, 73.2%]
	11	Logit	Registration time	58.8%	[52.3%, 65.3%]
Economist Models	12	Logit	McKenzie	63.9%	[57.4%, 70.4%]
	13	Logit	FaWo	67.1%	[60.7%, 73.6%]
	14	Logit	FaWo + BusScores	66.2%	[59.8%, 72.6%]
	15	OLS	FaWo + BusScores	65.7%	[59.4%, 72.1%]
	16	Probit	FaWo + BusScores	65.7%	[59.4%, 72.1%]
Machine Learning	17	LASSO	All	60.2%	[53.7%, 66.7%]
	18	SVM	All	66.2%	[59.8%, 72.6%]
	19	Boosting	All	63.9%	[57.6%, 70.2%]

Notes: outcome is whether an applicant who did not win the business plan competition is found to be operating a firm three years later. Out-of-sample accuracy is the ratio of true positives and true negatives to all observations, and is computed using the 20% hold-out sample (N=216). The numbers in brackets are bootstrapped 95 percent confidence intervals for hold-out prediction performance. Judges use scores from competition judges to predict the outcome. Single predictor models use a single survey measure. The economist models are the models of McKenzie (2015) and Fafchamps and Woodruff (2017) (abbreviated as FaWo) used in Tables 1-3. All models apart from the constant model also include a dummy for whether the applicant was for an existing firm compared to a new firm.

Table 5: Prediction (Out-of-sample) Accuracy for Total Employment

				Non-winners			Winners		
				MSE		R^2	MSE		R^2
		Model	Predictors	Mean	C.I.		Mean	C.I.	
	1	OLS	Constant	1.71	[1.54, 1.87]	0.0%	0.94	[0.68, 1.20]	0.0%
Judges	2	OLS	Application Score	1.68	[1.50, 1.86]	1.9%	0.86	[0.61, 1.10]	8.9%
	3	OLS	Business Score	1.68	[1.50, 1.86]	1.9%	0.85	[0.61, 1.10]	9.1%
	4	OLS	SubScore	1.65	[1.48, 1.83]	3.2%	0.86	[0.61, 1.11]	8.2%
Single Predictor	5	OLS	Gender	1.65	[1.47, 1.83]	3.2%	0.92	[0.66, 1.18]	1.9%
	6	OLS	Age	1.64	[1.46, 1.82]	3.9%	0.91	[0.66, 1.16]	3.3%
	7	OLS	Necessity Firm	1.70	[1.52, 1.88]	1.3%	0.92	[0.67, 1.18]	1.5%
	8	OLS	Grit	1.67	[1.48, 1.86]	2.4%	0.93	[0.67, 1.19]	1.3%
	9	OLS	Digit-span Recall	1.69	[1.50, 1.88]	1.7%	0.93	[0.67, 1.18]	1.5%
	10	OLS	Raven Test	1.61	[1.43, 1.80]	5.6%	0.94	[0.68, 1.20]	0.6%
	11	OLS	Registration time	1.68	[1.50, 1.85]	2.0%	0.93	[0.67, 1.20]	0.8%
Economist Models	12	OLS	McKenzie	1.59	[1.40, 1.78]	6.9%	0.91	[0.66, 1.17]	2.9%
	13	OLS	FaWo	1.54	[1.35, 1.72]	10.3%	0.93	[0.66, 1.20]	2.2%
	14	OLS	FaWo + BusScores	1.54	[1.35, 1.72]	10.3%	0.86	[0.60, 1.13]	8.1%
	15	OLS	FaWo + Baseline	1.56	[1.38, 1.74]	9.0%	0.86	[0.61, 1.11]	8.3%
Machine Learning	16	LASSO	All	1.54	[1.38, 1.70]	10.6%	0.85	[0.60, 1.09]	15.1%
	17	SVM	All	1.57	[1.39, 1.76]	8.4%	0.88	[0.64, 1.12]	6.2%
	18	Boosting	All	1.53	[1.35, 1.71]	10.3%	0.91	[0.65, 1.16]	3.8%

Notes: outcome is the inverse hyperbolic sine of total employment in the firm three years after applying, coded as zero for applicants not operating firms. Models are estimating separately for competition non-winners and winners. MSE is the out-of-sample mean-squared error computed using the 20% hold-out sample. The numbers in brackets are bootstrapped 95 percent confidence intervals for hold-out prediction performance. R^2 is the squared correlation between the fitted outcome and actual outcome in the 20% hold-out sample. Judges models use scores from competition judges to predict the outcome. Single predictor models use a single survey measure. The economist models are the models of McKenzie (2015) and Fafchamps and Woodruff (2017) (abbreviated as FaWo) used in Tables 1-3. All models apart from the constant model include a dummy for whether the applicant was for an existing firm compared to a new firm. Models 2-4 and 14 for winners also control for the total award received (in logs).

Table 6: Prediction (Out-of-sample) Accuracy for Monthly Profits

				Non-winners			Winners		
				MSE		R^2	MSE		R^2
		Model	Predictors	Mean	C.I.		Mean	C.I.	
	1	OLS	Constant	38.89	[37.18, 40.61]	0.0%	20.38	[15.37, 25.39]	0.0%
Judges	2	OLS	Application Score	39.02	[36.41, 41.62]	1.0%	20.32	[15.36, 25.28]	0.6%
	3	OLS	Business Score	39.04	[36.41, 41.67]	1.0%	20.30	[15.35, 25.24]	0.7%
	4	OLS	SubScore	38.90	[36.14, 41.65]	1.3%	20.18	[15.38, 24.99]	1.3%
Single Predictor	5	OLS	Gender	38.00	[35.34, 40.67]	2.6%	20.39	[15.45, 25.33]	0.5%
	6	OLS	Age	38.71	[35.98, 41.44]	1.5%	20.04	[15.16, 24.91]	1.8%
	7	OLS	Necessity Firm	39.26	[36.62, 41.90]	0.8%	19.92	[15.11, 24.74]	2.7%
	8	OLS	Grit	38.91	[36.25, 41.56]	1.2%	19.88	[15.06, 24.71]	2.8%
	9	OLS	Digit-span Recall	38.57	[35.90, 41.24]	1.7%	19.80	[14.98, 24.61]	3.4%
	10	OLS	Raven Test	37.28	[34.37, 40.19]	4.4%	19.89	[15.04, 24.75]	2.5%
	11	OLS	Registration time	39.15	[36.49, 41.80]	0.9%	19.98	[15.12, 24.83]	2.1%
Economist Models	12	OLS	McKenzie	37.99	[34.51, 41.48]	4.0%	20.90	[15.83, 25.98]	0.4%
	13	OLS	FaWo	34.54	[30.75, 38.33]	11.2%	20.79	[15.80, 25.77]	0.6%
	14	OLS	FaWo + BusScores	34.55	[30.78, 38.31]	11.2%	21.04	[16.00, 26.08]	0.3%
Machine Learning	15	LASSO	All	35.23	[32.86, 37.60]	11.1%	20.32	[15.30, 25.34]	0.9%
	16	SVM	All	33.79	[30.50, 37.08]	13.1%	20.07	[14.85, 25.29]	1.3%
	17	Boosting	All	35.41	[31.51, 39.32]	9.2%	21.07	[16.05, 26.08]	0.1%

Notes: outcome is the inverse hyperbolic sine of monthly profits in the firm three years after applying, coded as zero for applicants not operating firms. Models are estimating separately for competition non-winners and winners. MSE is the out-of-sample mean-squared error computed using the 20% hold-out sample. The numbers in brackets are bootstrapped 95 percent confidence intervals for hold-out prediction performance. R^2 is the squared correlation between the fitted outcome and actual outcome in the 20% hold-out sample. Judges models use scores from competition judges to predict the outcome. Single predictor models use a single survey measure. The economist models are the models of McKenzie (2015) and Fafchamps and Woodruff (2017) (abbreviated as FaWo) used in Tables 1-3. All models apart from the constant model include a dummy for whether the applicant was for an existing firm compared to a new firm. Models 2-4 and 14 for winners also control for the total award received (in logs).

Table 7: Prediction (Out-of-sample) Accuracy for Tail

Panel A: Tail Total Employment

			Non-winners			Winners		
			Accuracy		Recall	Accuracy		Recall
	Model	Predictors	Mean	C.I.		Mean	C.I.	
1	OLS	Constant	79.2%	[74.9%, 83.4%]	7.4%	82.9%	[79.2%, 86.7%]	10.5%
2	OLS	Business Score	81.0%	[76.5%, 85.5%]	14.8%	83.9%	[79.8%, 87.9%]	15.8%
3	OLS	Raven Test	81.9%	[77.6%, 86.3%]	18.5%	84.8%	[81.0%, 88.7%]	21.1%
4	OLS	McKenzie	81.0%	[76.4%, 85.7%]	14.8%	84.8%	[80.9%, 88.8%]	21.1%
5	OLS	FaWo	81.9%	[77.5%, 86.4%]	18.5%	84.8%	[80.8%, 88.8%]	21.1%
6	LASSO	All	85.6%	[81.4%, 89.9%]	33.3%	87.7%	[83.7%, 91.7%]	36.8%
7	SVM	All	81.9%	[77.4%, 86.4%]	18.5%	83.9%	[80.0%, 87.7%]	15.8%
8	Boosting	All	83.8%	[79.1%, 88.5%]	25.9%	88.6%	[84.8%, 92.5%]	42.1%

Panel B: Tail Monthly Profits

			Non-winners			Winners		
			Accuracy		Recall	Accuracy		Recall
	Model	Predictors	Mean	C.I.		Mean	C.I.	
1	OLS	Constant	79.2%	[74.6%, 83.9%]	12.9%	80.3%	[76.2%, 84.3%]	8.3%
2	OLS	Business Score	80.2%	[75.3%, 85.1%]	16.1%	81.3%	[76.7%, 85.8%]	12.5%
3	OLS	Raven Test	81.1%	[76.4%, 85.9%]	19.4%	84.1%	[79.8%, 88.5%]	25.0%
4	OLS	McKenzie	80.2%	[75.5%, 84.9%]	16.1%	81.3%	[76.9%, 85.6%]	12.5%
5	OLS	FaWo	81.1%	[76.3%, 85.9%]	19.4%	80.3%	[75.9%, 84.7%]	8.3%
6	LASSO	All	81.1%	[76.2%, 86.0%]	19.4%	80.3%	[76.1%, 84.5%]	8.3%
7	SVM	All	83.0%	[78.3%, 87.8%]	25.8%	83.2%	[78.8%, 87.5%]	20.8%
8	Boosting	All	80.2%	[75.1%, 85.3%]	16.1%	81.3%	[77.0%, 85.5%]	12.5%

Panel C: How much difference would this make for an investor taking a share in the top firms?

			Non-winners		Winners	
	Model	Predictors	Mean	C.I.	Mean	C.I.
1	OLS	Constant	2.04	[-3.26, 7.33]	4.76	[-7.64, 17.15]
2	OLS	Business Score	4.88	[1.44, 8.32]	4.08	[1.14, 7.02]
3	OLS	Raven Test	5.96	[1.39, 10.54]	19.39	[-2.36, 41.14]
4	OLS	McKenzie	5.40	[0.01, 10.79]	7.92	[-3.03, 18.86]
5	OLS	FaWo	6.74	[0.76, 12.71]	7.26	[-1.78, 16.29]
6	LASSO	All	6.07	[1.72, 10.41]	7.98	[-0.04, 15.99]
7	SVR	All	7.66	[3.05, 12.27]	10.58	[2.33, 18.82]
8	Boosting	All	6.76	[0.79, 12.74]	8.42	[0.20, 16.64]

Notes: outcome in Panel A is whether a firm is in the top 10% for total employment three years after applying (employment coded as zero for applicants not operating firms). Outcome in Panel B is whether a firm is in the top 10% for monthly profits three years after applying (profits coded as zero for applicants not operating firms). Models are estimating separately for competition non-winners and winners. Firms are sorted based on the employment or profit levels predicted in Tables 5 and 6 respectively, and the top 10% are assigned value one, zero otherwise. Out-of-sample accuracy is the ratio of true positives (1) and true negatives (0) to all observations, and is computed using the 20% hold-out sample. The numbers in brackets are bootstrapped 95 percent confidence intervals for hold-out prediction performance. The recall rate is the percentage of businesses correctly predicted to be in the top 10% over the total number of businesses in the 20% hold-out sample actually in the top 10%. Profits in Panel C are in 1,000,000 Nigerian Naira (1 Nigerian Naira equals around 0.0028 US$). For each model, we have taken the predicted profits computed in Table 6, sort firms in the hold-out sample based on these predicted values, selected the top 21 firms (equivalent to around 10% of the hold-out sample), and summed the profits from these selected firms.

Online Appendix (NOT FOR PUBLICATION)

A.1 Variable Description

A.1.1 Outcome Variables

Three-year Survival is an indicator variable equal to one if a respondent operates a business three years after the baseline interview.

Three-year Employment is the total number of wage or salaried workers, as well as casual or daily workers, in the firm. Operating firms have at least one worker (the owner). Partners and unpaid workers are not included in this figure. If the respondent is not operating a business, this variable is zero. If the respondent does not know the number of wage or casual employees, it is assumed that there are no such workers in the firm. The same imputation is done if the respondent cannot say or refuses to say the number of these workers. This variable is set to missing if the third-year survival indicator is missing

Three-year Sales are the total monthly sales (in Naira) of the respondent's business from all sources - including manufacturing, trade and services - in the month preceding the interview. Sales are set to zero if the respondent is not operating a business, or if she does not know the figure.

Three-year Profits are the total income (in Naira) the business earn in the month preceding the interview after paying all expenses including wages of employees, but not including any income paid to the owner. Profits are set to zero if the respondent is not operating a business at that time, or if she does not know the figure.

A.1.2 Baseline Predictors

For each variable, missing values (if any) are set to zero and a new binary variable is generated to indicate the observations that are missing.

Female is an indicator variable equal to one if the respondent is a woman, zero if he is a man.

Age is the respondent's age in years. Participants have to be 18 or older.

Education is measured using four indicator variables to specify if the respondent's highest educational level is:

- High school or lower
- Post-secondary school vocational training/OND/Technical colleges
- University/HND degree
- Post graduate degree

When such information is missing, the same variable from the first follow-up is examined (if available).

Regional Indicators is constructed for each of the following geographical areas:

- North-Central
- North-Eastern
- North-Western
- South-Eastern
- South-South

- South-Western

Marital status is measured using an indicator variable for each of the following options:

- Married or Living Together with a Partner
- Divorced/Separated
- Widowed
- Never Married and Never Lived Together

Family Composition. Three variables measure the number of respondents' children, brothers and sisters.

Language measures the main language spoken at home using five different indicator variables, one for each of the following options:

- Hausa
- Yoruba
- Igbo
- English
- Other

Internal migrant indicates whether the participant's state of origin is different from the state of residence.

Business background. Participants are asked in the baseline interview whether they have ever attended before the start of the competition a course covering:

- Business accounting
- Marketing
- Computer skills
- Vocational training

Outside Employment. Participants are asked in the baseline interview whether they are employed outside their proposed business. If yes, they are also asked whether they are working part-time or full-time.

Experience abroad. Participants are asked in the baseline interview whether they have ever worked or studied abroad. If yes, they are asked how many years they have spent abroad.

Risk aversion. Participants are asked in the baseline interview to choose between a lottery which pays a prize of 5 million if they win, or receiving 1 million with certainty. The possible options are:

1. Prefer lottery ticket
2. Prefer 1 million certain amount
3. Indifferent
4. Don't know

One of the above options has to be selected for each of the following scenarios:

- A lottery ticket which has a 10% chance of paying 5 million, vs receiving 1 million with certainty
- A lottery ticket which has a 20% chance of paying 5 million, vs receiving 1 million with certainty
- A lottery ticket which has a 30% chance of paying 5 million vs receiving 1 million with certainty
- A lottery ticket which has a 50% chance of paying 5 million vs receiving 1 million with certainty

- A lottery ticket which has a 70% chance of paying 5 million vs receiving 1 million with certainty
- A lottery ticket which has a 90% chance of paying 5 million vs receiving 1 million with certainty
- A lottery ticket which has a 100% chance of paying 5 million vs receiving 1 million with certainty

An indicator variable for each scenario-choice combination is created.

Motivation. Participants are asked in the baseline interview to specify how important different motivations for running their own business rather than being employed for a salary are for them. The possible options are:

1. Very Important
2. Important
3. Somewhat important
4. Not important
5. Don't know
6. Not applicable

The possible motivations are:

- The possibility to care for children or other family members while I work
- To be my own boss and not have to rely on others
- Flexibility in hours of work
- The possibility to earn more money
- Difficulty in finding a wage job
- The possibility that my business will grow in the future
- Running my own business is less boring/more exciting than salary/wage work

An indicator variable for each motivation-evaluation combination is created, as well as an overall indicator of motivation obtained through principal component analysis.

Self-confidence. Participants are asked in the baseline interview to rate their confidence in carrying out certain tasks. The possible options are:

1. Not at all confident
2. Somewhat confident
3. Quite confident
4. Very confident
5. Don't know
6. Not applicable

The tasks listed are:

- Sell a new product or service to a new client
- Find and hire good employees to expand your business
- Obtain a good price for buying business inputs
- Persuade a bank to lend you money to start your business
- Estimate accurately the costs of a new project
- Manage an employee who is not a family member

- Correctly value your business if you wanted to sell it
- Resolve a difficult dispute with a business client in another city
- Identify when it is time to stop making or selling a product that is not selling well

An indicator variable for each task-rating combination is created, as well as an overall indicator of self-confidence obtained through principal component analysis.

Registration type indicates whether the participant owns a business before applying to YouWIN!

Registration time records the time when the participant submits her initial application. This variable is recorded in Unix epoch time.

Assets. Participants are asked in the baseline interview to specify the number of items owned by their households for each of the following categories:

- Sewing machine
- Gas cooker
- Stove (electric)
- Stove gas (table)
- Stove (kerosene)
- Fridge
- Freezer
- Air conditioner
- Washing Machine
- Electric Clothes Dryer
- Bicycle
- Motorbike
- Cars and other vehicles
- Generator
- Hi-Fi (Sound System)
- Microwave
- TV Set
- Computer
- DVD Player
- Satellite Dish

These variables are categorical. An overall indicator of *Wealth* is obtained through principal component analysis. In addition, two indicator variables are set equal to one if the participant's household has an Internet connection in the house or owns any land, respectively.

Business sector. Participants are asked in the baseline interview to select - among 32 categories - the sector that their current business (for existing owners) or planned business (for start-ups) is going to be in. An indicator variable for each of the possible sectors is created. In addition, an indicator for *Manufacturing* combines tailoring, dressmaking or shoe making, furniture, craft, and other manufacturing. *Retail* combines market trader, street vendor/hawker, shop, street kiosk, or other retail trade.

Employee education. Participants who already owns a business at the time of the initial application are asked to specify how many individuals they employ by educational level. The possible categories are:

- No Education
- Primary Education
- Secondary Education
- Tertiary Education

Business type. A series of binary variables indicates whether the participants who already own a business have registered their business with the Corporate Affair Commission and if yes, which kind of status they have chosen among:

- Limited liability company
- Business name
- Other

Tax paid records the amount paid in taxed in the year preceding the initial application by the individuals who already own a business at that time.

Loans are a series of binary variables indicating - for the participants who already own a business at the time of the initial application – whether their business has a loan at that time from any of the following:

- Bank
- Microfinance Organization
- Cooperative society

Credit has been constructed by combining the above three indicators through principal component analysis.

Business Challenges are a series of binary variables indicating - for the participants who already own a business at the time of the initial application – whether they identifies the following as major challenges for their business:

- Lack of funds to start a business
- Lack of the required skills
- Lack of business experience
- Lack of interest in business

Projected outcomes record the expectations of participants who already own a business at the time of the initial application. Information are recorded regarding:

- Money needed to expand business
- Number of new people the business will employ
- Projected annual turnover

A.1.3 Text Analysis

Business name indicates whether the business name includes the owner's surname in it. Three related measures are constructed:

- An indicator variable equal to one in case of perfect match between the participant's surname and the firm name.
- An indicator variable equal to one in case the participant's name, surname, initials or subsets of the name or surname are in the firm name.
- A measure of the Levenshtein distance between participant's surname and the firm name using the Stata command *strdist*.

Length records the length of the letter of intent in the initial applications (truncated at 244 characters). Individuals are asked to answer the following questions:

- Why do you want to be an Entrepreneur?
- How did you get your business Idea?
- Where do you Intend to locate your Business and Why?
- Why will you succeed?

Length First Answer records the length of first answer "Why do you want to be an Entrepreneur?" in the initial applications (truncated at 244 characters).

Capital Letter records whether the letter of intent in the initial application is written all in capital letters.

Employer of labor indicates whether the letter of intent includes the words "employer of labor" or "employer of labour".

No competition indicates whether the participant's answer to the question "What is the competition? How will it change and why are you better?" contains "no competition".

A.1.4 Business Plan

First Round Score records the total marks (out of 100) given to the initial application. The evaluation criteria concern the quality and viability of the business idea, the potential for survival and job creation, the applicant's ability, business skills, passion and commitment. Applications are evaluated by the Enterprise Development Center (EDC). To ensure an impartial scoring process, all names and identifying information are removed from the applications. Each application is evaluated in around 10 minutes. There is a conscious decision to favor existing businesses.

Business Plan Total Score is the total score (out of 100 points) assigned by the markers to the applicant's business plan. A set of variables also records the scores assigned based on each of the following criteria, which were designed by the team leads from EDC/LBS and PwC:

- Ability to manage (10 points)
- Articulation (10 points)
- Capstone (10 points)
- Financial sources (5 points)

- Financial sustainability (10 points)
- Financial viability (5 points)
- Impact job creation (25 points)
- Risk assessment (10 points)
- Time to market (5 points)
- Understanding (10 points)

Financial details. A set of variables records the amount sought by the applicants in their business plans, as well as their stated financial contribution to the business.

Final status is a set of indicator variables recording whether the final status of the participants is national merit winners, zonal merit winners, ordinary winners, disqualified.

Final award is the total award actually paid to the YouWin winners (in logarithms).

A.1.5 First Follow-up Predictors

Life satisfaction. Participants are asked to evaluate their current life satisfaction on a scale from 1 to 10, as well as to state their expected life satisfaction 5 years later:

> Imagine for a minute that you are living the best life you can possibly imagine. Now imagine that your life is the worst it could possibly be. Imagine a ladder with 10 steps. Suppose we say that the top of the ladder (step 10) represents the best possible life for you and the bottom (step 1) represents the worst possible life for you. Which step on the ladder best represents where you personally stand at the present time?
>
> Think about your life five years from today. Which step best represents where you personally will be on the ladder five years from now?

Self-confidence /2. As in the baseline interview, participants are asked to rate their confidence in carrying on certain tasks. The possible options are:

1. Not at all confident
2. Somewhat confident
3. Confident
4. Very confident
5. No answer / Refuse to answer

The tasks listed are:

- Come up with an idea for a new business product or service
- Estimate accurately the costs of a new business venture
- Estimate customer demand for a new product or service
- Sell a product or service to a customer they are meeting for the first time
- Identify good employees who can help a business grow
- Inspire, encourage, and motivate employees
- Find suppliers who will sell them raw materials at the best price
- Persuade a bank to lend them money to finance a business venture

- Correctly value a business if they were to buy an existing business from someone else

An indicator variable for each task-rating combination is created, as well as an overall indicator of self-confidence obtained through principal component analysis.

A.1.6 Second Follow-up Predictors

Digit-span Recall. Participants are showed a card with four numbers marked on it, for ten seconds. The interviewer then waits for 10 seconds before asking them to repeat the numbers in reverse order. If they get the numbers correctly at the first attempt, the procedure is repeated with 5 numbers, otherwise the interviewer moves to the next section, and so on until 11 numbers. Participants are given an example with 3 numbers. If the interview is conducted over the phone, the interviewer reads the numbers slowly over the phone, and then asks the respondent to repeat them back in reverse order. A series of binary variables indicates whether the respondent succeeds in each of the recall exercises. A binary variable indicates whether the interview is conducted over the phone (91 cases).

Raven Test. Participants are showed a series of pictures. Each picture has figures in two rows and three columns. One figure is missing. Participants have to find the missing part required to complete a pattern (among 8 options). They are given 5 minutes. During that time, they are asked to do as many as possible (out of 12). They are told that they can skip a picture and come back later. Two examples are given in order to clarify the task. This test is conducted only for individuals interviewed in person, not by phone. An indicator variable for each option-picture combination is created.

Grit. Participants are asked how much they identify with certain descriptions of their personality. The possible answers are:

1. Very much like me
2. Mostly like me
3. Somewhat like me
4. Not much like me
5. Not like me at all

The statements are:

- I have overcome setbacks to conquer an important challenge
- New ideas and projects sometimes distract me from previous ones.
- My interests change from year to year.
- Setbacks don't discourage me.
- I have been obsessed with a certain idea or project for a short time but later lost interest.
- I am a hard worker.
- I often set a goal but later choose to pursue a different one.
- I have difficulty maintaining my focus on projects that take more than a few months to complete.
- I finish whatever I begin.
- I have achieved a goal that took years of work.
- I become interested in new pursuits every few months.
- I am diligent.

An indicator variable for each option-description combination is created. An averaged indicator of grit is also computed.

Ability (McKenzie, 2015) is constructed by applying principal component analysis to the maximum correct numbers in the digit-span recall exercises, as well as the number of correct answers in the Raven test.

Ability (Fafchamps-Woodruff, 2017) is constructed by combining three indicators through principal component analysis: the number of correct answers in the Raven test, the maximum correct numbers in the digit-span recall exercises, and the respondent's educational level.

A.2 Human Approaches to Predicting Business Success: Additional Results

Table A1: Can Human Experts Predict Sales After 3 Years?

	Non-Winners			Winners			
	(1) Judges	(2) McKenzie	(3) Fafchamps	(4) Judges	(5) Judges	(6) McKenzie	(7) Fafchamps
Business Plan score	-0.0061	-0.0139	-0.0142	-0.0024	-0.0085	-0.0138	-0.0114
	(0.0196)	(0.0204)	(0.0199)	(0.0127)	(0.0132)	(0.0148)	(0.0144)
Female		-2.1592***	-1.8667***			-0.8560**	-0.7131*
		(0.5252)	(0.5165)			(0.4261)	(0.4256)
Age		0.1630***	0.1168***			0.0275	0.0175
		(0.0445)	(0.0439)			(0.0373)	(0.0365)
Graduate Education		-1.0281				0.1303	
		(0.8576)				(0.5890)	
Worked abroad		0.6583				-0.9308	
		(0.8441)				(0.5778)	
Lottery choice		0.0039				-0.3252	
		(0.4002)				(0.2907)	
Wealth		0.0175				0.0972**	
		(0.0780)				(0.0429)	
Grit		-0.3697				0.0629	
		(0.3854)				(0.2727)	
Agriculture		-0.3840				-0.4033	
		(0.4929)				(0.3713)	
IT		-0.5886				-0.4774	
		(0.6901)				(0.4336)	
Ability		0.7588***	0.5503**			-0.1925	-0.1837
		(0.2300)	(0.2200)			(0.1619)	(0.1515)
Manufacture		0.5160	0.7886			-0.0373	0.2436
		(0.5326)	(0.4810)			(0.3860)	(0.3648)
Retail			1.1057				1.7743***
			(1.0513)				(0.5247)
Self-confidence			0.1348				0.0090
			(0.1220)				(0.0833)
Motivation			0.0595				-0.1375
			(0.1468)				(0.1016)
Happiness			0.0698				0.0702
			(0.1653)				(0.1298)
Optimism			0.4525**				-0.1093
			(0.1876)				(0.1375)
Credit			0.2958*				0.0670
			(0.1627)				(0.1977)
Existing	3.0602***	2.4878***	0.9168	1.0565***	1.0309***	0.7446**	0.4872
	(0.5071)	(0.5432)	(0.6123)	(0.2872)	(0.2886)	(0.3631)	(0.4319)
Log(Total award paid)					0.5965**	0.5699*	0.6043**
					(0.2931)	(0.2944)	(0.2956)
Regional Indicators	No	Yes	Yes	No	No	Yes	Yes
Sample Size	1,058	1,058	1,058	1,036	1,036	1,036	1,036
P-value		0.000	0.000			0.000	0.002
Adjusted R^2	0.037	0.097	0.133	0.012	0.014	0.018	0.019
MSE	43.67	40.96	39.29	20.47	20.41	20.35	20.32

Robust Standard Errors in parentheses. *, **, *** denote significance at the 10, 5, and 1 percent levels respectively. See Appendix A.1 for variable definitions. Dependent variable is inverse hyperbolic sine of sales 3 years after applying, coded as zero for applicants without an operating firm. P-value is for testing that set of controls added by McKenzie (2015) or Fafchamps and Woodruff (2017) (abbreviated as FaWo) are jointly zero.

Table A2: Robustness to Including Judge and Region Fixed Effects

	Non-winners				Winners			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Survival	Employment	Sales	Profits	Survival	Employment	Sales	Profits
Panel A: Without Judge Fixed Effects								
Business Plan score	-0.0002	-0.0023	-0.0061	-0.0045	-0.0000	0.0881***	-0.0024	-0.0030
	(0.0014)	(0.0168)	(0.0196)	(0.0180)	(0.0006)	(0.0312)	(0.0127)	(0.0126)
Sample Size	1,077	1,077	1,058	1,058	1,051	1,051	1,036	1,036
Adjusted R^2		0.012	0.037	0.034		0.020	0.012	0.010
Panel B: With Judge Fixed Effects								
Business Plan score	-0.0012	0.0014	-0.0132	-0.0147	-0.0004	0.1251***	-0.0097	-0.0082
	(0.0016)	(0.0176)	(0.0218)	(0.0197)	(0.0008)	(0.0340)	(0.0144)	(0.0141)
Sample Size	1,067	1,077	1,058	1,058	953	1,051	1,036	1,036
Adjusted R^2		0.002	0.032	0.032		0.025	0.031	0.022
Panel C: With Region Fixed Effects								
Business Plan score	-0.0001	0.0126	-0.0008	0.0036	-0.0003	0.1183***	-0.0070	-0.0041
	(0.0016)	(0.0177)	(0.0212)	(0.0193)	(0.0008)	(0.0369)	(0.0140)	(0.0138)
Sample Size	1,077	1,077	1,058	1,058	1,051	1,051	1,036	1,036
Adjusted R^2		0.014	0.034	0.032		0.030	0.009	0.007

Robust standard errors in parentheses, *, **, *** denote significance at the 10, 5, and 1 percent levels respectively. See Appendix A.1. for variable definitions. Employment, Sales and Profits are unconditional outcomes, coded as zero for non-operating firms.

Table A3: Do Sub-Scores of Judges Predict Outcomes?

	Non-Winners				Winners			
	(1) Survival	(2) Employment	(3) Sales	(4) Profits	(5) Survival	(6) Employment	(7) Sales	(8) Profits
Score: ability to manage	0.030***	0.400***	0.347***	0.294**	0.002	0.228	0.095	0.054
	(0.010)	(0.119)	(0.131)	(0.123)	(0.005)	(0.177)	(0.089)	(0.088)
Score: articulation	-0.003	-0.090	-0.165	-0.144	0.002	-0.158	0.160	0.141
	(0.013)	(0.140)	(0.179)	(0.165)	(0.006)	(0.223)	(0.117)	(0.117)
Score: capstone	-0.009	-0.178	-0.198	-0.037	-0.009	0.182	-0.198	-0.228*
	(0.014)	(0.198)	(0.191)	(0.175)	(0.007)	(0.238)	(0.123)	(0.125)
Score: financial sources	0.011	0.077	0.050	0.026	-0.003	0.414*	-0.021	0.034
	(0.012)	(0.132)	(0.158)	(0.146)	(0.006)	(0.223)	(0.107)	(0.107)
Score: financial sustainability	-0.011	-0.072	-0.164	-0.115	-0.004	-0.110	-0.029	-0.018
	(0.011)	(0.128)	(0.146)	(0.138)	(0.006)	(0.173)	(0.086)	(0.089)
Score: financial viability	0.008	0.004	0.208	0.147	0.001	-0.147	-0.100	-0.112
	(0.017)	(0.191)	(0.224)	(0.211)	(0.008)	(0.316)	(0.139)	(0.139)
Score: job creation	0.000	0.068*	0.008	0.002	-0.000	0.189***	-0.010	0.021
	(0.003)	(0.035)	(0.040)	(0.037)	(0.001)	(0.059)	(0.023)	(0.025)
Score: risk assessment	-0.002	-0.044	-0.008	0.013	-0.001	-0.042	-0.087	-0.060
	(0.007)	(0.081)	(0.094)	(0.087)	(0.003)	(0.141)	(0.059)	(0.060)
Score: time to market	-0.003	-0.057	-0.022	-0.080	0.005	-0.323	0.093	0.092
	(0.010)	(0.122)	(0.141)	(0.131)	(0.006)	(0.237)	(0.101)	(0.102)
Score: understanding	-0.010	-0.133	0.002	-0.086	0.006	-0.094	0.058	-0.001
	(0.012)	(0.139)	(0.162)	(0.148)	(0.006)	(0.196)	(0.107)	(0.109)
Existing	0.172***	1.032*	2.549***	2.413***	0.025	2.121***	0.692**	0.707**
	(0.042)	(0.552)	(0.556)	(0.519)	(0.019)	(0.641)	(0.342)	(0.342)
Log(Total award paid)					0.025*	3.608***	0.585**	0.392
					(0.014)	(0.433)	(0.296)	(0.300)
Constant		3.515***	6.573***	5.916***		-49.088***	2.344	4.264
		(0.958)	(1.155)	(1.062)		(6.899)	(4.561)	(4.629)
Sample Size	1,077	1,077	1,058	1,058	1,051	1,051	1,036	1,036
Adjusted R^2		0.020	0.037	0.033		0.062	0.015	0.010
P-value that scores jointly zero	0.339	0.039	0.418	0.625	0.891	0.050	0.472	0.566

Robust Standard errors in parentheses. *, **, *** denote significance at the 10, 5, and 1 percent levels respectively. P-value is for F-test that all ten sub-scores are jointly unrelated to the business outcome three years after application.

Table A4: Predicting Outcomes for Non-Winners Conditional on Operating a Firm

	Employment			Sales			Profits		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Judges	McKenzie	FaWo	Judges	McKenzie	FaWo	Judges	McKenzie	FaWo
Business Plan score	-0.0015	0.0015	0.0133	-0.0057	-0.0010	-0.0019	-0.0035	0.0086	0.0089
	(0.0237)	(0.0262)	(0.0260)	(0.0137)	(0.0150)	(0.0146)	(0.0137)	(0.0153)	(0.0149)
Female		-0.0900	0.1945		-0.6231	-0.5917		-0.3258	-0.1434
		(0.8291)	(0.8688)		(0.4898)	(0.5002)		(0.4839)	(0.4900)
Age		0.0649	0.0477		0.0445	0.0316		0.0048	-0.0043
		(0.0590)	(0.0576)		(0.0366)	(0.0359)		(0.0354)	(0.0355)
Graduate Education		1.7346			-1.0596			-1.0168	
		(1.3208)			(0.8020)			(0.7696)	
Worked abroad		0.1126			1.0058			1.2066*	
		(1.2587)			(0.7147)			(0.6615)	
Lottery choice		1.1042**			0.4946			0.3414	
		(0.4994)			(0.3403)			(0.3320)	
Wealth		0.1927*			0.0733			0.0113	
		(0.1072)			(0.0625)			(0.0565)	
Grit		-0.3765			0.0110			-0.1281	
		(0.5093)			(0.3318)			(0.3248)	
Agriculture		-0.4910			0.3381			0.1937	
		(0.7322)			(0.4249)			(0.4051)	
IT		-1.6523**			0.1593			-0.3260	
		(0.6767)			(0.4819)			(0.5861)	
Ability		0.1621	0.1560		0.2644*	0.2390*		0.1981	0.1300
		(0.2750)	(0.2386)		(0.1570)	(0.1449)		(0.1706)	(0.1506)
Manufacture		-0.9742	-0.5383		0.7376*	0.6008		0.3268	0.2296
		(0.6507)	(0.6013)		(0.4221)	(0.3660)		(0.4098)	(0.3481)
Retail			0.4716			0.6569			0.3917
			(1.1551)			(0.7945)			(0.7088)
Self-confidence			0.0320			0.0446			-0.0337
			(0.1819)			(0.1085)			(0.1079)
Motivation			-0.0700			-0.2204*			-0.1962*
			(0.1523)			(0.1264)			(0.1181)
Happiness			0.3130*			0.2105*			0.2241*
			(0.1798)			(0.1194)			(0.1209)
Optimism			-0.0850			0.1251			0.1126
			(0.2621)			(0.2074)			(0.1863)
Credit			-0.1595			0.1398*			0.1314**
			(0.1614)			(0.0716)			(0.0638)
Existing	0.3867	0.6156	-0.1350	0.8023**	0.8309**	0.2874	0.6643*	0.9444**	0.3440
	(0.5644)	(0.5888)	(0.7419)	(0.3354)	(0.3595)	(0.3992)	(0.3546)	(0.3935)	(0.4307)
Regional Indicators	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Sample Size	629	629	629	610	610	610	610	610	610
P-value		0.000	0.000		0.000	0.000		0.000	0.000
Adjusted R^2	-0.002	0.017	0.004	0.007	0.019	0.035	0.004	0.011	0.038

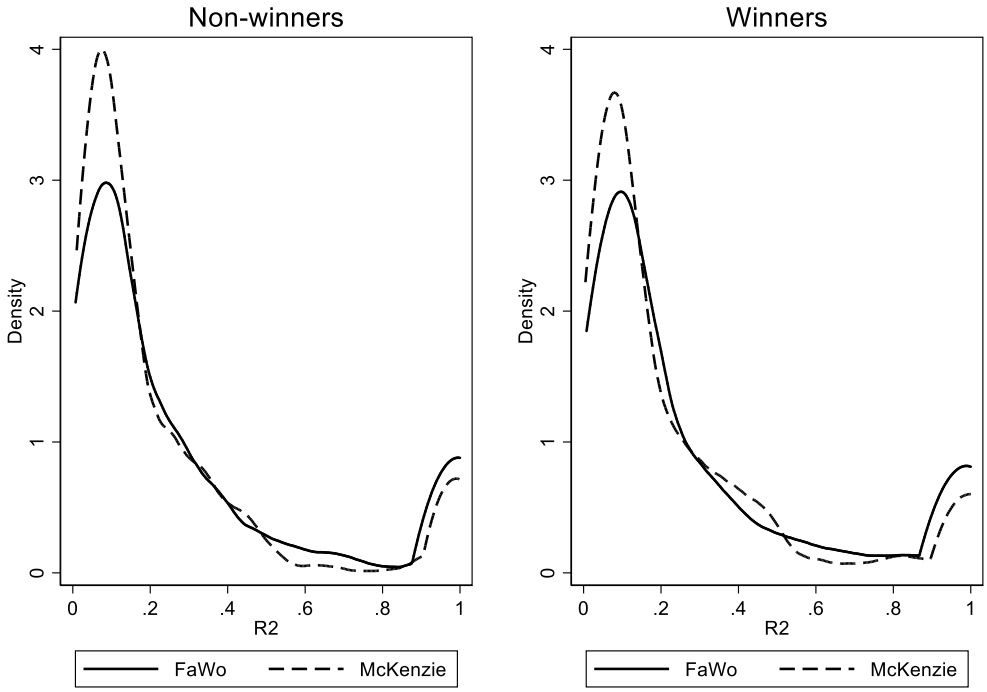
Robust Standard Errors in parentheses. *, **, *** denote significance at the 10, 5, and 1 percent levels respectively. P-value is for testing that set of controls added by McKenzie (2015) or Fafchamps and Woodruff (2017) (abbreviated as FaWo) are jointly zero.

A.3 Machine Learning: Technical Appendix

The aim of this appendix is to provide a quick overview of how the machine learning algorithms used in the paper have been calibrated, as well as additional technical details.

A.3.1 Inputs

Figure A1: Do ML inputs provide additional information?



Notes: This figure shows that the inputs used in the ML algorithms contain more information than the variables used in the models by McKenzie (2015) or Fafchamps and Woodruff (2017) (abbreviated as FaWo). Each ML input has been regressed on the McKenzie (2015) or Fafchamps and Woodruff (2017) set of variables (OLS for non-binary variables, Probit for binary variables). This figure shows kernel densities of the resulting R^2 (or pseudo-R^2 for binary variables), for winners and non-winners separately. Winners include both non-experimental and experimental winners, and non-winners include control group semi-finalists as well as a sample of firms submitting business plans which had first round application scores just above the cut-off for getting selected for business plan training.

Table A5: Canonical analysis. Economist models vs ML set of inputs

	McKenzie vs ML		Fafchamps and Woodruff vs. ML	
Order	Non-winners	Winners	Non-winners	Winners
1	0.9927***	0.9843***	0.9949***	0.9975***
2	0.9831***	0.9757***	0.9838***	0.9672***
3	0.9781***	0.9667***	0.8104***	0.9494***
4	0.8372***	0.9055***	0.7636***	0.7572***
5	0.7766***	0.8175***	0.7209***	0.7067***
6	0.7340***	0.7424***	0.6868***	0.6809***
7	0.7085***	0.7074***	0.6682***	0.6517***
8	0.6754***	0.6595***	0.5307***	0.5652***
9	0.5682***	0.6264***	0.5048***	0.5282***
10	0.5570***	0.5771***	0.4500***	0.5170***
11	0.5160***	0.5265***	0.4327***	0.4422***
12	0.4739***	0.5014***	0.4040*	0.4267***
13	0.4265**	0.4715***	0.3930	0.3936**
14	0.4070	0.4543***	0.3710	0.3922
15	0.3785	0.4407***	0.3609	0.3790
16	0.3415	0.3965*	0.3309	0.3441
17	0.3302	0.3870	0.3184	0.3377
18	0.3098	0.3458	0.2965	0.3175
19	0.2909	0.3340	0.2841	0.3143
20	0.2588	0.3006	0.2298	0.3039
21	0.2278	0.2904		

Notes: This table reports the canonical correlations from describing the relationship between the set of input used in the economist models and the ones used in the ML algorithms. The latter does not include in this case variables that are perfectly collinear with those used in the economist models. s. As explained in the Stata 15 Manual (command *canon*), given these two sets of variables, say $X = (x_1, x_2, \dots, x_K)$ for the variables in the economist models and $Y = (y_1, y_2, \dots, y_K)$ for those in the ML algorithms, the goal is to find linear combinations of X and Y so that the correlation between the linear combinations is as high as possible. That is, letting $\hat{x}_1 = \beta_{11}x_1 + \beta_{12}x_2 + \cdots + \beta_{1k}x_k$ and $\hat{y}_1 = \gamma_{11}y_1 + \gamma_{12}y_2 + \cdots + \gamma_{1L}y_L$ be the linear combinations, the first canonical correlation is the maximum correlation between $\hat{x}_1$ and $\hat{y}_1$ as functions of the β's and the γ's. The second canonical correlation coefficient is defined as the correlation between $\hat{x}_2 = \beta_{21}x_1 + \beta_{22}x_2 + \cdots + \beta_{2k}x_k$ and $\hat{y}_2 = \gamma_{21}y_1 + \gamma_{22}y_2 + \cdots + \gamma_{2L}y_L$. This correlation is maximized subject to the constraints that $\hat{x}_1$ and $\hat{x}_2$, along with $\hat{y}_1$ and $\hat{y}_2$, are orthogonal and that $\hat{x}_1$ and $\hat{y}_2$, along with $\hat{x}_2$ and $\hat{y}_1$, are also orthogonal. The third and further correlations are defined similarly. There are $m = min(K, L)$ such correlations. Models are estimating separately for competition non-winners and winners, as well as for the McKenzie (2015) and Fafchamps and Woodruff (2017) set of variables. *, **, *** denote significance at the 10, 5, and 1 percent levels respectively (Wilks' lambda). The stars next to the first canonical correlation coefficient indicate whether all canonical correlations are significant. The stars next to the second canonical correlation coefficient indicate whether canonical correlations 2, 3, …, *m* are significant; and so on.

A.3.2 Data cleaning

The usual caveat in these techniques is to normalize with zero mean and unit variance all the variables, or to restrict their domain between zero and one, so that the regularization is not inflated by the different scale of the variables. As also mentioned in Guenther and Schonlau (2016), both normalization or rescaling methods should work correctly. Following Mullainathan and Spiess (2017), the regularization need to be done using only information, such as mean and standard deviation, obtained from the training sample, not the hold-out sample.

When implementing the ML algorithms with a continuous variable (employment, sales, profits), we have adjusted the predictions to take into account that employment and sales levels cannot be negative. Therefore, both during the cross-validation procedure and in the final estimation, we have imputed all negative predictions with the minimum value of the dependent variable in the training sample (zero per employment and sales, slightly negative in a few cases for profits) before computing the in-sample and out-of-sample R^2 and MSE.

Given the limited sample size, we have use the inverse hyperbolic sine transformation when implementing the ML algorithms with a continuous variable (employment, sales, and profits) to reduce the risk of in-sample outliers driving the choice of the parameters during cross-validation, thus potentially leading to over-fitting. In addition, we have adjusted extremely high predictions. More specifically, we have replaced predictions that exceeded twice the maximum value of the dependent variable in the training sample with such threshold value. In most cases, no prediction exceeds such threshold.

A.3.3 LASSO

We have implemented LASSO in Stata using the package *elasticregress* developed by Wilbur Townsend. We have included all possible predictors as inputs in the algorithms. We have then used a grid-search among around 70 values (mainly between 0 and 1) to find the optimal penalization term λ, that is the one that maximizes accuracy or minimized the MSE in the 5-fold cross-validation procedure described in Section 4.2.

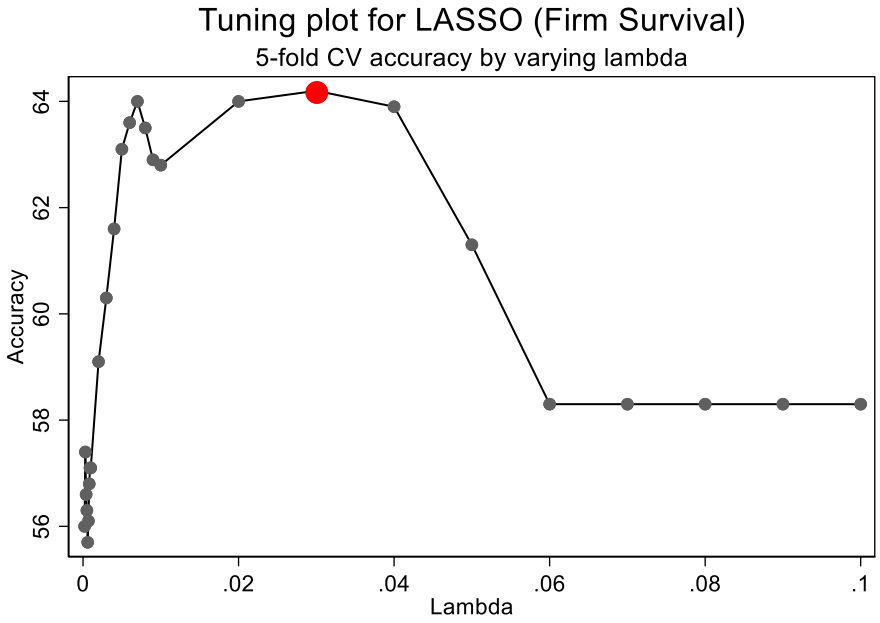
As shown in Table A6, the range of possible values for the penalization term λ is large, and the selected values are in the interior of this range. Figure A2 plots the relation between λ and the 5-fold average CV accuracy when predicting firm survival. This figure shows that the choice of λ substantially affects the LASSO performance. The same conclusion is reached by plotting the relation between λ and the 5-fold average CV MSE for employment (Figure A3) and sales (Figure A4) among non-winning firms. It is also possible to compare the in-sample and out-of-sample accuracy rates in Tables 4-6, A10 and A11-14 to verify that the 5-fold CV average is close to the out-of-sample performance, thus confirming that the algorithm is not over-fitting or underfitting the training data.

By default, the LASSO command in Stata automatically drops perfectly collinear variables. Table A7 also includes perfectly collinear variables – in addition to the aggregate variables already included in the main analysis - among the set of possible inputs in LASSO. Results are similar to those in Tables 4-6 and A10.

Table A6: LASSO tuning

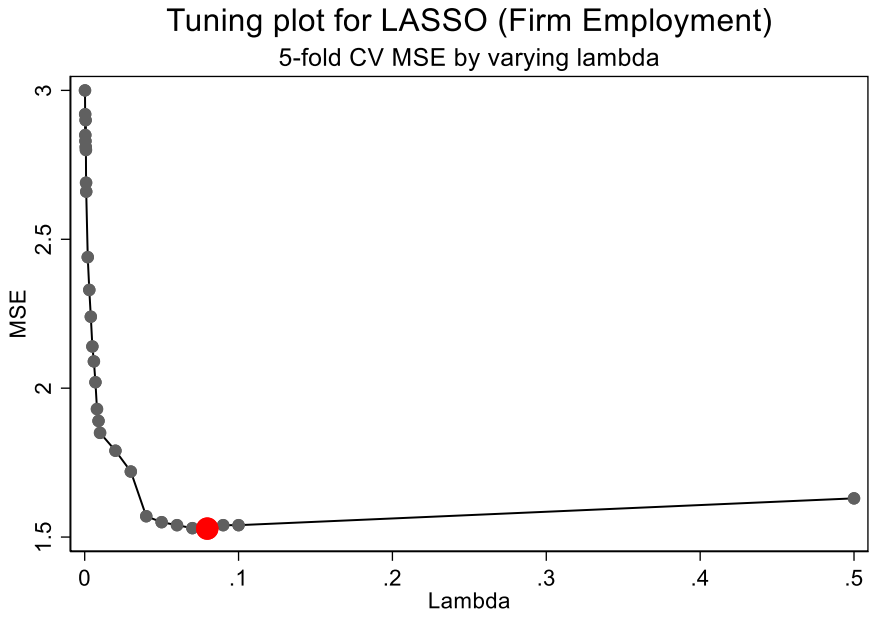
	λ				
Outcome	Min	Max	#values	Selected	#Predictors
Survival	0	50	70	0.03	24
Employment No Win	0	50	70	0.08	26
Employment Win	0	50	70	0.08	11
Sales No Win	0	50	70	0.5	20
Sales Win	0	50	70	0.5	2
Profits No Win	0	50	70	0.5	14
Profits Win	0	50	70	0.5	2

Figure A2: LASSO Tuning for Business Survival among Non-winners



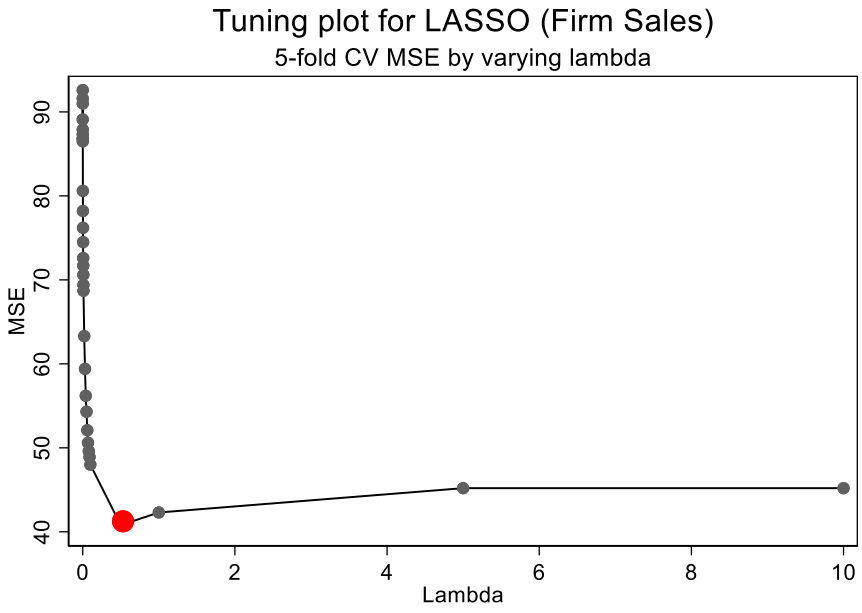
Note: for visual reason, the plot shows only a subset of the possible values for the penalization term λ.

Figure A3: LASSO Tuning for Total Employment among Non-winners



Note: for visual reason, the plot shows only a subset of the possible values for the penalization term λ.

Figure A4: LASSO Tuning for Monthly Sales among Non-winners



Note: for visual reason, the plot shows only a subset of the possible values for the penalization term λ.

Table A7: LASSO with perfectly collinear variables as possible inputs

Outcome	Group	Accuracy	MSE	R^2
Survival	Non-winners	59.7% [53.3%, 66.1%]		
Employment	Non-winners		1.52 [1.35, 1.68]	11.6%
Employment	Winners		0.84 [0.60, 1.09]	15.5%
Sales	Non-winners		40.81 [37.59, 44.02]	13.4%
Sales	Winners		20.08 [14.05, 26.11]	0.5%
Profits	Non-winners		34.76 [32.35, 37.17]	12.7%
Profits	Winners		20.32 [15.30, 25.34]	0.9%

Note: Out-of-sample accuracy is the ratio of true positives and true negatives to all observations, and is computed using the 20% hold-out sample. MSE is the out-of-sample mean-squared error computed using the 20% hold-out sample. The numbers in brackets are bootstrapped 95 percent confidence intervals for hold-out prediction performance. R^2 is the squared correlation between the fitted outcome and actual outcome in the 20% hold-out sample. The set of possible predictors does not include the initial application scores, the business scores, or the total amount awarded to winners. The set of possible predictors does include perfectly collinear variables.

A.3.4 Support Vector Machines

We have implemented Support Vector Machines and Support Vector Regression in Stata using the command *svmachines* (Guenther and Schonlau, 2016). We have included all possible predictors as inputs variables. A grid-search has been used to find the optimal parameters, that is the ones that maximize accuracy or minimized the MSE in the 5-fold cross-validation procedure described in Section 4.2.

In case of continuous outcome variables, we have considered 9 possible values for each parameter (the penalization term C, the kernel smoothing parameter γ, and the bandwidth ε), as well as one kernel (Guassian) for a total of 729 combinations ($1x9^3$). A similar grid-search has been conducted with the binary outcome variable survival, although in that case it has been necessary to calibrate only two parameters (the penalization term C and the kernel smoothing parameter γ) combined with the same Gaussian kernels. We have considered 16 possible values for each parameter, as well as one kernels (Gaussian) for a total of 256 combinations ($1x16^2$). Using sigmoid kernel provides similar results. As shown in Table A8, the selected values are in the interior of the provided intervals.

Table A8: SVM tuning

		C				γ				ε			
Outcome	Kernel	Min	Max	#val	Sel	Min	Max	#val	Sel	Min	Max	#val	Sel
Survival	Normal	5e-03	1e+05	16	1	5e-07	10	16	0.01	NA	NA	NA	NA
Employment No Win	Normal	5	5e+04	9	5000	1e-09	1e-05	9	5e-08	5e-07	5e-03	9	5e-05
Employment Win	Normal	500	5e+06	9	5e+05	1e-10	1e-06	9	5e-09	1e-03	10	9	1
Sales No Win	Normal	0.5	5000	9	5	5e-07	5e-03	9	5e-04	1e-07	1e-02	9	5e-04
Sales Win	Normal	0.5	5000	9	5	1e-04	1	9	0.05	1e-03	10	9	1
Profits No Win	Normal	5e-02	500	9	5	1e-04	1	9	0.05	1e-03	10	9	0.5
Profits Win	Normal	0.5	5000	9	5	1e-04	1	9	0.05	1e-03	10	9	1

A.3.5 Boosted Regression

Boosting has been implemented in Stata using the command *boost* (Schonlau, 2005). A grid-search has been used to find the optimal parameters, that is the ones that maximize accuracy or minimized the MSE in the in the 5-fold cross-validation procedure described in Section 4.2.

In addition to the usual boosting parameters (number of trees and interactions), we have introduced the option for the algorithm to select a shrinking parameter to avoid over-fitting. This reduces the contribution of each additional tree, thus decreasing the impact of an over-fitted tree. The cost of this procedure is a substantial increase in computational time, since it is necessary to increase the number of iterations in order to compensate for these smaller steps. Bagging is an additional technique which we have used to reduce the variance of the final prediction without influencing the bias. At each iteration, we have used only a random subset of the train set (50% of it) to build the tree.

As shown in Table A9, we have considered trees with up to 8 splits, up to 150 iterations (i.e. number of trees) and two values for the shrinkage parameter (1 and 0.1), for a total of 2400 combinations (8*150*2). In addition to this, for the binary dependent variable survival we have also considered two error distribution (Gaussian and Logistic), thus computing 5800 combinations for each outcome-fold pair.

Table A9: Boosting tuning

		# tree splits				# trees				Shrinkage			
Outcome	Distribution	Min	Max	#val	Sel	Min	Max	#val	Sel	Min	Max	#val	Sel
Survival	Logistic	1	8	8	1	1	150	150	134	0.1	1	2	0.1
Employment No Win	Normal	1	8	8	1	1	150	150	94	0.1	1	2	0.1
Employment Win	Normal	1	8	8	8	1	150	150	19	0.1	1	2	0.1
Sales No Win	Normal	1	8	8	1	1	150	150	113	0.1	1	2	0.1
Sales Win	Normal	1	8	8	6	1	150	150	13	0.1	1	2	0.1
Profits No Win	Normal	1	8	8	1	1	150	150	135	0.1	1	2	0.1
Profits Win	Normal	1	8	8	1	1	150	150	43	0.1	1	2	0.1

A.4 Machine Learning: Additional Results

Table A10: Prediction (Out-of-sample) Accuracy for Monthly Sales

				Non-winners			Winners		
				MSE		R^2	MSE		R^2
		Model	Predictors	Mean	C.I.		Mean	C.I.	
	1	OLS	Constant	46.17	[43.80, 48.55]	0.0%	20.01	[14.00, 26.02]	0.0%
Judges	2	OLS	Application Score	46.89	[43.37, 50.41]	0.6%	19.99	[13.98, 26.00]	0.5%
	3	OLS	Business Score	46.93	[43.36, 50.49]	0.6%	20.01	[14.00, 26.02]	0.5%
	4	OLS	SubScore	46.45	[42.78, 50.13]	1.1%	20.01	[14.10, 25.92]	0.7%
Single Predictor	5	OLS	Gender	45.55	[41.97, 49.14]	2.0%	19.78	[13.91, 25.65]	1.2%
	6	OLS	Age	46.35	[42.78, 49.91]	1.2%	19.71	[13.81, 25.62]	1.5%
	7	OLS	Necessity Firm	47.17	[43.47, 50.87]	0.5%	19.59	[13.77, 25.41]	2.3%
	8	OLS	Grit	46.38	[42.99, 49.78]	1.1%	19.63	[13.76, 25.50]	2.0%
	9	OLS	Digit-span Recall	46.15	[42.70, 49.60]	1.4%	19.55	[13.73, 25.37]	2.4%
	10	OLS	Raven Test	43.96	[40.51, 47.41]	5.0%	19.72	[13.81, 25.63]	1.5%
	11	OLS	Registration time	47.10	[43.53, 50.66]	0.5%	19.35	[13.59, 25.11]	3.6%
Economist Models	12	OLS	McKenzie	44.22	[40.20, 48.24]	5.3%	20.61	[14.47, 26.75]	0.2%
	13	OLS	FaWo	41.13	[36.67, 45.60]	10.9%	20.42	[14.38, 26.47]	0.5%
	14	OLS	FaWo + BusScores	41.23	[36.81, 45.65]	10.7%	20.78	[14.62, 26.94]	0.2%
Machine Learning	15	LASSO	All	40.97	[37.78, 44.17]	13.0%	20.08	[14.05, 26.11]	0.5%
	16	SVM	All	43.23	[37.88, 48.58]	9.8%	19.75	[13.49, 26.02]	1.6%
	17	Boosting	All	40.80	[36.62, 44.98]	11.4%	20.53	[14.48, 26.57]	0.2%

Notes: outcome is the inverse hyperbolic sine of monthly sales in the firm three years after applying, coded as zero for applicants not operating firms. Models are estimating separately for competition non-winners and winners. MSE is the out-of-sample mean-squared error computed using the 20% hold-out sample. The numbers in brackets are bootstrapped 95 percent confidence intervals for hold-out prediction performance. R^2 is the squared correlation between the fitted outcome and actual outcome in the 20% hold-out sample. Judges models use scores from competition judges to predict the outcome. Single predictor models use a single survey measure. The economist models are the models of McKenzie (2015) and Fafchamps and Woodruff (2017) (abbreviated as FaWo) used in Tables 1-3. All models apart from the constant model include a dummy for whether the applicant was for an existing firm compared to a new firm. Models 2-4 and 14 for winners also control for the total award received (in logs).

Table A11: Prediction (In-sample) Accuracy for Business Survival among Non-winners.

				Accuracy	
		Model	Predictors	80% sample	5-fold CV
	1	Logit	Constant	58.3%	
Judges	2	Logit	Application Score	58.3%	
	3	Logit	Business Score	58.3%	
	4	Logit	SubScore	59.7%	
Single Predictor	5	Logit	Gender	61.2%	
	6	Logit	Age	59.3%	
	7	Logit	Necessity Firm	60.7%	
	8	Logit	Grit	60.0%	
	9	Logit	Digit-span Recall	59.2%	
	10	Logit	Raven Test	63.0%	
	11	Logit	Registration time	58.3%	
Economist Models	12	Logit	McKenzie	64.7%	
	13	Logit	FaWo	66.6%	
	14	Logit	FaWo + BusScores	66.8%	
	15	OLS	FaWo + BusScores	66.4%	
	16	Probit	FaWo + BusScores	66.7%	
Machine Learning	17	LASSO	All	69.0%	64.2%
	18	SVM	All	79.3%	66.2%
	19	Boosting	All	73.2%	65.9%
		Full sample mean	58.4% (N=1,077)		

See notes in Table 4.

Table A12: Prediction (In-sample) Accuracy for Total Employment

				Non-winners			Winners		
				MSE		R^2	MSE		R^2
		Model	Predictors	80% sample	5-fold CV	80% sample	80% sample	5-fold CV	80% sample
	1	OLS	Constant	1.63		0.0%	0.95		0.0%
	2	OLS	Application Score	1.57		3.5%	0.89		5.8%
Judges	3	OLS	Business Score	1.57		3.5%	0.89		5.9%
	4	OLS	SubScore	1.54		5.0%	0.88		7.2%
	5	OLS	Gender	1.55		4.7%	0.93		1.6%
	6	OLS	Age	1.55		4.8%	0.93		2.2%
	7	OLS	Necessity Firm	1.56		4.1%	0.93		1.6%
Single Predictor	8	OLS	Grit	1.55		4.4%	0.93		2.1%
	9	OLS	Digit-span Recall	1.55		4.8%	0.93		2.1%
	10	OLS	Raven Test	1.52		6.3%	0.93		2.1%
	11	OLS	Registration time	1.57		3.5%	0.93		1.9%
	12	OLS	McKenzie	1.45		10.6%	0.91		4.3%
Economist Models	13	OLS	FaWo	1.40		13.7%	0.90		5.2%
	14	OLS	FaWo + BusScores	1.40		13.7%	0.86		9.4%
	15	OLS	FaWo + Baseline	1.38		15.0%	0.86		9.1%
	16	LASSO	All	1.35	1.53	20.1%	0.87	0.90	10.5%
Machine Learning	17	SVM	All	1.41	1.47	13.4%	0.82	0.91	15.5%
	18	Boosting	All	1.21	1.42	27.6%	0.63	0.88	46.8%
			Full sample mean	1.34 (N=1,077)			2.68 (N=1,051)		

See notes in Table 5.

Table A13: Prediction (In-sample) Accuracy for Monthly Sales

				Non-winners			Winners		
				MSE		R^2	MSE		R^2
		Model	Predictors	80% sample	5-fold CV	80% sample	80% sample	5-fold CV	80% sample
	1	OLS	Constant	45.09		0.0%	20.86		0.0%
	2	OLS	Application Score	42.78		5.1%	20.46		1.9%
Judges	3	OLS	Business Score	42.77		5.2%	20.45		2.0%
	4	OLS	SubScore	42.50		5.8%	20.27		2.8%
	5	OLS	Gender	42.44		5.9%	20.05		1.7%
	6	OLS	Age	42.15		6.5%	20.57		1.4%
	7	OLS	Necessity Firm	42.60		5.5%	20.53		1.6%
Single Predictor	8	OLS	Grit	42.38		6.0%	20.58		1.3%
	9	OLS	Digit-span Recall	42.35		6.1%	20.53		1.6%
	10	OLS	Raven Test	41.05		9.0%	20.51		1.7%
	11	OLS	Registration time	42.76		5.2%	20.46		1.9%
	12	OLS	McKenzie	39.38		12.7%	19.86		4.8%
Economist Models	13	OLS	FaWo	38.06		15.6%	19.92		4.5%
	14	OLS	FaWo + BusScores	38.02		15.7%	19.76		5.3%
	15	LASSO	All	38.39	40.78	17.6%	20.68	20.81	2.6%
Machine Learning	16	SVM	All	39.92	41.80	14.6%	7.22	20.71	92.8%
	17	Boosting	All	32.58	39.02	30.3%	16.41	20.29	31.0%
			Full sample mean	7.02 (N=1,058)			12.09 (N=1,036)		

See notes in Table A10.

Table A14: Prediction (In-sample) Accuracy for Monthly Profits

				Non-winners			Winners		
				MSE		R^2	MSE		R^2
		Model	Predictors	80% sample	5-fold CV	80% sample	80% sample	5-fold CV	80% sample
	1	OLS	Constant	38.77		0.0%	22.11		0.0%
	2	OLS	Application Score	37.02		4.5%	21.79		1.5%
Judges	3	OLS	Business Score	37.02		4.5%	21.79		1.5%
	4	OLS	SubScore	36.79		5.1%	21.63		2.2%
	5	OLS	Gender	36.79		5.1%	21.68		2.0%
	6	OLS	Age	36.61		5.6%	21.83		1.3%
	7	OLS	Necessity Firm	36.92		4.8%	21.85		1.2%
Single Predictor	8	OLS	Grit	36.71		5.3%	21.84		1.2%
	9	OLS	Digit-span Recall	36.74		5.2%	21.84		1.2%
	10	OLS	Raven Test	35.80		7.7%	21.73		1.7%
	11	OLS	Registration time	37.00		4.6%	21.80		1.4%
	12	OLS	McKenzie	34.43		11.2%	20.93		5.3%
Economist Models	13	OLS	FaWo	33.30		14.1%	20.98		5.1%
	14	OLS	FaWo + BusScores	33.29		14.1%	20.90		5.5%
	15	LASSO	All	34.24	35.69	14.2%	21.72	22.11	3.3%
Machine Learning	16	SVM	All	4.27	34.66	92.1%	6.62	22.11	94.6%
	17	Boosting	All	27.97	34.74	30.3%	19.50	21.71	16.8%
			Full sample mean	6.39 (N=1,058)			10.68 (N=1,036)		

See notes in Table 6.

Table A15: Alternative Set of Possible Predictors (Out-of-sample Performance)

Panel A: Original Results

	Survival	Employment				Sales				Profits			
	Non-winners	Non-winners		Winners		Non-winners		Winners		Non-winners		Winners	
Model	Accuracy	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2
Judge Score	58.8%	1.68	1.9%	0.85	9.1%	46.93	0.6%	20.01	0.5%	39.04	1.0%	20.30	0.7%
	[52.3%, 65.3%]	[1.50, 1.86]		[0.61, 1.10]		[43.36, 50.49]		[14.00, 26.02]		[36.41, 41.67]		[15.35, 25.24]	
McKenzie	63.9%	1.59	6.9%	0.91	2.9%	44.22	5.3%	20.61	0.2%	37.99	4.0%	20.90	0.4%
	[57.4%, 70.4%]	[1.40, 1.78]		[0.66, 1.17]		[40.20, 48.24]		[14.47, 26.75]		[34.51, 41.48]		[15.83, 25.98]	
FaWo	67.1%	1.54	10.3%	0.93	2.2%	41.13	10.9%	20.42	0.5%	34.54	11.2%	20.79	0.6%
	[60.7%, 73.6%]	[1.35, 1.72]		[0.66, 1.20]		[36.67, 45.60]		[14.38, 26.47]		[30.75, 38.33]		[15.80, 25.77]	
LASSO	60.2%	1.54	10.6%	0.85	15.1%	40.97	13.0%	20.08	0.5%	35.23	11.1%	20.32	0.9%
	[53.7%, 66.7%]	[1.38, 1.70]		[0.60, 1.09]		[37.78, 44.17]		[14.05, 26.11]		[32.86, 37.60]		[15.30, 25.34]	

Panel B: Exclude Gender

	Survival	Employment				Sales				Profits			
	Non-winners	Non-winners		Winners		Non-winners		Winners		Non-winners		Winners	
Model	Accuracy	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2
McKenzie	64.3%	1.62	5.4%	0.92	2.4%	45.79	3.2%	20.39	0.4%	39.21	2.2%	20.43	1.1%
	[58.1%, 70.6%]	[1.43, 1.81]		[0.66, 1.17]		[41.61, 49.96]		[14.30, 26.49]		[35.56, 42.86]		[15.48, 25.37]	
FaWo	65.7%	1.56	8.9%	0.93	2.1%	42.28	8.8%	20.23	0.7%	35.38	9.3%	20.31	1.4%
	[59.5%, 71.9%]	[1.38, 1.75]		[0.66, 1.20]		[37.71, 46.85]		[14.23, 26.23]		[31.48, 39.28]		[15.43, 25.19]	
LASSO	62.0%	1.55	9.8%	0.85	15.1%	41.16	12.4%	20.08	0.5%	35.23	11.1%	20.32	0.9%
	[55.8%, 68.3%]	[1.39,1.71]		[0.60, 1.09]		[37.96, 44.36]		[14.05, 26.11]		[32.86, 37.60]		[15.30, 25.34]	

Panel C: Exclude Age

	Survival	Employment				Sales				Profits			
	Non-winners	Non-winners		Winners		Non-winners		Winners		Non-winners		Winners	
Model	Accuracy	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2
McKenzie	59.3%	1.63	5.0%	0.93	1.6%	44.78	4.5%	20.54	0.3%	38.41	3.4%	20.94	0.3%
	[52.7%, 65.9%]	[1.44, 1.82]		[0.67, 1.19]		[40.79, 48.77]		[14.43, 26.64]		[34.94, 41.87]		[15.87, 26.02]	
FaWo	65.3%	1.56	9.2%	0.94	1.2%	41.41	10.3%	20.31	0.6%	34.69	10.9%	20.79	0.6%
	[58.9%, 71.7%]	[1.37, 1.74]		[0.67, 1.21]		[36.90, 45.93]		[14.31, 26.32]		[30.90, 38.48]		[15.83, 25.76]	
LASSO	57.4%	1.54	10.6%	0.85	15.1%	40.86	13.4%	20.08	0.5%	35.23	11.1%	20.32	0.9%
	[50.8%, 64.1%]	[1.38, 1.70]		[0.60, 1.09]		[37.66, 44.06]		[14.05, 26.11]		[32.86, 37.59]		[15.30, 25.34]	

Panel D: Exclude Age and Gender

	Survival	Employment				Sales				Profits			
	Non-winners	Non-winners		Winners		Non-winners		Winners		Non-winners		Winners	
Model	Accuracy	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2
McKenzie	60.6%	1.66	3.4%	0.93	1.1%	46.43	2.4%	20.29	0.5%	39.69	1.7%	20.43	1.0%
	[54.1%, 67.2%]	[1.47, 1.86]		[0.67, 1.19]		[42.36, 50.50]		[14.23, 26.34]		[36.12, 43.27]		[15.50, 25.36]	
FaWo	64.8%	1.58	7.7%	0.94	1.1%	42.58	8.2%	20.08	0.9%	35.54	9.0%	20.28	1.5%
	[58.5%, 71.2%]	[1.40, 1.77]		[0.67, 1.22]		[38.00, 47.16]		[14.13, 26.04]		[31.66, 39.43]		[15.43, 25.14]	
LASSO	62.0%	1.55	9.8%	0.85	15.1%	41.12	12.5%	20.08	0.5%	35.23	11.1%	20.32	0.9%
	[55.6%, 68.5%]	[1.39, 1.71]		[0.60, 1.09]		[37.91, 44.34]		[14.05, 26.11]		[32.86, 37.59]		[15.30, 25.34]	

Note: Out-of-sample accuracy is the ratio of true positives and true negatives to all observations, and is computed using the 20% hold-out sample. MSE is the out-of-sample mean-squared error computed using the 20% hold-out sample. The numbers in brackets are bootstrapped 95 percent confidence intervals for hold-out prediction performance. R^2 is the squared correlation between the fitted outcome and actual outcome in the 20% hold-out sample.

Table A16: Include Business Scores (for winners and non-winners) and Total Award Amount (for winners) to Set of Possible Predictors

	Survival	Employment				Sales				Profits			
	Non-winners	Non-winners		Winners		Non-winners		Winners		Non-winners		Winners	
Model	Accuracy	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2
LASSO	61.6%	1.54	10.6%	0.85	14.8%	40.97	13.0%	20.08	0.5%	35.23	11.1%	20.32	0.9%
	[55.1%, 68.0%]	[1.38, 1.70]		[0.60, 1.09]		[37.78, 44.17]		[14.05, 26.11]		[32.86, 37.60]		[15.30, 25.34]	
Business scores selected?	Yes (Total mark, fin sources)	No		No (Only award amount)		No		No		No		No	

Note: Out-of-sample accuracy is the ratio of true positives and true negatives to all observations, and is computed using the 20% hold-out sample. MSE is the out-of-sample mean-squared error computed using the 20% hold-out sample. The numbers in brackets are bootstrapped 95 percent confidence intervals for hold-out prediction performance. R^2 is the squared correlation between the fitted outcome and actual outcome in the 20% hold-out sample.

Table A17: Use only set of inputs from Fafchamps and Woodruff (2017)

	Survival	Employment				Sales				Profits			
	Non-winners	Non-winners		Winners		Non-winners		Winners		Non-winners		Winners	
Model	Accuracy	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2
FaWo	67.1%	1.54	10.3%	0.93	2.2%	41.13	10.9%	20.42	0.5%	34.54	11.2%	20.79	0.6%
	[60.7%, 73.6%]	[1.35, 1.72]		[0.66, 1.20]		[36.67, 45.60]		[14.38, 26.47]		[30.75, 38.33]		[15.80, 25.77]	
Boost	63.9%	1.53	10.3%	0.91	3.8%	40.80	11.4%	20.53	0.2%	35.41	9.2%	21.07	0.1%
	[57.6%, 70.2%]	[1.35, 1.71]		[0.65, 1.16]		[36.62, 44.98]		[14.48, 26.57]		[31.51, 39.32]		[16.05, 26.08]	
Boost (FaWo)	66.7%	1.51	11.5%	0.94	2.0%	40.54	11.9%	20.72	0.1%	34.55	11.1%	20.87	0.1%
	[60.4%, 72.9%]	[1.34, 1.69]		[0.68, 1.21]		[36.58, 44.49]		[14.64, 26.80]		[31.12, 37.97]		[15.80, 25.94]	
SVM	66.2%	1.57	8.4%	0.88	6.2%	43.23	9.8%	19.75	1.6%	33.79	13.1%	20.07	1.3%
	[59.8%, 72.6%]	[1.39, 1.76]		[0.64, 1.12]		[37.88, 48.58]		[13.49, 26.02]		[30.50, 37.08]		[14.85, 25.29]	
SVM (FaWo)	66.2%	1.58	9.5%	0.92	4.1%	44.35	8.0%	21.59	1.3%	36.58	9.4%	22.55	0.8%
	[59.7%, 72.7%]	[1.37, 1.78]		[0.66, 1.19]		[38.54, 50.17]		[14.01, 29.17]		[31.73, 41.42]		[15.53, 29.57]	

Note: Out-of-sample accuracy is the ratio of true positives and true negatives to all observations, and is computed using the 20% hold-out sample. MSE is the out-of-sample mean-squared error computed using the 20% hold-out sample. The numbers in brackets are bootstrapped 95 percent confidence intervals for hold-out prediction performance. R^2 is the squared correlation between the fitted outcome and actual outcome in the 20% hold-out sample. The FaWo, Boost and SVM models are the same as reported in Tables 4-6 and A10. The Boost (FaWo) and SVM (FaWo) models use the same set of inputs of the FaWo model.

Table A18: Ensembles (Out-of-sample Performance)

		ML		ML + Raven		ML + Judges		ML + FaWo	
Outcome	Group	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2
Employment	Non-winners	1.52	11.2%	1.52	11.1%	1.52	11.3%	1.51	11.5%
		[1.34, 1.70]		[1.34, 1.70]		[1.34, 1.70]		[1.33, 1.69]	
Employment	Winners	0.84	12.2%	0.84	12.4%	0.83	13.3%	0.84	12.6%
		[0.60, 1.08]		[0.60, 1.08]		[0.58, 1.07]		[0.60, 1.08]	
Sales	Non-winners	40.62	11.7%	40.62	11.7%	40.55	11.8%	40.37	12.2%
		[36.57, 44.66]		[36.58, 44.67]		[36.51, 44.60]		[36.28, 44.47]	
Sales	Winners	19.77	1.2%	19.78	1.1%	19.74	1.3%	19.81	1.0%
		[13.79, 25.75]		[13.80, 25.77]		[13.78, 25.71]		[13.83, 25.79]	
Profits	Non-winners	34.44	11.3%	34.59	11.0%	34.54	11.1%	34.33	11.6%
		[30.80, 38.08]		[30.95, 38.23]		[30.90, 38.17]		[30.67, 37.99]	
Profits	Winners	20.48	0.2%	20.42	0.3%	20.49	0.2%	20.55	0.2%
		[15.48, 25.48]		[15.44, 25.40]		[15.49, 25.50]		[15.56, 25.54]	

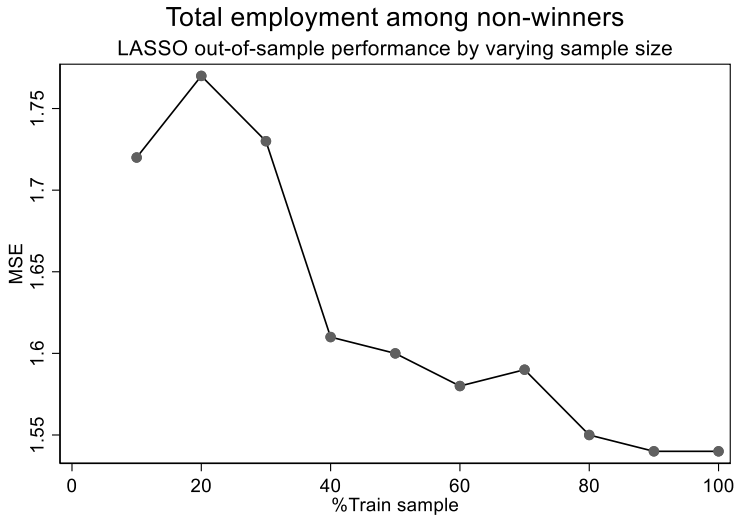
Note: This table presents four ensembles. The first one combines the predictions from the three main ML algorithms (LASSO, SVM, Boosting), the second one combines the three ML algorithms and a simple model using the Raven test scores, the third one combines the three ML algorithms and a model using the judges sub-scores, the fourth one combines the three ML algorithms and the predictions from the Fafchamps and Woodruff (2017) (abbreviated as FaWo) model. Ensembles are fitted using the procedure described in Mullainathan and Spiess (2017). MSE is the out-of-sample mean-squared error computed using the 20% hold-out sample. The numbers in brackets are bootstrapped 95 percent confidence intervals for hold-out prediction performance. R^2 is the squared correlation between the fitted ensemble outcome and actual outcome in the 20% hold-out sample

Table A19: Elastic Net (Out-of-sample Performance)

Outcome	Group	Accuracy	MSE	R^2
Survival	Non-winners	60.6% [54.0%, 67.3%]		
Employment	Non-winners		1.54 [1.38, 1.70]	10.6%
Employment	Winners		0.85 [0.60, 1.09]	15.1%
Sales	Non-winners		40.14 [36.79, 43.49]	14.7%
Sales	Winners		20.40 [14.30, 26.50]	0.4%
Profits	Non-winners		34.74 [32.15, 37.34]	12.1%
Profits	Winners		20.17 [15.23, 25.11]	1.0%

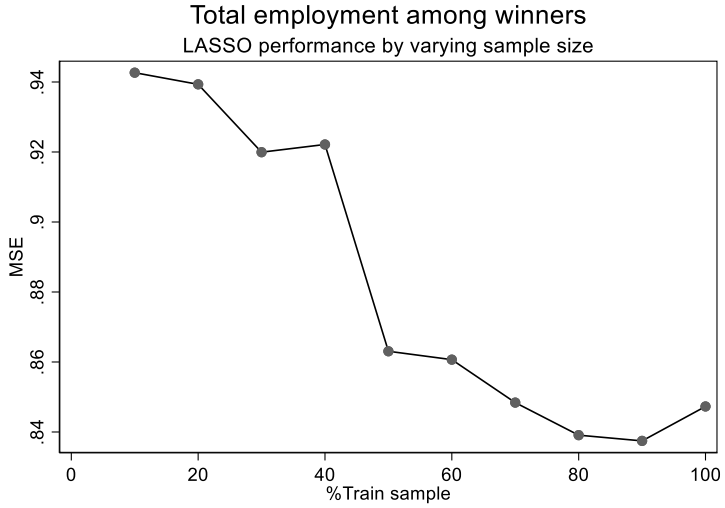
Note: Out-of-sample accuracy is the ratio of true positives and true negatives to all observations, and is computed using the 20% hold-out sample. MSE is the out-of-sample mean-squared error computed using the 20% hold-out sample. The numbers in brackets are bootstrapped 95 percent confidence intervals for hold-out prediction performance. R^2 is the squared correlation between the fitted outcome and actual outcome in the 20% hold-out sample. The set of predictors does not include the initial application scores, the business scores, or the total amount awarded to winners.

Figure A5: LASSO performance by varying sample size training set. Non-winners



Note: this figure plots the out-of-sample performance of LASSO. The outcome variable is total employment among non-winners. The model is estimating using a varying percentage of the training sample. 100% of the training sample represents 80% of the original sample. The out-of-sample MSE is always computed using the original 20% hold-out sample. The out-of-sample performances reported using 100% of the training sample are the same as the one reported in Model 16 Table 5 (Non-winners).

Figure A6: LASSO performance by varying sample size training set. Winners



Note: this figure plots the out-of-sample performance of LASSO. The outcome variable is total employment among winners. The model is estimating using a varying percentage of the training sample. 100% of the training sample represents 80% of the original sample. The out-of-sample MSE is always computed using the original 20% hold-out sample. The out-of-sample performances reported using 100% of the training sample are the same as the one reported in Model 16 Table 5 (Winners).

Table A20: Variables most commonly selected by Boosting

		Total Employment		Sales		Profits	
Predictors	Count	NoWin	Win	NoWin	Win	NoWin	Win
PCA ability (education, raven, digit-span recall)	6	2.02	4.17	1.61	5.89	0.75	5.43
Registration Time	6	3.38	3.48	4.54	2.34	2.92	1.92
# Workers with tertiary education	6	7.76	3.10	5.15	2.53	1.56	8.69
# Workers with secondary education	6	4.81	10.06	8.19	2.24	8.15	8.40
Age	6	0.70	4.01	2.69	8.41	2.82	8.96
# Brothers	5	2.44	0.00	0.80	2.43	1.13	4.96
# Sisters	5	2.34	0.63	1.77	1.03	0.62	0.00
PCA ability (raven, digit-span recall)	5	2.56	3.02	0.95	0.00	1.49	2.47
# Workers with primary education	5	7.96	0.46	1.30	1.72	1.67	0.00
Expected amount of money needed to expand	5	0.83	1.93	0.58	1.29	0.59	0.00
Taxes paid last year	5	0.00	1.38	1.92	10.49	1.07	8.62
Has AC	5	2.34	0.75	0.53	1.13	0.63	0.00
Wealth (household assets) index	5	1.43	0.00	1.60	3.49	1.40	3.47
Self-confidence (first follow-up)	4	7.55	0.94	5.46	0.00	4.79	0.00
Grit: “I become interested in new pursuits every few months” (Very much like me)	4	2.55	0.45	0.65	0.00	1.25	0.00
Raven test	4	2.01	1.59	1.49	0.00	2.06	0.00
Length first answer in application	4	0.89	1.45	2.80	0.00	2.96	0.00
“Which step on the ladder best represents where you personally stand at the present time?”	4	0.71	1.77	1.45	0.00	1.15	0.00
# Years living abroad	4	1.21	2.24	0.00	2.36	0.40	0.00
“Which step best represents where you personally will be on the ladder five years from now?”	4	2.91	0.95	4.43	0.00	5.56	0.00
Self-confidence (baseline)	4	1.72	2.94	2.51	0.00	3.02	0.00
# Workers with secondary education	4	0.00	1.00	0.88	0.00	1.30	4.25
Grit score	4	0.00	0.99	0.78	1.71	1.79	0.00
Gender	4	3.21	0.00	0.49	0.00	0.42	1.75

Notes: NoWin and Win denote Non-winners and Winners respectively. Count denotes the number of times (out of 6) that the variable is chosen by boosting in predicting employment, sales, and profit outcomes. The remaining columns then show the average influence the variable has on the prediction. 182 variables are selected at least once by Boosting.

Table A21: Variables most commonly selected by LASSO

Predictors	Count
Confidence: “Correctly value a business if you were to buy an existing business from someone else” (Very confident)	3
Use expressions “employer of labour” or “employer of labor” in application	3
Raven test: Question 1 Answer 8 (correct)	3
Grit: “I become interested in new pursuits every few months” (Very much like me)	3
Raven test: Question 11 Answer 2 (incorrect)	3
Raven test: Question 11 Answer 1 (correct)	1
PCA ability (education, raven, digit-span recall) missing	2
Number of children	2
Loan from cooperative missing	2
Education level: secondary education or lower	2
Age	2
“Which step best represents where you personally will be on the ladder five years from now?” (8/10)	2
Gender	2
Education level: university degree	2
Business sector: other repair services	2
Confidence: “Sell a new product or service to a new client” (Quite confident)	2
Confidence: “Estimate accurately the costs of a new project” (Very confident)	2
# Workers with secondary education missing	2
Length all answers in application	2

Notes: Count denotes the number of times (out of 6) that the variable is chosen by LASSO in predicting employment, sales, and profit outcomes among winners and non-winners. 32 additional inputs are selected only once.