Limited Partners versus Unlimited Machines;
Artificial Intelligence and the Performance of Private Equity Funds

Borja Fernández Tamayo[1], Reiner Braun[2], Florencio López-de-Silanes*[3],

Ludovic Phalippou[4], and Natalia Sigrist[5]

December 15th, 2023

We assemble a proprietary dataset of 395 private equity (PE) fund prospectuses to analyze fund performance and fundraising success. We analyze both quantitative and qualitative information contained in these documents using econometric methods and machine learning techniques. PE fund performance is unrelated to quantitative information, such as prior performance, and measures of document readability. Measures of fundraising success, in contrast, are correlated to most fund characteristics but are not related to future performance. Meanwhile, machine learning tools can use qualitative information to predict future fund performance: the performance spread between the funds within the top and bottom terciles of predicted probability of success is about 25%. Our findings support the view that in opaque and non-standardized markets, investors fail to incorporate qualitative information in their asset manager selection process, but do incorporate salient quantitative information.

Keywords: Private equity, performance predictability, natural language processing, machine learning

[1] Université Côte d'Azur, SKEMA Business School and Unigestion SA. Email at borja.fernandeztamayo@skema.edu
[2]Technische Universität München School of Management. Email at reiner.braun@tum.de
[3]Université Côte d'Azur, SKEMA Business School and NBER. Corresponding Author. Email at florencio.lopezdesilanes@skema.edu
[4]University of Oxford, Said Business School. Email at ludovic.phalippou@sbs.ox.ac.uk
[5]Unigestion, SA. Email at nsigrist@unigestion.com
*Corresponding author

We thank participants at the International Accounting & Finance Doctoral Symposium (2021), Private Markets Research Conference (2022), Private Equity Research Consortium (2022), SKEMA Business School, HEC Paris, and Benelux CF Network (2023) for their comments and suggestions. We acknowledge and appreciate the data access, cooperation and useful discussions with members of Unigestion, S.A. We are also grateful to SKEMA Business School for providing financial support of the project. An earlier version of the paper circulated after the Private Equity Research Consortium conference in November 2022 under the title "Selecting Private Equity Funds using Machine Learning."

1. Introduction

Private Equity (PE) markets have exploded in the past two decades. PE assets under management have multiplied more than tenfold since 2004 reaching $8 trillion in 2022. PE also represents an increasing proportion of most institutional investor portfolios. These investors dominate the demand side of the market and spend considerable time and resources making PE investment decisions, probably as a result of the large commitment and multi-year nature of such investments (Da Rin and Phalippou (2017)). Albeit its growing importance in financial markets, PE fundraising success and performance remain understudied, possibly due to the scarcity of information.

The abundant literature on mutual fund flows and performance may help understand PE markets. The empirical evidence on mutual fund flows seems consistent with models that predict investors flowing to funds with good past performance and decreasing returns to fund scale (Berk & Green (2004)). One could argue that since a large proportion of investment in public and private markets is made by the same institutions, similar forces may be driving these two markets. Indeed, funds in PE have also been shown to flow to past performance (Chung et al., (2012)) and there are decreasing returns to scale (e.g., Braun et al., (2023), Lopez-de-Silanes et al., (2015))

On the other hand, the pervasiveness of asymmetric information in PE may render the investors' task more difficult. Unlike mutual fund databases, private equity datasets are thin. Moreover, PE fund managers also have considerable freedom to frame their track records at the time of fundraising (Brown et al., (2019); Barber & Yasuda (2017)). In addition, although they provide quantitative and qualitative information to prospective investors at the time of fundraising, its presentation is non-standardized and relies heavily on textual form or conversations. Investors may find it difficult to form accurate expectations on this basis, and even harder to communicate their recommendation other investment committee members (Malenko et al., 2023). Finally, and more broadly, PE investors may face difficulties in incorporating intangible information. Edmans (2022) argues that when information is uncertain, difficult to process or likely to affect discounted future cash-flows in the longer term, investors may fail to incorporate it in their choices (Edmans (2011); Cohen et al., (2013)

In this paper, we collect a unique dataset of close to 400 PE fundraising prospectuses and analyze the impact of quantitative and qualitative information contained in these documents on fundraising success and performance. For the first time in the PE literature, we use econometric methods, Natural Language Processing (NLP) and machine learning algorithms together to analyze the impact of these two kinds of information. Fundraising success, a measure of investor demand, can be used to identify

the information incorporated by LPs. Meanwhile, ultimate fund performance predictability allows us to identify which information is not priced in by investors.

We start by collecting the quantitative information extracted from private placement memorandums (PPMs): past performance, vintage year, fund size, and fund sequence. In addition, we compute two common measures of readability proposed in the literature but never used in the PE context: the number of words in relevant PPM sections and the total number of PPM pages (Loughran and McDonald (2014)). Our first set of econometric tests analyzes the impact of these variables on investors' demand, for which we construct two different measures. The first proxy, which is new, is the time it takes to raise a fund, i.e. from fundraising launch until final closing. The second measure is oversubscription, which is computed as the ratio of the realized fund size at the end of the fundraising period to the fund size targeted by the PE firm at the start of the fundraising effort (Hochberg, Ljungqvist, and Vissing-Jørgensen (2014); Barber and Yasuda (2017); Brown et al. (2019)). Our results are consistent with investors processing quantitative information in their capital allocation decision. PE firm reputation (as proxied by size and number of funds previously raised) as well as past performance are significantly related to fundraising success. We also find that funds with longer prospectuses are those that take longer to fundraise.

We then analyze whether quantitative information predicts future performance (i.e., whether investors really spotted fund manager skills). In line with existing literature (Harris, Jenkinson, and Kaplan (2023)), we find that none of these variables are related to future performance. A new result, however, is that fundraising success is also unrelated to future performance. At first, this might appear surprising as sought-after funds should be more skilled and perform better. However, this set of findings is consistent with the Berk and Green (2004) model where institutional investors incorporate all the readily available information into their demand. This is also consistent with the empirical evidence showing diseconomies of scale (Braun et al., (2023); Lopez-de-Silanes et al., (2015)).

Next, we proceed to the central contribution of our paper: the analysis of qualitative information. Are investors processing qualitative information, given that this information is not readily comparable from one fund to the other, and not easily processed and transmitted? To test for this hypothesis, we combine NLP and machine learning techniques and focus on the strategy section of the prospectuses in which GPs elaborate on their planned activities.[6]

The strategy section of the funds in our sample contains an average of 2,774 words and conveys considerable information about what the fund manager plans to do. The richness of this information

[6] The investment strategy section contrasts with the other sections of the PPM, which lawyers tend to largely copy-paste across PPMs. For details on the role played by lawyers in drafting a PPM please refer to: https://www.lexisnexis.com/lexis-practice-advisor/the-journal/b/lpa/posts/drafting-and-reviewing-the-key-

contrasts with the more limited and generic set of quantitative fund characteristics used before. We apply the Term Frequency - Inverse Document Frequency (TF-IDF) approach, a well-established method NLP. This method produces a score indicating how unique a given term (i.e., feature) is in each document processed. An analysis of terms across time shows that these vectors have been remarkably stable, particularly since 2003. Our data also shows that the investment strategy section of the PPM is not significantly influenced by the lawyers advising the PE firm.[7]

We carry out our analysis using two common machine learning techniques: Lasso Regression, which is a linear approach, with coefficient estimates that can be interpreted in a standard way, and Gradient Boosting, which is a non-linear approach with coefficient estimates that are not readiy innterpretable.[8] The objective of our algorithms is to predict fund outperformance based on the TF-IDF vectors.[9] To measure outperformance, we benchmark each fund's TVPI to the median TVPI of the funds with the same investment strategy and vintage year in the Preqin database. Our outperformance measure is constructed as a binary indicator which takes the value one if the fund outperforms and zero otherwise.

In any PE context, the long holding period and the imperfect intermediary investment valuations represent an important challenge to perform out-of-sample tests. For example, in order to carry out a standard out-of-sample test, we would need to train the algorithm on funds raised before 1990, measure their performance as of 2000, test the algorithm on funds raised during the 2000s, and measure their performance as of 2020. This procedure is generally not possible, given the lack of available PPMs in the 1980s. It may also not be desirable in light of the dramatic change in the industry in the past 40 years, and the waste of potentially valuable information.

Therefore, in order to address the trade-off between statistical power and a look-ahead bias, we conduct two sets of tests. First, we show results when we train the algorithms on funds before raised between 1999 (the start of our sample) and 2013, as well as between 2003 (when PPMs become homogenous) and 2013, and report out-of-sample goodness of fit on the funds raised in 2014-2016, with all fund performance measured as of 2022. The second test is purely out of sample: we train

documentation-for-a-private-equity-fund-and-its-offering. In addition to the mostly legal sections devoted to the offer and its risks, PPMs also contain fund managers' background and selected case studies. Due to the extent of this paper, we analyse these sections in a separate paper.

[7] Law firms in our sample have a PPM similarity of between 30 and 50%, suggesting that the responsibility for what is written in the strategy section of the PPM lies on the PE firm. We also compare the investment strategy sections of two adjacent funds from the same PE firm. We find a similarity of more than 70% for more than half of them.

[8] We also use Random Forest, another common machine learning algorithm, and present the results in the Appendix.

[9] Given the relatively small size of our sample, compared to other machine learning settings, we restrict ourselves to a binary indicator of outperformance.

algorithms on funds raised before 2007, measure their (intermediary) outperformance as of 2013, and use the 2014-2016 funds for the out-of-sample test.[10]

We find that the machine learning algorithms are remarkably effective at predicting fund performance, especially Gradient Boosting (GB). The accuracy of the algorithms improves when we use funds raised after 2003 for training and does not materially deteriorate when we only train on the 2003-2007 funds with performance as of 2013. Remarkably, GB correctly classifies 75% of the outperformers in the top tercile and higher performance in the top tercile does not translate into increased downside risk. Our results are also economically significant: The spread between the overall average TVPI and the TVPI of the top quartile of funds selected by the algorithm is 0.23 points, equivalent to about 4% per annum.

Unlike traditional econometric models, machine learning algorithms do not provide a straightforward interpretation of the marginal influence of features (i.e. independent variables) on outcome variables. However, developments in the explainable AI (XAI) and interpretable ML have started to improve the interpretability of models (e.g., Lundberg and Lee (2017), Ribeiro, Singh, and Guestrin (2016), and Vilone and Lungo (2020) for a review). We use these methods to gain insights into our algorithms, and to quantify the contribution of each feature in predicting fund performance. An analysis looking into the most impactful features predicting performance shows terms such as "operational (and) financial (expertise)," "network relationship," or "investment criteria." This analysis gives us some confidence that our algorithms are picking up economically interpretable constructs.

In the rest of the paper, we perform a series of ancillary tests that put the machine learning results in the context of our previous analysis using traditional econometric methods, and provide support for the robustness of the main findings of the paper. One could argue that the comparison of machine learning-based results using qualitative information and results using quantitative information through linear regression models is not as meaningful because the former models are able to capture non-linear complexities. For this reason, we perform tests using machine learning models trained with standard quantitative fund variables and/or fundraising success variables. Our results show that the predictive power of these models for future fund performance is substantially lower than that of the algorithms using qualitative information. Finally, we also show that the qualitative information exploited by the algorithms does not simply mirror some quantitative fund characteristics. We find

[10] In untabulated results, we find that most of the funds outperforming in their sixth year are still outperforming in their tenth year. For this reason, we believe that using six years does not represent a major loss of statistical power.

that the probabilities of fund success produced by the algorithms using qualitative information as input are uncorrelated with standard quantitative variables, such as reputation or past performance.

Our findings indicate that qualitative information is a valuable tool to learn about fund manager skills and maybe one of the reasons why some LPs are better at this exercise than others (Lerner, Schoar, and Wongsunwai (2007); Sensoy, Wang, and Weisbach (2014). In addition, following the Berk and Green (2004) theory, one interpretation of the results is that when allocating capital, investors do not incorporate available qualitative information but do incorporate the quantitative information provided to them. As a direct consequence, quantitative information is unrelated to future performance, but qualitative information is. As in public markets (see, e.g., Cohen et al., 2013, Edmans, 2011), even the most professional investors find it hard to incorporate qualitative information in their decisions.

Our study provides the first empirical analysis assessing the impact of the readability measures of fund manager disclosures and the application of textual analysis and machine learning in markets characterized by non-standardized disclosure and inherent information asymmetries. Our results contribute to four different strands of the existing literature in finance.

First, we complement papers on return predictability in private markets. As in Harris et al. (2023), we find that buyout fund performance is poorly predicted by traditional quantitative variables, such as the track record available at fundraising time. Our results are also in line with PE papers raising concerns about the reliability of returns reported to investors when raising a new fund (e.g., Barber and Yasuda (2017); Jenkinson et al., (2020)), and with other work arguing that prior returns are not an important issue for sophisticated investors (Brown et al., (2019); Robinson and Sensoy (2013, 2016)). We supplement this literature using a new fundraising success measure and introducing the analysis of qualitative (textual) data. We show that the average investor finds it difficult to extract information about fund manager skills' heterogeneity using traditional quantitative measure, but that analyzing qualitative content can actually help. We therefore offer qualitative information as a potential explanation of why some investors seem to be persistently good at selecting funds (Cavagnaro et al., 2019). Our paper is also connected to Perfetto and Sigrits (2019), Kruglikov and Forthun (2022), and the recent Cavagnaro, Kong and Wang (2023). These studies train machine learning algorithms on quantitative variables computed from Preqin to try to identify fund

outperformance.[11] Our paper shows that qualitative information works much better for prediction than standard quantitative fund variables using traditional econometric methods and ML techniques.

Second, our paper expands the literature on document readability of disclosures in financial markets. Previous papers have examined the association between readability, fund flows, and future firm performance in public markets (Li (2008); Loughran and McDonald (2014); Loughran and McDonald (2016)). To the best of our knowledge, we are the first to empirically analyze the informational value of document readability in private markets. We do not find readability proxies to be significantly correlated with fundraising success or fund performance in private markets. We posit that, as in security issuance (La Porta et al., (2006)), one potential explanation behind these results may be the lack of disclosure standardization.

Third, our paper expands the nascent and rapidly growing literature using textual analysis in finance which has started to exploit public firms' 10K filings and earnings call transcripts. Hoberg and Phillips (2010) are attributed to be the first to use textual analysis to cluster firms according to product markets. Since then, these methods have been used for diverse objectives including categorizing corporate goals using letters to shareholders (Rajan et al., (2022)) and identifying risk factors raised by firms in their annual reports (Lopez-Lira (2023)).[12] In the area of private markets, the only other paper exploiting textual data that we are aware of is Biesinger et al. (2021), who apply this method to investigate value creation. We contribute to this literature showing the potential to extract meaningful patterns from textual data in private market disclosures. We find such qualitative information could be valuable in assessing PE manager skills beyond quantitative measures studied in previous papers (e.g., Kaplan and Schoar (2005), Braun, Jenkinson, and Stoff (2017)).

Fourth, our paper is connected to the rising literature applying machine learning techniques to identify human biases and improving selection in both asset pricing (e.g., Bianchi et al. (2021); Ke, Kelly, and Xiu (2019); Gu et al. (2020)) and corporate finance (Li et al. (2021), Bubb and Catan (2020), Erel et al. (2021)). Erel et al. (2021), for example, use algorithms to predict director performance and also show that firms are more likely to nominate male candidates as directors than what a trained machine learning model would. Davenport (2022) and Lyonnet and Stern (2022) have used a similar method to compare investor choices in venture capital to algorithms' predictions. The

[11] These studies only use fund characteristics (i.e., quantitative information) as input, and they do not conduct (strict) out-of-sample tests as we do here. In addition, they use IRR to measure fund performance, which may be problematic if IRR is influenced by fund manager characteristics. For example, a GP that always uses a lot of subscription lines, will have high IRRs across its successive funds. A final difference is that since our algorithms do not rely on fund manager track records, they can also be used to predict first-time fund outperformance.

[12] For an extensive overview on textual analysis in several topics in accounting and finance see Loughran and McDonald (2016) and (2020).

thrust of our findings suggests that investors in private markets might be biased towards more established PE firms.

We organize the rest of the paper as follows. After this introduction, Section 2 presents our new dataset of 395 PE funds. In Section 3, we analyze the role of quantitative information and document readability in explaining fundraising success using traditional econometric methods. We also analyze whether a more successful fundraising campaign helps predict ultimate fund success in terms of returns. The second part of Section 3 expands the analysis presenting the main results of the paper including the study of qualitative information using textual analysis and machine learning techniques to predict PE fund success. In the last part of Section 3, we also provide some economic interpretation of the algorithms. In Section 4, we carry out robustness tests addressing potential biases and alternative explanations, and well as ancillary tests connecting quantitative and qualitative information. Section 5 concludes.

2. Data

2.1. Data collection

When raising capital, PE firms start by sending a Private Placement Memorandum (PPM) to potential investors. PPMs are confidential, legally-binding documents that provide a wealth of information including key details about the investment opportunity. Although there is no explicit or implicit industry standard for PE fund disclosures, most of the PPM content is similar across funds. The typical PPM contains a summary of the terms, a discussion of the risks in private equity investments, legal disclaimers, manager biographies, and a discussion of selected past investment. From PPMs, one can extract the standard fund characteristics used in the literature, such as past performance, the number of funds the PE firm has raised in the past, and the target amount to be raised. Importantly for this paper, PPMs also contain a section where GPs describe their investment strategy.

We source PPMs from a large global institutional investor based in Europe who is mostly focused on growth capital and leveraged buy-out funds, and slightly more focused on European funds. The proprietary database we assembled using the investor's paper and electronic archives, consists of 951 PPMs received between 1999 and 2020. Table 1 shows how we arrive at our final sample of 395 funds. Since we use performance as of June 2022, we only keep funds raised by 2016. We also require

that the fund has a minimum size, is present in Preqin dataset, and remove funds that are focusing on venture capital and real assets.[13]

Table 1 also provides statistics on the number of funds, the average and median fund size for our sample and two other samples, which we call the "Preqin Summary Fund Sample (SFS)" and the "Preqin Cash Flow Sample (CFS)," as we apply the various filters to arrive to the final sample of funds we use in our analysis. SFS is a dataset containing (nearly all) key summary statistics for a large set of funds (fund size, geographic focus, investment type, latest reported performance summary measure). CFS is a smaller dataset containing funds for which Preqin has the complete set of cash flows in and out of the funds and the time-series of their Net Asset Values. The table shows that our final sample is comparable to that of the "Preqin Cash Flow Sample" in terms of fund size.

<Insert Table 1 here>

Table 2 shows the various sources of fund performance data for our sample. Our data provider invested in 100 of the 395 funds in our final sample. A total of 34 of the remaining 295 funds are in the Preqin CFS dataset. Hence, we have the full time series of cash flows and NAVs for 134 funds. For the rest of the funds, we only have a performance summary statistics. For 61 of these funds, the data comes from internal records from our data provider (e.g., PPMs of subsequent funds) while for the remaining 200 funds, the data comes from the Preqin SFS dataset.

Since we do not have the time-series of cash flows for all the funds, the main performance measure we use is the Total Value to Paid-in (TVPI) net of fees. TVPI is the ratio of the sum of all capital distributions plus the last reported NAV over the total amount of capital called. Note that 'capital called' includes any called management fees, but NAV is usually gross of unrealized carried interest. Note that as returns are reported in different currencies, it is important to carefully convert them all into a common currency. We choose the Euro because most of the funds in our sample are in this currency. Appendix Table A2 provides the details of the conversion procedure.

Perhaps surprisingly, in PPMs, TVPI of preceding funds is rarely reported. Instead, it is a gross of fees measure that is reported called Multiple of Invested Capital (MOIC). MOIC is the ratio of the sum of all capital distributions plus the last reported NAV over the total amount of capital *invested* (not capital *called*). In practice, MOIC and TVPI are highly correlated and relatively close to one

[13] Preqin is a commercial dataset that is widely used by academics (e.g., Cavagnaro et al., 2019, Gupta and van Nieuwerburgh, 2021, Stafford, 2022). Using the Preqin classification, we exclude funds belonging to one of the following categories: Natural Resources, Special Situations, Secondaries, Distressed Debt, Co-Investment, Mezzanine, Infrastructure, Secondaries, Venture Debt, Fund of Funds, Real Estate, and Venture Capital.

another in terms of magnitude because most of the management fees are not called.[14] We also report the Internal Rate of Return (IRR) because we could use it for robustness. Finally, we benchmark each fund TVPI: we identify the set of funds in the Preqin SFS dataset with the same vintage year, and same investment type, and take the median TVPI of that set of funds as the benchmark. Excess TVPI is equal to a given fund's TVPI minus its benchmark.

<Insert Table 2>

2.2. Fund characteristics

Panel A of Table 3 shows descriptive statistics of fund characteristics, as well as measures of fundraising success and PPM readability. The panel starts with fund performance measures. The average TVPI is 1.8x and the average IRR is 14% for the funds in our final sample. These figures are comparable to those in other datasets covering a similar time window (e.g., Cavagnaro et al., 2019). The fact that about half (52%) of the 395 funds in our sample outperform their benchmark provides further confidence on the representativeness of our data.

Panel A also provides summary statistics for different measures of fundraising success. Our main measure is the duration of the fundraising period, which is novel in the literature. We compute this variable as the number of months between the date when the PE firm starts marketing the fund and the final closing date of the fundraising campaign. The median fund in our sample spends one year raising its capital. We observe substantial differences across funds: the fund in the 75th percentile spent one and a half year (19 months), while the fund in the 25th percentile raised its capital in half a year (6 months).

Another measure of fundraising success, and one that has been used in the literature, is the oversubscription ratio, i.e., the ratio of final to target fund size (see e.g., Hochberg, Ljungqvist, and Vissing-Jørgensen, 2014). The oversubscription ratio for our sample of funds has a mean (median) value of 1.05 (1.07). This means that the average fund in our sample is oversubscribed by 5%. This value is relatively close to the average oversubscription documented in Hochberg, Ljungqvist, and Vissing-Jørgensen (2014) for an earlier time period. The standard deviation of 0.31 of this measure is indicative of significant variation in oversubscription ratios.

The following rows of Panel A provide descriptive statistics on other fund characteristics. The average fund size in our sample is €1 billion, and 25% of our funds raised more than €815 million.

[14] Management fees are offset against the fees taken directly out of the portfolio companies (Phalippou, Rauh and Umber, 2018).

We have 75 first-time funds in our sample, and the median fund in our data is the third fund raised within the same investment type by a given PE firm.

A large body of the literature has studied the relationship between fundraising success and performance reported at the time of fundraising (e.g., Hochberg, Ljungqvist, Vissing-Jørgensen (2014); Barber and Yasuda (2017); Brown et al. (2019); Phalippou (2010)). A MOIC is provided in 318 of the 320 PPMs of a second (or higher) fund generation. On average MOIC is 1.6x, and the interquartile range is 1.2x to 1.9x. Note that MOIC is lower than TVPI despite it being more 'gross of fees' than TVPI. One reason is that MOIC usually increases significantly during the first five years of a fund's life, and the preceding fund is, on average, in its third year at the start of fundraising.

Finally, Panel A shows statistics on PPM readability measures. Whilst many studies rely on lexicas to determine complexity of language, we follow Loughran and McDonald (2014) who criticize lexica-based approaches based on the argument that the complexity of words may be closely related to the complexity of the underlying business rather than to the actual readability of the document itself. For this reason, they use the size of 10-K filings and show that larger files are associated with high return volatility, earnings forecast errors, and earnings forecast dispersions. Although our setting is somewhat different (i.e., unregulated and non-standardized disclosure),[15] we follow Loughran and McDonald and compute two simple readability proxies: the number of words in the PPM's strategy section and the number of pages of the PPM. The traditional assumption in this field is that a larger number of words or pages translates into lower readability. The average PPM in our sample is 84 pages and about forty thousand words long, and the average strategy section contains 2774 words.

Panel B of Table 3 shows the correlation matrix between the eleven fund characteristics shown in Panel A. Several novel and interesting results are worth noting. First, there is a high correlation among the four proxies for fund performance, the two proxies for fundraising success, and the two proxies for experience (size and sequence). Of particular interest is the fact that the proxies for fundraising success are highly correlated with each other. This fact makes us more confident about the validity of our measure to proxy for the unobservable investor demand for a fund at the time of fundraising. Similarly, there is a positive correlation between the number of pages of the PPM and the number of words in the strategy section, which is consistent with the view that document length is a proxy for a common characteristic, namely, readability.

The number of pages of the PPM is positively correlated with fund size and to a lesser extent with fund sequence. Hence, more established funds tend to submit longer documents. The number of words

[15] There is no standard document or regulating organization (like the SEC in public markets) publishing such documents in private markets. In fact, different PPMs are compiled (and provided) using different word processing software.

in the strategy section is also correlated with these two fund characteristics but much less so. Longer PPMs tend to have longer strategy sections as well. Interestingly, both the number of words and the length of the PPM are negatively correlated with past performance: High (past) performers write shorter documents. Maybe people with poorer track records have more explaining to do or compensate with more selected case studies, longer biographies etc. We also observe that future performance, in contrast, is positively related to the length of the strategy section, but negatively related to the length of the PPM. These correlations are small though. It is also interesting to note that funds with shorter documents spend less time fundraising.

<Insert Table 3>

2.3 Numerical representation of qualitative information

A key goal of our study is to numerically represent the text in the investment strategy section of the PPM so that it can be used to predict performance. To do so, we use the Term Frequency - Inverse Document Frequency (TF-IDF) approach, which is an established method in computational linguistics. The idea is that terms that appear in most documents are less likely to help discriminating between funds. For this reason, the TF-IDF measures the originality of a term; it compares the number of times a term appears in a document with the number of documents the term appears in. The method produces a score indicating how characteristic a given term (or feature) is in each document processed. Each PPM in our sample is represented by a vector of TF-IDF scores.

To implement the TF-IDF method in the strategy section of the PPM, we carry out the following steps. We first *stem* words in the text corpse. Second, as single words are usually not meaningful, we keep only two or three adjacent words (respectively called bigrams or trigrams). There are 37,001 unique bigrams and trigrams (i.e., terms) across the 395 strategy sections in our sample, and about half of these terms appear in more than one document. Third, we compute the scaled TF-IDF of a term *i* in document *j* as proposed by Pedregosa et al. (2011), which is a slightly modified version of the original measure proposed by Salton and Buckley (1988):

$$TFIDF\ (i,j) = \frac{TF\ (i,j).(\ln(N) - \ln\ (N_i))}{\sqrt{(\sum(TF(i,j).(\ln(N) - \ln\ (N_i))^2}}$$

That is, the frequency of term *i* in document *j* ($TF(i,j)$) is weighted by the ratio of the total number of documents in the corpse (*N*, here 395 PPMs) to the number of documents containing the term i at least once (N_i). This means that the term frequency in each document is penalized by its frequency across documents. A high TF-IDF score thus indicates that a term is particularly characteristic for a given document. For example, if a term appears twice in the focal document, and is not used in any

of the other 394 documents, its TF-IDF would be equal to 0.5. A term that occurs 1,000 times but appears in all of the 395 documents would have a score of close to zero.

Table 4 shows the most common stemmed bigrams (Panel A) and trigrams (Panel B). This list of terms suggests that our approach picks up concepts that are meaningful in the PE context. We see terms associated with deal sourcing (e.g., *investment opportunity, proprietary deal flow, investment criteria*), value creation (e.g., *value creation, management team, buy and build strategy*), and decision making (e.g., *investment process, due diligence*). *Portfolio company* has the highest average TF-IDF score among frequently used terms: it is mentioned 4,065 times in total and at least once in 92% of the PPMs. Overall, the high standard deviation of TF-IDF scores across PPMs for each term indicates that the terms used differ significantly across documents.

<Insert Table 4>

Investment strategies may have changed over the two decades we span. To analyze this possibility, we compute the average of all TF-IDF scores within a vintage year and calculate the cosine similarities between all annual vectors.[16] Table 5 shows that the wording used in the strategy section of the PPMs has remained remarkably stable since 2003, but do seem to differ before that vintage year. The cosine similarities among PPMs after 2003 have values above 0.6. After 2006, values are above 0.7, and most of them have values above 0.8. We believe that the stability of wording over time alleviates concerns regarding our approach to train algorithms using a body of historic PPMs to predict future fund success. However, using PPMs written in 2003 or earlier probably introduces additional noise.

<Insert Table 5>

A potential concern is that the strategy section of the PPM is written by the lawyers advising the PE firm rather than by the fund managers themselves. Figure 1 reassures us that this is not the case. In Panel A, we compute the similarity of wording across all PPMs advised by a given law firm. The majority of law firms have a PPM similarity between 30 and 50%. In Panel B, we plot the distribution of cosine similarities between two adjacent funds from the same PE firm. For more than half of them, we find a similarity of more than 70%. In our view, this empirical pattern confirms that the responsibility for what is written in the strategy section lies with the PE firm.

<Insert Figure 1>

[16] To avoid that extremely uncommon or extremely common bigrams and trigrams influence the comparison, we filter out bigrams and trigrams that appear in less than 1% of all PPM Strategy sections (i.e., in 4 PPMs) and those that appears in more than 99% of all PPM Strategy sections (i.e., in 391 PPM).

3. Empirical results

3.1 Fundraising success and fund characteristics

In this section, we analyze whether investor demand is related to quantitative fund characteristics. Table 6 presents OLS regression results where the dependent variable is the natural logarithm of the number of months a PE firm takes to raise a given fund. We posit that funds experiencing high demand are able to raise capital more rapidly. All specifications control for investment type, targeted region, and vintage year.

There are three main findings that emerge from Table 6. First, fundraising success is positively correlated with firm reputation as proxied by the natural logarithms of fund sequence and fund size.[17] In terms of fund size, for the interquartile range (€177 vs. €815 million) the coefficient from the first specification in Panel A translates into a 12 percent shorter fundraising period. The strength of the effect for the interquartile range of fund sequence is a bit smaller at 9 percent (2nd vs. 4th fund generation).

The second finding that emerges from the table is that readability measures are *negatively* related to fundraising success, but weakly so. Increasing the total number of PPM pages from 62 to 101 pages or the number of words in the strategy section of the PPM from 1,829 to 3,724 words, both equivalent to the inter quartile range across funds, increases fundraising time by 21% and 8% respectively. These are new and interesting results in light of the literature. It has been argued that in standardized public markets, complex language may be used to intentionally increase investors' information processing cost (Bloomfield, 2002; Hoberg and Lewis, 2017; Brown, Crowley, and Elliott, 2020). We note that this effect is not present if we use oversubscription as the dependent variable. Hence, longer documents do not predict whether a fund will attract more capital than it has targeted but is closely linked to how long it takes to raise the targeted amount.[18]

The third finding pertains to the impact of past performance on fundraising success. The interim performance of the GP's previous fund at the time of fundraising is positively associated with fundraising success. All else being equal, at the time of fundraising, a fund manager with a previous-fund MOIC of 1.9x closes a fund 6% faster than a competitor manager with a previous-fund MOIC

[17] Since one could expect a decreasing marginal effect of fund sequence and fund size on fundraising success, we use the natural logarithm of these two measures in all regression specifications.

[18] Our somewhat weaker results might result from the non-standardized nature of disclosure in private markets. Additionally, it is also possible that the nature of the information analysis process carried out by institutional investors in private markets is also important. These investors typically carry out a thorough analysis of only a few funds per year. To the extent that they engage in substantial screening for each fund they analyze, the additional costs caused by low readability may not end up amounting to much for fund manager selection decisions.

of 1.2x. The difference of 0.7x is equivalent to the inter quartile range across the 318 funds in our sample with a predecessor fund.[19] Results are similar in Panel B but with reverse signs since oversubscription is negatively correlated to the time it takes to raise a fund. We interpret this result ad evidence that investors process the manager's past performance when making allocation decisions.

Overall, the results in Table 6 show a consistent pattern linking standard fund characteristics to fundraising success. We interpret these results as suggesting that the average investor takes into account quantitative information provided in the PPM in her fund manager selection decision.

< Insert Table 6 >

3.2 Fund Performance and fund characteristics

Building on the financial intermediation theory of Berk and Green (2004), we conjecture that the strength of their capital demand should be related to past performance, but not to fund future performance. However, private markets are characterized by non-standardized disclosures and significant information asymmetries between managers and investors which may interfere in the capital allocation process (Robinson and Sensoy, 2013). In addition, PE firms may be voluntarily capping their fund size (e.g., Hochberg, Ljungqvist, and Vissing-Jørgensen, 2014). These facts may therefore impact the correlation between the fund characteristics that are related to fundraising success and subsequent fund performance.

Table 7 presents the results of such an analysis. The dependent variable in all specifications is the fund TVPI and the set of regressors includes the fund characteristics used in Table 6, as well as our fundraising success variables which proxy variables for the strength of demand. Results in this table show that the fund characteristics that are related to fundraising success are not related to future fund performance. An interesting exception is that the number of pages of the PPM is related to future performance: the longer the PPM the worse the returns (equivalent to a -0.13 lower TVPI for the interquartile range change of 39 PPM pages). In addition, Table 7 shows that the speed of fundraising and the fund oversubscription ratio are also unrelated to future performance.[20] Findings are similar whether we use TVPI as a measure of fund performance (Panel A) or TVPI over its benchmark (excess TVPI; Panel B). These results are therefore very much in line with Berk and Green (2004) theory.

[19] In un-tabulated results, we find that fund managers with both previous funds outperforming their benchmark raise capital faster. The sample is smaller (243 funds) and the effect is significant at the 10% level test.

[20] In unreported tests, we report the regressions shown in Table 5 using IRR as alternative measures of fund performance. The results are similar. Note that the finding that the performance at the time of fundraising is unrelated to final fund performance may be surprising at first sight. However, the same result has been found in the literature (e.g., Harris, Jenkinson, Kaplan, and Stucke (2023)).

< Insert Table 7 >

3.3 Training Machine Learning Algorithms

If standard fund characteristics and demand strength turn out to be poor predictors of future fund performance, the next logical question is if other types of information, and in particular qualitative information, are related to future performance. In the next section, we explore the potential of natural language processing techniques combined with machine learning algorithms to process the *qualitative content* of investment strategy descriptions to predict fund performance.

Our setting is characterized by a combination of a large number of terms and few observations of fund performance. For this reason, we use a machine learning approach and apply two of the most traditional techniques in this field: Lasso regression and Gradient Boosting. These techniques differ from traditional regression methods in several respects. While Lasso helps mitigating overfitting and multicollinearity in linear regressions, Gradient Boosting has the ability to capture non-linearities in the data.

We follow the standard design of a machine learning exercise. We use about 80% of the sample (323 funds raised *before* the end of 2013) to train the algorithms and apply five-times repeated ten-fold cross-validation as a re-sampling procedure to estimate the parameters of the different algorithms.[21] The models so obtained are then applied to the remaining sample, that is those funds raised between 2014 and 2016 (the test sample).

Our prediction variable is a dummy variable that takes the value one if a fund outperforms the median fund from the same investment type and vintage year in the Preqin database, and zero otherwise. For each fund in the test sample, our method produces an outperformance probability, which we call "predicted probability" and a fund is said to be *predicted* to outperform if the predicted probability is greater than 50%.

We assess the algorithms' goodness of fit using two standard measures in the machine learning literature. The first measure is Balanced Accuracy, which is the average of the ratio of true positives

[21] Cross-validation is the most widely used method for estimating prediction error. The advantage of using repeated cross-validation, instead of only cross-validation, is that it reduces the variance of the cross-validation estimator. Since the size of our training sample is relatively small for prediction purposes, we set K=10 (Hastie, Tibshirani, and Friedman, 2009). Appendix Figure A1 depicts the implementation process of cross-validation for an example of cross-validation with K=5 (i.e., five-fold cross validation).

(TP) over the sum of true positives and false positives (FP), and the ratio of true negatives (TN) over the sum of true negatives and false negatives (FN):

$$Balanced\ Accuracy = \left[\frac{TP}{TP + FP} + \frac{TN}{TN + FN}\right]/2$$

The second standard metric we use in this field is the area under the receiver operating characteristic curve (ROC-AUC). A model whose predictions are all wrong has a ROC-AUC of 0, while a model whose predictions are all correct has a ROC-AUC of 1. The algorithm has a higher predictive power than pure randomness when ROC-AUC is above 0.5.

3.4 Goodness-of-fit

Tables 8 and 9 present the core results of the paper. In Panel A, we train the algorithms using the funds whose vintage is from 1999 to 2013 and the out-of-sample test is carried out on fund of vintage years 2014 to 2016. Since the wording of the PPMs is more similar from 2003 onwards (section 2.3), Panel B carries out the same exercise but changing the training sample to funds whose vintage year is between 2003 and 2013. For each of the two panels of the table, we show the ROC-AUC and the Balanced Accuracy, both for the training sample (in-sample) and for the test sample (Pseudo out-of-sample) using the two different algorithms of Lasso and Gradient Boosting.

In Panel A, all algorithms perform well as shown by the fact that the goodness of fit is always above 0.5. Lasso has the lowest goodness of fit in-sample, but the highest out-of-sample goodness of fit. Results are somewhat different in Panel B. In line with the observation of more similar language starting in 2003, goodness of fit is higher across the board. In-sample fit is similar across the two algorithms. Meanwhile, out of sample, Gradient Boosting has the highest ROC-AUC and balanced accuracy.

A key difference between Lasso and Gradient Boosting is that the latter allows for non-linear combination of predictors. Although the simple linear algorithm (Lasso) performs reasonably well already, the results in Table 8 show that there is value added by allowing for non-linearities.

< Insert Table 8 >

Figure 2 shows the Receiving Operating Characteristic curve (ROC) using the Gradient Boosting algorithm. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings for the 2012-2014 test sample. The blue line represents the algorithm discrimination ability for different thresholds. The diagonal red line shows the results of random

guesses for different thresholds. Comparing the two lines shows that with the exception of a small range of high thresholds, Gradient Boosting has a higher TPR than FPR across thresholds.

In Table 9, we try to obtain more intuitive measures of accuracy. We split the 72 test-sample funds into terciles based on the predicted probability of outperformance by the Lasso (Panel A), and the Gradient Boosting (Panel B) algorithms. There are three interesting findings that emerge from this table. First, both Lasso and Gradient Boosting attribute very low (high) probabilities of outperformance to the bottom (top) tercile of funds. The mean probability of outperformance predicted by Lasso (Gradient Boosting) is 0.08 (0.05) for the bottom tercile and 0.93 (1.00).for the top-tercile of funds.

Second, column 4 shows that the two algorithms correctly predict a very high percentage of outperformers (underperformers) in the top (bottom) terciles. The Gradient Boosting algorithm, for example, correctly classifies 75% of the outperformers in the top tercile and 67% of the underperformers in the bottom tercile. The classification of funds using the Lasso algorithm is slightly less accurate than the one obtained with Gradient Boosting, showing once again that non-linearities play an important role.

Third, funds predicted to outperform do not exhibit higher downside risk. For example, the last column of the TVPI results for Gradient Boosting (Panel B) shows that only 4% of the funds with the highest probability of outperforming (top tercile) have lost money. In contrast, 42% of the funds with the lowest probability of outperforming (bottom tercile) have lost money. The standard deviation of TVPI within each tercile, which is another measure of risk, is nearly identical across terciles: 0.40, 0.46, and 0.52 respectively. Although risk is difficult to measure, particularly in private equity, these results indicate that the funds predicted to outperform are unlikely to be the riskiest ones.

It is worth noting that we trained the algorithms to maximize the AUC, which means that they are trained to maximize their ability to rank and order funds based on their predicted probabilities, and not to predict the magnitude of the outperformance. A fund with a TVPI 0.1x above its benchmark has the same label (i.e., outperformer) as a fund that outperforms its benchmark by 1.5x, for example. As a result, algorithms may correctly classify funds but that may not lead to a significant average outperformance. The last set of columns in Table 9 show that the top-tercile funds have a significantly higher average TVPI than the funds ranked in the bottom tercile. Gradient Boosting, for example, generates a mean TVPI 1.66 for the bottom-tercile funds compared a mean TVPI of 2.07 for the top-tercile funds. Results are siimilar with Lasso and when using Excess TVPI.

<Insert Table 9>

Figure 3 provides further illustration of the economic significance of our results. The funds used for this graph are the 72 test-sample funds raised between 2014 and 2016. The figure shows the average performance of sub-samples of funds sorted according to three different selection criteria. The three lines plot the average TVPI of sub-samples of funds sorted by: the predicted probability of outperformance generated by Gradient Boosting (solid line); the inverse of the number of months to raise the fund (dotted line); and preceeding fund performance (dashed line). The x-axis shows the fraction of funds eliminated from each sub-sample of funds: it starts at 0% (i.e., all funds are included) and goes up to 75% (i.e., only the top quartile of funds in terms of each of the three selection criteria). The y-axis shows the average TVPI for each sub-sample of funds. To generate the solid line, for example, we remove the fund with the lowest predicted probability of outperformance generated by Gradient Boosting and calculate the average performance of the remaining funds in the portfolio. We repeat this procedure fund by fund as we move to the right of the graph, hence moving up the rank of predicted probability of outperformance.[22]

Using the predicted probabilities of outperformance of the Gradient Boosting algorithm to sort funds, the solid line shows that at 0% on the x-axis, the unconditional average TVPI is 1.86. As we move further to the right in the graph, we take out funds from the calculation and see that the average TVPI increases steadily. When all the bottom-quartile funds in terms of predicted probability of outperformance are taken out (i.e., at 25% on the x-axis), the average TVPI reaches 1.95. When the bottom half of the funds is taken out, average TVPI is at 1.98, and if we only keep the top quartile of funds, average TVPI reaches nearly 2.09. Hence, the spread between the overall average TVPI and the TVPI of the top quartile of funds selected by the algorithm is 0.23. Since PE investments are held 5 years on average, this spread represents close to a 4% annual excess rate of return.

The dotted line in the figure shows the average TVPI of fund portfolios obtained with the same procedure outlined above, but sorting funds by the inverse of the number of months it took to fundraise (i.e., the top quartile is comprised of the funds that were raised the fastest), while the dashed line shows average TVPI when sorting by preceeding fund performance. Had we selected the top quartile funds according to fundraising speed or preceeding fund performance, the average TVPI would have been 1.95 and 1.72, respectively. These numbers are much lower than the 2.09 obtained

[22] Following this procedure, the y-axis coordinate of the point whose x-axis coordinate is at 0% corresponds to the average performance of all the funds in the sample of the 72 funds raised between 2014 and 2016. The y-axis value of the point whose x-axis coordinate is at 25% (50%) corresponds to the average performance of the remaining funds after elimination of the bottom quartile (half) of funds in terms of predicted probability of outperformance. The y-axis value of the point whose x-axis coordinates is at 75% corresponds to the average performance of top-quartile funds sorted according to the predicted probability of outperfromance.

with our Gradient Boosting algorithm. Funds that are typically called "top-quartile" are those funds with the best preceeding performance. If we had followed this criterion as the investment guide, the selection would have delivered a TVPI below the unconditional average, consistent with the regression results shown in previous sections.

<Insert Figure 3>

3.5 Isolating the most important terms

As in most papers using machine learning techniques, we do not focus on specific terms because what is picked up by the algorithms are non-linear combinations of terms. Yet, below we follow some of the techniques used in other papers (e.g., Lundberg and Lee (2017), Ribeiro, Singh, and Guestrin (2016), and Vilone and Lungo (2020)) to get a general sense of the terms that matter most, without drawing definitive conclusions.

Figure 4 shows the terms with the highest Shapley Additive exPlanations (SHAP) values when we apply the Gradient Boosting algorithm using the 2003-2013 funds as the training set to predict outperformance. The method of analysis in the figure was developed by Lundberg and Lee (2017) and is frequently used in the machine learning literature. An intuitive interpretation of a SHAP value is the difference between the predicted outcome when the term is included and when it is not.

Panel A of Figure 4 displays the most important terms for predicting outperformance in our sample. The figure provides three distinct pieces of information concerning the terms. First, terms are ranked from top to bottom by their mean absolute SHAP value (i.e., their average importance across all PPMs). Second, the horizontal position of a point represents the SHAP value, i.e. the effect of an term on predicted outperformance for each PPM in our sample. Points to the right have a strong positive impact, while those to the left have a strong negative effect. SHAP values are vertically stacked for cases of high-density areas. Finally, the red and blue colors in this panel correspond to the TF-IDF values of each term and PPM: red (blue) indicates a high (low) TF-IDF raw value.

Panel A helps us visualize the complex non-linear nature of machine learning approaches. The effect of an term on predicted outperformance varies across PPMs in strength as well as in direction. For this reason, we cannot simply state whether a fund will perform well (or poorly) because its PPM strategy section includes frequently a specific term. In other words, the pleasant interpretation of a linear regression is not applicable to these models. For example, many low TF-IDF values (blue) of the term "operational-and-financial" have a moderate negative effect (left-hand side) on predicted outperformance. However, in many PPMs a low TF-IDF is associated with a positive effect (right-hand side) on predicted performance.

A simpler graph is presented in Panel B of Table 4. This figure is similar to that in Erel et al, (2021); it lists the terms with the highest median absolute SHAP values. The terms are the same as in Panel A, but they are ranked according to their median absolute SHAP value. We use median values (instead of mean values) because several terms have extreme values (as Panel A shows). The colors in Panel B have a different meaning than in Panel A: terms in blue (red) have a positive (negative) median value. These values correspond to their median position on the x-axis in Panel A.

< Insert Figure 4 >

Overall, the terms found to be important in predicting outperformance are economically meaningful concepts and seem consistent with the existing literature. The creation of shareholder value by resolving agency problems, for example, has been central to explaining the nature of LBOs (Jensen, 1989). Similarly, empirical (Acharya et al., 2013; Guo et al. 2011) and theoretical (Malenko and Malenko, 2015) studies support the view that operational improvements, governance or financial engineering are positively correlated with success.

In one of the most comprehensive analysis of the value creation activities in private equity, Biesinger, Bircan and Ljungqvist (2021) group activities into five categories according to their frequency: operational improvements, top-line growth, governance engineering, financial engineering, and cash management (mostly reduction in working capital requirements). The first three activities are among the top terms in Panel B of Figure 4, and they all have a positive effect on predicted probability of outperformance. The figure also illustrates that the term “operational and financial” has the highest (and positive) median value effect. This is directly related to the aforementioned measure of operational value creation and financial engineering. Biesinger, Bircan and Ljungqvist (2021) underline that although most PE firms plan to create value in their portfolio companies, only the implementation of associated measures is positively correlated with returns in their sample.

Other terms shown in Panel B may be associated with the manner in which value creation is carried out. For example, we find that company development (“development of the portfolio”, positive median SHAP value) in cooperation with the management team (“relationship with the management team”, positive) are some of the most important terms. Apparently, “team experience” (positive) is beneficial for both management teams and fund managers. PE firms often involve “senior executives” (positive) in the development of value creation plans. More specifically, such plans may be based on exploiting "economies of scale" (positive) or pursuing a strategy of (organic) growth ("organic growth", "grow the business"), which translates into shareholder value and, ultimately, fund outperformance.

Interestingly, generic terms such as "investment criteria," have a negative median impact on predicted outperformance. Most of the terms positively correlated with outperformance are consistent with the pre- and post-investment sources of value creation that PE fund managers expect to influence deal returns (Gompers et al., 2016). One notable difference is the positive median effect of having access to proprietary deals ("proprietary investment opportunities"), which is presumably associated with buying at cheap valuations (Gompers et al., 2016). In line with what the industry believes, entry pricing per se has a negative median value ("entry valuation") in our analysis. Nevertheless, practitioners consider low entry pricing to be an important source of value creation in about half of their deals.

Finally, we note that only two textual terms could be linked to higher risk. As a whole, the list of most important textual terms in Figure 4 reassures us that risk is not the driving force behind the positive correlation between algorithmic prediction and ultimate performance.

4 Additional Results

4.1 Out-of-sample tests

The analysis presented in Tables 8 and 9 kept as many funds in the sample as possible. However, this choice results in a look-ahead bias since the algorithms are trained using performance information at the end of the sample period, and therefore use information that was not available at the end of the training period.

In Table 10, we conduct an exercise that reduces sample size but eliminates the look-ahead bias. The performance of the funds used in the training sample is measured as of December 31, 2013. This means we need to eliminate the funds that are still in their investment period at the end of the year 2013, which reduces the training samples by about half. The test sample continues to be the 72 funds from vintage years 2014 to 2016, with performance measured at the end of the year 2022.

Table 10 presents measures of in-sample and out-of-sample goodness-of-fit for the two algorithms (Lasso and Gradient Boosting; one column each) and for two sets of training data (Panels A and B). For Panel A the training dataset is the 141 funds raised between 1999 and 2007 whilst in Panel B it is the 122 funds raised between 2003 and 2007.

< Insert Table 10 >

Comparing with the larger training sample (Table 8), we see a lower out-of-sample goodness of fit for all algorithms. This is expected as the training sample size has decreased by more than half. The Lasso algorithm loses some predictability, but Gradient Boosting continues to produce a good out-of-sample fit especially when the training sample is limited to funds raised after 2003 (Panel B). Gradient Boosting also obtains the highest AUC (0.63) and balanced accuracy (0.60). Appendix Figure 2A provides further robustness tests.[23]

< Insert Table 11 >

We now repeat Table 9 with this pure out of sample estimation and results are shown in Table 11. We observe that the two clusters of low and high probability are still present, and in fact, are more extreme. For Gradient Boosting, the mean probability in the low and high terciles is basically 0% and 100%, respectively.

As with our findings for statistical accuracy measures, the percentage of correctly predicted outperformers (underperformers) decreases. Yet, if we retained the top tercile according to Gradient Boosting, we would end up with 67% of the funds outperforming, which is statistically significantly different from 50%. In the bottom tercile, it is less good but still better than random: 58% end up being underperformers indeed.

Another way to see these positive results is with average performance per tercile. If we select the top tercile according to Gradient Boosting, the average TVPI is 1.99, which is statistically significantly above the unconditional average of 1.86. Also, the average TVPI is only 1.65 for the funds in the bottom tercile. A 0.35 spread in realized performance between the two terciles is economically very large, about 6% per year (compounded over 5 years).

These results confirm that Gradient Boosting can predict outperformance, based on textual information. In contrast, Lasso is not as effective on this smaller sample. The findings support again the view that allowing for non-linearity is important.

4.2 Other ML predictions: quantitative information and fundraising success

So far, we have shown that standard fund characteristics (i.e., quantitative variables) are not correlated to subsequent fund performance, but that the qualitative information, particularly the one processed by a Gradient Boosting algorithm, does predict fund outperformance. However, it is possible that Gradient Boosting (or Lasso) could predict outperformance using fund characteristics.

[23] Figure A2 shows the ROC calculated using the Gradient Boosting algorithm trained on 122 funds raised between 2003 and 2007 and tested on the 72 funds raised between 2014 and 2016.

To analyze this possibility, instead of the TF-IDF scores, we train the machine learning models using the quantitative variables used in Table 5. The results of this analysis are displayed in Panel A of Table 12. The findings show that fund characteristics have no predictive power in our sample. These results are consistent with the findings of simple OLS regressions in Table 7.

< Insert Table 12 >

Next, we train the algorithms to predict fundraising success (i.e., being faster than the median fund in our sample) using the same qualitative information we used to predict fund performance in previous tables. Panel B of Table 12 shows no predictive power. One possible interpretation of this result is that investors do not take into account the information contained in the strategy section when forming their demand for a fund. A comparison with the results in Panel B of Table 8 shows that the algorithms are better at predicting fund outperformance than fundraising success. This is consistent with the idea that investors do not pay enough attention to qualitative data or are unable to fully use its predictive power (Cohen, Malloy, and Nguyen, 2020).

4.3 Drivers of Machine Learning Probabilities of Outperformance

The results presented so far imply that qualitative information can provide informational value beyond what quantitative information can do. In this section, we try to bring together the findings on qualitative and quantitative information. In section 3.5, we show which words play an important role in the probabilities generated by the Gradient Boosting algorithm. In this section, we study whether some fund characteristics, including demand are related to the Gradient Boosting probabilities.

Table 13 shows OLS regressions where the probability of outperformance predicted by the Gradient Boosting algorithm is the dependent variable. The explanatory variables are the same fund characteristics used in Table 5. Table 13 shows that no fund characteristic, including the strength of demand, is actually correlated with the generated probabilities. These results suggest that the machine learning algorithm picks up something uncorrelated with what the standard quantitative measures are picking up. In addition, the fact that the probabilities are not correlated with fund size or differ across regions or industries provides us further assurance that the algorithm is not selecting funds with a particular risk profile.[24]

< Insert Table 13 >

[24] Results are similar when we use the probability of success obtained from Lasso regression; see Appendix Table A4.

In Table 14, we show OLS regression results where the dependent variable is the natural logarithm of the number of months a PE firm takes to raise a given fund. These tests are similar to the analysis carried out in Table 6 where we control for the traditional quantitative fund characteristics and our readability measures. However, we add the probability of success predicted by the Gradient Boosting algorithm to each specification. In all regressions, the coefficient on this variable is statistically insignificant and close to zero. These results suggest that the algorithm picks up something that is not processed by investors and, thus, is not embodied in fundraising success. Again, results are similar when we use the probability of success obtained from Lasso regression (see Appendix Table A5).

< Insert Table 14>

4.4 Machine Learning selection versus Market selection

Another way to display our results on the impact of quantitative and qualitative information in PE fund selection is Figure 5. The figure provides similar information to Figure 3, but presents results based on the number of funds selected by investors per year. Figure 5 depicts the size-weighted average TVPI of the funds selected by the Gradient Boosting algorithm against those that are able to raise capital more successfully (i.e., perceived to be successful by investors) using the funds raised between 2014 and 2016. To compare the selection of Gradient Boosting versus that of investors, we use the out-of-sample probabilities resulting from training the algorithm with performance information available as of December 31st 2013 (as in Panel B of Table 10). As in previous analyses, we use the number of months needed to close a fund to proxy for fundraising success. The red line depicts the size-weighted average TVPI of the 72 funds raised between 2014 and 2016.

< Insert Figure 5 >

The figure shows that across all portfolio sizes, size-weighted average TVPI of the Gradient Boosting is higher than the size-weighted average TVPI of the funds with the fastest fundraising speed and the size-weighted average TVPI. The numbers in the graph imply that an investor committing capital to the top five funds per year selected by Gradient Boosting would have achieved a TVPI of 2.3x, whereas an investor putting capital into the top five funds with the fastest fundraising time would have generated a TVPI of 2.1x. Interestingly, this graph also shows that the funds picked by the market, proxied by the months in fundraising, have a TVPI close to the size-weighted average TVPI. This finding is consistent with the idea that the market has a hard time identifying the most successful funds.

5 Conclusion

Our study relies on the use of qualitative information provided to PE investors to predict future fund performance. Applying standard textual analysis and machine learning techniques, we provide the first empirical analysis of readability and qualitative information in private markets. Our approach may be particularly useful if we consider that non-standardized disclosures and inherent information asymmetries characterize this market.

We start the paper by conducting a traditional econometric analysis of the determinants of fundraising success using standard proxies for PE firm reputation and past performance. Our results show a positive association of most of these variables with success in PE raising capital. We complement this analysis looking at readability measures. Although readability measures have been shown to matter in public market disclosure, we do not find them to be consistently associated with fundraising success in PE. This may be due to the non-standardized nature of disclosure in private markets or the thorough analysis carried out by investors.

Since investors seem to consider quantitative factors in their PE capital allocation decisions, we analyze if these quantitative factors are good predictors of subsequent fund performance. Our results show that this is not the case: traditional quantitative factors and document readability proxies are poor predictors of fund outperformance. In addition, we do not find that our fundraising success proxies are correlated with subsequent fund outperformance. This pattern of findings seems difficult to reconcile with Cavagnaro et al. (2019) who show that some investors seem to make persistently good fund manager choices. One could argue that these investors may be exploiting a different set of signals to learn about differential PE manager ability.

In this paper, we test for the first time if qualitative information provided by the fund managers as they lay out their investment strategy is useful to extract such a set of signals. We are the first to use Natural Language Process (NLP) techniques and machine learning algorithms to analyze whether this kind of qualitative information predicts returns. Results show that approaches exploiting the qualitative information disclosed to investors in PPMs have important predictive power for subsequent fund outperformance. Our study provides evidence of the value of applying these new technologies to process qualitative information in private markets. Future studies may address some of the limitations of our current work. One such limitation, which we are currently working on, is the inclusion of factors capturing the experience and background of the investment teams in the analysis.

We hope that our study demonstrates the potential of new methods and motivates owners of proprietary qualitative data to cooperate with researchers and put together larger samples that can be

analyzed. There are certainly additional potential improvements as the disciplines of textual analysis and machine learning continue their rapid growth and provide even more powerful methods to be applied in private capital markets. We believe our findings have important implications and real-world applications for investors in private markets. Our results suggest that there are signals of differential ability buried in qualitative information which can be exploited with the methods used in this study.

Bibliography

Acharya, V. V., Gottschalg, O. F., Hahn, M., & Kehoe, C. (2013). Corporate governance and value creation: Evidence from private equity. *The Review of Financial Studies*, *26*(2), 368-402.

Barber, B. M., & Yasuda, A. (2017). Interim fund performance and fundraising in private equity. *Journal of Financial Economics*, 124(1), 172-194.

Berk, J. B., & Green, R. C. (2004). Mutual fund flows and performance in rational markets. *Journal of Political Economy*, 112(6), 1269-1295.

Bianchi, D., Büchner, M., & Tamoni, A. (2021). Bond risk premiums with machine learning. *The Review of Financial Studies*, 34(2), 1046-1089.

Biesinger, M., Bircan, C., & Ljungqvist, A. (2020). Value creation in private equity. *European Bank of Reconstruction and Development Working Paper Number 242.*

Bloomfield, R. J. (2002). The "incomplete revelation hypothesis" and financial reporting. *Cornell University Working Paper*. Available at SSRN 312671.

Braun, R., Dorau, N., Jenkinson, T. & Urban, D. (2023). Size, returns, and value added: Do private equity firms allocate capital according to manager skill? Available at SSRN 3475460.

Braun, R., Jenkinson, T., & Stoff, I. (2017). How persistent is private equity performance? Evidence from deal-level data. *Journal of Financial Economics*, *123*(2), 273-291.

Brown, G. W., Gredil, O. R., & Kaplan, S. N. (2019). Do private equity funds manipulate reported returns? *Journal of Financial Economics*, 132(2), 267-297.

Brown, N. C., Crowley, R. M., & Elliott, W. B. (2020). What are you saying? Using topic to detect financial misreporting. *Journal of Accounting Research*, *58*(1), 237-291.

Bubb, R., & Catan, E. M. (2022). The party structure of mutual funds. *The Review of Financial Studies*, 35(6), 2839-2878.

Cavagnaro, D. R., Sensoy, B. A., Wang, Y., & Weisbach, M. S. (2019). Measuring institutional investors' skill at making private equity investments. *The Journal of Finance*, 74(6), 3089-3134.

Cohen, L., Malloy, C., & Nguyen, Q. (2020). Lazy prices. *The Journal of Finance*, *75*(3), 1371-1415.

Cohen, L., Diether, K., & Malloy, C. (2013). Misvaluing innovation. *The Review of Financial Studies*, 26(3), 635-666.

Da Rin, M., & Phalippou, L. (2017). The importance of size in private equity: Evidence from a survey of limited partners. *Journal of Financial Intermediation*, *31*, 64-76.

Davenport, D. (2022). Predictably bad investments: Evidence from venture capitalists. Available at SSRN 4135861.

Edmans, A. (2011). Does the stock market fully value intangibles? Employee satisfaction and equity prices. *The Journal of Financial Economics*, 101(3), 621-640.

Edmans, A. (2022). The end of ESG. *Financial Management*, 52(1), 3-17,

Erel, I., Stern, L. H., Tan, C., & Weisbach, M. S. (2021). Selecting directors using machine learning. *The Review of Financial Studies*, 34(7), 3226-3264.

Gompers, P., & Lerner, J. (1998). Venture capital distributions: Short-run and long-run reactions. *The Journal of Finance*, 53(6), 2161-2183.

Gompers, Paul, Steven N. Kaplan, & Vladimir Mukharlyamov (2016), What do private equity firms say they do? *Journal of Financial Economics,* 121, 449–476.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223-2273.

Gupta, A. & van Nieuwerburgh, S. (2021) Valuing private equity investments strip by strip. *The Journal of Finance*, 76(7). pp. 3255-3307.

Guo, Shourun, Edith S. Hotchkiss, & Weihong Song (2011). Do Buyouts (Still) Create Value? *The Journal of Finance*, 66 (2), pp. 479-517.

Harris, R. S., Jenkinson, T., Kaplan, S. N., & Stucke, R. (2023). Has persistence persisted in private equity? Evidence from buyout and venture capital funds, *Journal of Corporate Finance*, *81*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and Hober prediction. Springer Science & Business Media.

Hochberg, Y. V., Ljungqvist, A., & Lu, Y. (2007). Whom you know matters: Venture capital networks and investment performance. *The Journal of Finance*, 62(1), 251-301.

Hochberg, Y. V., Ljungqvist, A., & Vissing-Jørgensen, A. (2014). Informational holdup and performance persistence in venture capital. *The Review of Financial Studies*, 27(1), 102-152.

Hoberg, G., & Phillips, G. (2010). Product market synergies and competition in mergers and acquisitions: A text-based analysis. *The Review of Financial Studies*, 23(10), 3773-3811.

Jenkinson, T., Landsman, W. R., Rountree, B. R., & Soonawalla, K. (2020). Private equity net asset values and future cash flows. *The Accounting Review*, 95(1), 191-210.

Jensen, Michael C. (1989), Eclipse of the Public Corporation, *Harvard Business Review*, 67, pp. 61-73.

Kaplan, S. N., & Schoar, A. (2005). Private equity performance: Returns, persistence, and capital flows. *The Journal of Finance*, 60(4), 1791-1823.

Kruglikov, N., & Forthun, A. (2022). *Predicting Private Equity Fund Performance with Machine Learning* (Master's thesis).

Ke, Z. T., Kelly, B. T., & Xiu, D. (2019). Predicting returns with text data. *University of Chicago, Becker Friedman Institute for Economics Working Paper No. 2019-69*. Available in SSRN 3389884.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45(2-3), 221-247.

Li, K., Mai, F., Shen, R., & Yan, X. (2021). Measuring corporate culture using machine learning. *The Review of Financial Studies*, 34(7), 3265-3315.

La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2006) What Works in Securities Laws? *The Journal of Finance*, **LXI**, No. 1, February, pp. 1-32.

Lerner, J., Schoar, A., & Wongsunwai, W. (2007). Smart institutions, foolish choices: The limited partner performance puzzle. *The Journal of Finance*, *62*(2), 731-764.

Lopez-de-Silanes, F., Phalippou, L., & Gottschalg, O. (2015). Giants at the gate: Investment returns and diseconomies of scale in private equity. *Journal of Financial and Quantitative Analysis*, 50(3), 377-411.

Lopez-Lira, A. (2023). Risk factors that matter: Textual analysis of risk disclosures for the cross-section of returns. *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper.*

Loughran, T., & McDonald, B. (2014). Measuring readability in financial disclosures. *The Journal of Finance*, 69(4), 1643-1671.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4), 1187-1230.

Loughran, T., & McDonald, B. (2020). Textual analysis in finance. *Annual Review of Financial Economics*, 12, 357-375.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Lyonnet, V., & Stern, L. H. (2022). Venture Capital (Mis) allocation in the Age of AI. *Fisher College of Business Working Paper, (2022-03), 002*.

Malenko, A, Malenko, N. (2015). A theory of LBO activity based on repeated debt-equity conflicts, *Journal of Financial Economics*, 117(3), 607-627.

Malenko, A., Nanda, R., Rhodes-Kropf, M., & Sundaresan, S. (2023). Catching outliers: Committee voting and the limits of consensus when financing innovation. Available at SSRN 3866854.

Perfetto, M. (2019). Quantitative Selection of Private Equity Funds. École Polytechnique Fédérale de Lausanne, Master Thesis in Financial Engineering.

Phalippou, L. (2010). Venture capital funds: Performance persistence and flow-performance relation, *Journal of Banking and Finance* 34, 568-577

Phalippou, L., Rauch, C., & Umber, M. (2018). Private equity portfolio company fees. *Journal of Financial Economics*, *129*(3), 559-585.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *The Journal of machine Learning Research, 12, 2825-2830.*

Rajan, R. G., Ramella, P., & Zingales, L. (2022). What Purpose Do Corporations Purport? Evidence from Letters to Shareholders. *NBER Working Paper No. w31054*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic Interpretability of Machine Learning. *arXiv preprint arXiv:1606.05386*.

Robinson, D. T., & Sensoy, B. A. (2013). Do private equity fund managers earn their fees? Compensation, ownership, and cash flow performance. *The Review of Financial Studies*, 26(11), 2760-2797.

Robinson, D. T., & Sensoy, B. A. (2016). Cyclicality, performance measurement, and cash flow liquidity in private equity. *Journal of Financial Economics*, 122(3), 521-543.

Salton, G., & Buckley, C. (1988). Term-weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5), 513-523.

Sensoy, B. A., Wang, Y., & Weisbach, M. S. (2014). Limited partner performance and the maturing of the private equity industry. *Journal of Financial Economics*, *112*(3), 320-343.

Stafford, E. (2022). Replicating private equity with value investing, homemade leverage, and hold-to-maturity accounting. *The Review of Financial Studies*, 35(1), pp. 299-342.

Vilone, G., & Longo, L. (2020). Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*.

Table 1. Sample construction and comparison with Preqin datasets

The table details the construction of our sample of funds and compares our sample at each stage with the two different Preqin datasets.

	Our sample			Preqin Summary Sample			Preqin Cash-Flow Sample		
	# Funds	Fund Size (EURm)		# Funds	Fund Size (EURm)		# Funds	Fund Size (EURm)	
	(%)	Average	Median	(%)	Average	Median	(%)	Average	Median
Initial sample from archives of data provider	941	836	291	37,798	242	54	9,272	560	176
1. Vintage year 1999-2020, size>€5mn, listed in Preqin	737	836	291	28,309	265	66	8,580	572	183
2. Vintage year 1999-2016	589	791	282	18,728	237	66	6,743	472	165
(2/1)	(80%)			(66%)			(79%)		
3. Buyout, Growth Capital, Turnaround, Balanced	505	879	322	6,682	402	117	2,799	726	260
(3/2)	(86%)			(36%)			(42%)		
4. Investing in Europe, Asia, or North America	503	878	322	6,147	420	121	2,601	757	271
(4/3)	(100%)			(92%)			(93%)		
5. PPM includes an Investment Strategy section	501	881	322	6,147	420	121	2,601	757	271
(5/4)	(100%)			(100%)			(100%)		
6. TVPI or IRR available at age 6 or later	395	1025	354	6,147	420	121	2,350	810	283
(6/5)	(79%)			(100%)			(90%)		

Table 2: Information Sources for Fund Performance

The table shows the sources of the data of fund performance for our final sample of funds.

	# Funds (%)	Fund Size (EURm) Average	Median
Fund Cash Flows from Internal Sources	100 (25%)	1316	507
Fund Cash Flows from Preqin Database	34 (9%)	1826	1047
Summary Fund Performance from Internal Sources	61 (15%)	260	219
Preqin Fund Performance Sample	200 (51%)	976	319
All	395 (100%)	1025	354

Table 3: Fund Characteristics

The table provides descriptive statistics for our final sample of funds (Panel A) and correlations among the main variables used in the paper (Panel B). Definitions of all the variables are provided in Appendix Table A1. Note that Preceding Fund Performance is measured by MOIC.

Panel A: Descriptive Statistics

		N_Obs	Mean	Std Dev.	p25	Median	p75
Fund Performance:							
1	TVPI	395	1.8	0.7	1.4	1.7	2.1
2	Excess TVPI	395	0.1	0.7	-0.3	0.0	0.4
3	Excess TVPI > 0	395	52%				
4	IRR	395	14%	13%	8%	14%	21%
Fundraising Success:							
5	Months in Fundraising	395	13.8	9.2	6.2	12.0	19.3
6	Oversubscription	395	105%	31%	87%	107%	125%
Fund Characteristics:							
7	Fund Size	395	1025	1913	176	354	815
8	Fund Sequence	395	3.3	2.0	2.0	3.0	4.0
9	Preceding Fund Performance	318	1.6	0.6	1.2	1.5	1.9
PPM Readability:							
10	Number of words (strategy section)	395	2774	1281	1829	2644	3724
11	Number of pages of the PPM	395	84	36	62	78	101

Panel B: Correlation matrix

		1	2	3	4	5	6	7	8	9	10
1	TVPI										
2	Excess TVPI	0.98									
3	Excess TVPI > 0	0.68	0.70								
4	IRR	0.89	0.85	0.60							
5	Months in Fundraising	0.02	0.01	0.00	0.05						
6	Oversubscription	-0.05	-0.03	0.03	-0.06	-0.35					
7	Fund Size	0.11	0.11	0.24	0.08	0.01	0.16				
8	Fund Sequence	0.03	0.02	0.03	0.02	-0.14	0.13	0.41			
9	Preceding Fund Performance	-0.04	-0.05	-0.02	-0.02	-0.13	0.12	-0.12	-0.25		
10	Number of words (strategy section)	0.05	0.03	0.01	0.06	0.14	-0.03	0.13	0.07	-0.06	
11	Number of pages of the PPM	-0.01	-0.04	0.02	-0.01	0.08	0.09	0.39	0.18	-0.11	0.47

Table 4. Most Common stemmed bigrams and trigrams

Panel A shows the most common stemmed bigrams (i.e., combinations of two adjacent stemmed words) in our sample. Panel B displays the most common trigrams (i.e., combinations of three adjacent stemmed words). For each stemmed biagram or trigram, the table reports the number of observations, the percentage of documents containing the stemmed bigram or trigram, and the average and standard deviation of the TF-IDF score.

	Number of observations across all PPMs	% PPM with this term	TF-IDF	
			Average	Std. Dev.
Panel A: bigrams				
portfolio company	4,065	92.41%	1.53%	0.69%
manag team	2,752	92.15%	1.30%	0.57%
invest opportune	1,602	87.85%	1.09%	0.56%
due dilig	2,075	87.09%	1.18%	0.62%
privat equity	1,422	84.56%	1.01%	0.61%
invest strategi	878	82.28%	0.84%	0.55%
valu creation	1,423	72.91%	0.99%	0.76%
target company	869	69.11%	0.84%	0.71%
compani manag	653	68.35%	0.76%	0.66%
invest process	669	66.84%	0.77%	0.67%
deal flow	818	65.82%	0.89%	0.81%
long term	529	60.51%	0.67%	0.64%
invest company	461	60.51%	0.69%	0.69%
fund invest	726	58.99%	0.83%	0.84%
cash flow	569	58.48%	0.69%	0.70%
Panel B: trigrams				
due dilig process	404	49.62%	0.50%	0.58%
portfolio compani manag	293	39.24%	0.46%	0.66%
privat equiti firm	212	33.16%	0.36%	0.56%
manag portfolio compani	186	32.66%	0.38%	0.61%
compani manag team	204	32.41%	0.37%	0.59%
privat equiti invest	190	28.10%	0.34%	0.61%
attract invest opportun	166	27.59%	0.31%	0.55%
proprietari deal flow	158	26.08%	0.33%	0.60%
valu portfolio compani	129	23.54%	0.28%	0.56%
privat equiti fund	129	22.53%	0.30%	0.61%
decis make process	128	21.01%	0.29%	0.60%
buy build strategi	168	20.76%	0.27%	0.57%
potenti invest opportun	109	19.75%	0.33%	0.72%
fund portfolio compani	120	19.49%	0.24%	0.52%
privat equiti investor	105	18.73%	0.25%	0.55%

Table 5. Similarity of bigrams and trigrams across years

The table shows the cosine similarity score of vectors representing the annual average frequency of stemmed terms across years. The number of PPMs in each vintage year is shown in the last row of the table.

Year	'99	'00	'01	'02	'03	'04	'05	'06	'07	'08	'09	'10	'11	'12	'13	'14	'15
'00	0.37																
'01	0.48	0.40															
'02	0.51	0.42	0.35														
'03	0.46	0.62	0.47	0.49													
'04	0.46	0.52	0.49	0.45	0.64												
'05	0.47	0.58	0.50	0.51	0.71	0.66											
'06	0.54	0.65	0.57	0.57	0.78	0.74	0.80										
'07	0.52	0.65	0.54	0.58	0.78	0.76	0.81	0.88									
'08	0.51	0.62	0.51	0.56	0.74	0.70	0.80	0.88	0.87								
'09	0.49	0.64	0.49	0.58	0.76	0.68	0.78	0.86	0.88	0.85							
'10	0.52	0.55	0.55	0.49	0.68	0.67	0.73	0.81	0.78	0.78	0.73						
'11	0.52	0.59	0.51	0.58	0.75	0.68	0.78	0.85	0.86	0.82	0.83	0.75					
'12	0.50	0.60	0.50	0.55	0.74	0.72	0.77	0.87	0.88	0.87	0.85	0.75	0.83				
'13	0.51	0.59	0.50	0.55	0.72	0.73	0.77	0.86	0.88	0.88	0.84	0.75	0.83	0.88			
'14	0.47	0.55	0.47	0.52	0.71	0.67	0.77	0.84	0.85	0.87	0.81	0.74	0.81	0.85	0.86		
'15	0.52	0.60	0.52	0.56	0.75	0.72	0.79	0.88	0.88	0.86	0.85	0.80	0.85	0.89	0.88	0.85	
'16	0.51	0.57	0.50	0.54	0.75	0.68	0.76	0.84	0.87	0.85	0.84	0.77	0.85	0.86	0.87	0.83	0.87
#obs	*3*	*6*	*5*	*5*	*11*	*13*	*22*	*39*	*49*	*39*	*25*	*20*	*27*	*27*	*32*	*25*	*28*

Table 6: Fundraising Success

This table presents OLS regression results. Standard errors are in parentheses and clustered at the PE firm level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively. Definitions of all the variables are provided in Appendix Table A1.

Panel A: The dependent variable is the natural logarithm of the number of months it took to raise the fund

	(1)	(2)	(3)	(4)	(5)
Fund Size (ln)	-0.087**	-0.052	-0.089**	-0.086**	-0.086**
	(0.037)	(0.041)	(0.039)	(0.040)	(0.041)
Fund Sequence (ln)		-0.143*	-0.153*	-0.143*	-0.259**
		(0.084)	(0.081)	(0.081)	(0.129)
Number of PPM Pages (ln)			0.378***	0.310**	0.380**
			(0.117)	(0.128)	(0.155)
Number of words (strategy section) (ln)				0.107	0.049
				(0.076)	(0.081)
Preceding Fund Performance					-0.148**
					(0.070)
Other control variables					
Investment type	Yes	Yes	Yes	Yes	Yes
Targeted region	Yes	Yes	Yes	Yes	Yes
Vintage year	Yes	Yes	Yes	Yes	Yes
Observations	395	395	395	395	318
R-squared	0.124	0.132	0.155	0.159	0.160

Panel B: The dependent variable is the fund oversubscription ratio

	(1)	(2)	(3)	(4)	(5)
Fund Size (ln)	0.094***	0.098***	0.103***	0.102***	0.087***
	(0.013)	(0.014)	(0.015)	(0.015)	(0.016)
Fund Sequence (ln)		-0.020	-0.019	-0.022	-0.037
		(0.026)	(0.026)	(0.026)	(0.034)
Number of PPM Pages (ln)			-0.046	-0.024	-0.052
			(0.044)	(0.052)	(0.051)
Number of words (strategy section) (ln)				-0.035	-0.002
				(0.035)	(0.036)
Preceding Fund Performance					0.055*
					(0.031)
Other control variables					
Investment type	Yes	Yes	Yes	Yes	Yes
Targeted region	Yes	Yes	Yes	Yes	Yes
Vintage year	Yes	Yes	Yes	Yes	Yes
Observations	395	395	395	395	318
R-squared	0.318	0.319	0.321	0.324	0.294

Table 7: Fund performance

This table presents OLS regression results. Standard errors are in parentheses and clustered at the PE firm level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively. Definitions of all the variables are provided in Appendix Table A1.

Panel A: The dependent variable is TVPI

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Fund Size (ln)	0.002	-0.007	-0.007	0.011	0.049	0.051	0.046	
	(0.029)	(0.034)	(0.034)	(0.037)	(0.038)	(0.038)	(0.039)	
Fund Sequence (ln)		0.037	0.036	0.045	-0.034	-0.026	-0.022	
		(0.071)	(0.071)	(0.069)	(0.087)	(0.088)	(0.089)	
Number of PPM Pages (ln)			-0.177*	-0.200*	-0.248**	-0.260**	-0.258**	
			(0.105)	(0.108)	(0.117)	(0.119)	(0.119)	
Number of words (strategy section) (ln)				0.036	0.080	0.078	0.078	
				(0.075)	(0.083)	(0.083)	(0.083)	
Preceding Fund Performance					-0.055	-0.050	-0.053	
					(0.065)	(0.064)	(0.064)	
Months in Fundraising (ln)						0.031	0.036	-0.003
						(0.042)	(0.045)	(0.045)
Oversubscription Rate							0.064	0.009
							(0.137)	(0.118)
Other control variables								
Investment type	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Targeted region	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Vintage year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	395	395	395	395	318	318	318	395
R-squared	0.105	0.106	0.112	0.113	0.106	0.107	0.108	0.105

Panel B: The dependent variable is Excess TVPI

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Fund Size (ln)	-0.001	-0.011	0.007	0.008	0.048	0.051	0.046	
	(0.029)	(0.034)	(0.037)	(0.037)	(0.038)	(0.038)	(0.040)	
Fund Sequence (ln)		0.040	0.045	0.049	-0.033	-0.025	-0.021	
		(0.071)	(0.069)	(0.069)	(0.087)	(0.088)	(0.088)	
Number of PPM Pages (ln)			-0.184*	-0.208*	-0.253**	-0.266**	-0.265**	
			(0.107)	(0.109)	(0.117)	(0.119)	(0.119)	
Number of words (strategy) (ln)				0.038	0.079	0.077	0.077	
				(0.076)	(0.083)	(0.083)	(0.083)	
Preceding Fund Performance					-0.067	-0.062	-0.064	
					(0.065)	(0.064)	(0.064)	
Months in Fundraising (ln)						0.033	0.038	-0.002
						(0.042)	(0.045)	(0.045)
Oversubscription Rate							0.053	-0.012
							(0.138)	(0.119)
Other control variables								
Investment type	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Targeted region	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Vintage year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	395	395	395	395	318	318	318	395
R-squared	0.074	0.075	0.082	0.083	0.089	0.090	0.091	0.074

Table 8: Predicting Fund Performance (Pseudo out of sample)

The table shows the goodness of fit of two machine learning algorithms. In-sample fit refers to the goodness of fit of the training sample. The test sample consists of funds raised between 2014 and 2016. Fund performance is measured by TVPI as of June 2022 for all funds. In Panel A, the training sample consists of the 323 funds raised between 1999 and 2013. In Panel B, the training sample consists of the 304 funds raised between 2003 and 2013. See Appendix Table A3 for details on the methodology.

Panel A: Training on 1999-2013 funds

	Lasso	Gradient Boosting
In-sample Fit:		
Area Under Curve	0.57	0.62
Balanced Accuracy	0.53	0.60
Pseudo Out-of-sample Fit:		
Area Under Curve	0.60	0.59
Balanced Accuracy	0.60	0.58

Panel B: Training on 2003-2013 funds

	Lasso	Gradient Boosting
In-sample Fit:		
Area Under Curve	0.59	0.61
Balanced Accuracy	0.58	0.58
Pseudo Out-of-sample Fit:		
Area Under Curve	0.62	0.66
Balanced Accuracy	0.57	0.59

Table 9: Predicted probability of outperformance by terciles (Pseudo out-of-sample)

This table shows the pseudo out-of-sample accuracy of the Lasso (Panel A) and Gradient Boosting (Panel B) algorithms by terciles. The psuedo out-of-sample consists of the funds raised between 2014 and 2016. Fund performance is measured by TVPI as of June 2022 for all funds. Each panel shows statistics of the predicted probability of outperformance, the percentage of outperforming funds, statistics of TVPI, and the mean Excess TVPI. The panels also show the results of statistical tests to see whether funds in the high tercile of predicted probability of outperformance ultimately outperform those funds in the low tercile of predicted probability of outperformance: T-tests for mean TVPI and Excess TVPI, Wilcoxon rank-sum tests for median TVPIs, and Chi-square tests for shares of under- and outperformers, respectively. *, **, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively. Definitions of all the variables are provided in Appendix Table A1.

Panel A: Lasso

	Predicted Probability of Outperformance			Outperformers	TVPI				Excess TVPI
	Mean	Min	Max	Percent	Mean	Median	SD	Bottom 25% (Percent)	Mean
Total	0.50	0.00	1.00	51%	1.86	1.77	0.52	25%	0.06
Low	0.08	0.00	0.23	42%	1.62	1.72	0.45	42%	-0.18
Medium	0.49	0.24	0.81	46%	1.79	1.75	0.33	25%	-0.01
High	0.93	0.81	1.00	67%*	2.16***	2.01***	0.59	8%***	0.37***

Panel B: Gradient Boosting

	Predicted Probability of Outperformance			Outperformers	TVPI				Excess TVPI
	Mean	Min	Max	Percent	Mean	Median	SD	Bottom 25% (Percent)	Mean
Total	0.65	0.00	1.00	51%	1.86	1.77	0.52	25%	0.06
Low	0.05	0.00	0.35	33%	1.66	1.70	0.50	42%	-0.13
Medium	0.89	0.36	1.00	46%	1.85	1.79	0.46	29%	0.05
High	1.00	1.00	1.00	75%***	2.07***	1.93***	0.52	4%***	0.26***

Table 10. Predicting Fund Performance (Pure out of sample)

This table shows the goodness of fit of two machine learning algorithms. In-sample fit refers to the goodness of fit of the training sample. The test sample consists of funds raised between 2014 and 2016. Fund performance is measured by TVPI as of June 2022 for the funds in the test sample, and is measured by TVPI as of December 2013 for the funds in the training sample. In Panel A, the training sample consists of the funds raised between 1999 and 2007. In Panel B the training sample consists of the funds raised between 2003 and 2007.

Panel A: Training on 1999-2007 funds

	Lasso	Gradient Boosting
In-sample Fit:		
Area Under Curve	0.64	0.56
Balanced Accuracy	0.62	0.53
Out-of-sample Fit:		
Area Under Curve	0.47	0.56
Balanced Accuracy	0.47	0.53

Panel B: Training on 2003-2007 funds

	Lasso	Gradient Boosting
In-sample Fit:		
Area Under Curve	0.60	0.63
Balanced Accuracy	0.56	0.60
Out-of-sample Fit:		
Area Under Curve	0.53	0.64
Balanced Accuracy	0.50	0.56

Table 11: Predicted probability of outperformance by terciles (Pure out of sample)

This table shows the pure out-of-sample accuracy of the Lasso (Panel A) and Gradient Boosting (Panel B) algorithms by terciles. The pure out-of-sample consists of the funds raised between 2014 and 2016. Fund performance is measured by TVPI as of June 2022 for all funds. Each panel shows statistics of the predicted probability of outperformance, the percentage of outperforming funds, statistics of TVPI, and the mean Excess TVPI. The panels also show the results of statistical tests to see whether funds in the high tercile of predicted probability of outperformance ultimately outperform those funds in the low tercile of predicted probability of outperformance: T-tests for mean TVPI and Excess TVPI, Wilcoxon rank-sum tests for median TVPIs, and Chi-square tests for shares of under- and outperformers, respectively. *, **, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively. Definitions of all the variables are provided in Appendix Table A1.

Panel A: Lasso

	Predicted Probability of Outperformance			Outperformers Percent	TVPI				Excess TVPI
	Mean	Min	Max		Mean	Median	SD	Bottom 25% (Percent)	Mean
Total	0.48	0.00	1.00	0.51	1.86	1.77	0.52	25%	0.06
Low	0.11	0.01	0.26	0.50	1.88	1.83	0.64	25%	0.07
Medium	0.46	0.29	0.68	0.50	1.80	1.73	0.54	29%	0.02
High	0.88	0.70	0.99	0.54	1.89	1.79	0.34	21%	0.09

Panel B: Gradient Boosting

	Predicted Probability of Outperformance			Outperformers Percent	TVPI				Excess TVPI
	Mean	Min	Max		Mean	Median	SD	Bottom 25% (Percent)	Mean
Total	0.49	0.00	1.00	0.51	1.86	1.77	0.52	25%	0.06
Low	0.00	0.00	0.02	0.42	1.65	1.67	0.54	42%	-0.13
Medium	0.47	0.02	0.98	0.46	1.93	1.80	0.50	8%	0.14
High	1.00	0.99	1.00	0.67*	1.99**	1.95**	0.45	25%	0.17**

Table 12: Other ML Predictions

The table shows the goodness of fit of two machine learning algorithms. In-sample fit refers to the goodness of fit of the training sample. The test sample consists of the funds raised between 2014 and 2016. Fund performance is measured by TVPI as of June 2022 for all funds. Fundraising Speed is measured by the number of months needed to reach the final closing of the fund. In Panel A, algorithms are trained on 295 funds using quantitative (salient) information as features to predict outperformance (i.e., past performance, fund size, and fund sequence). In Panel B, algorithms are trained on 303 funds using TF-IDF vectors as features to predict Fundraising Speed being faster than the sample median.

Panel A: Predicting Fund Performance using Quantitative Information

	Lasso	Gradient Boosting
In-sample Fit:		
Area Under Curve	0.54	0.58
Balanced Accuracy	0.52	0.55
Pseudo Out-of-sample Fit:		
Area Under Curve	0.54	0.50
Balanced Accuracy	0.59	0.48
Correctly predicted outperformers	0.37	0.46
Correctly predicted underperformers	0.81	0.50

Panel B: Predicting Fundraising Speed using Qualitative Information

	Lasso	Gradient Boosting
In-sample Fit:		
Area Under Curve	0.57	0.58
Balanced Accuracy	0.54	0.54
Pseudo Out-of-sample Fit:		
Area Under Curve	0.51	0.47
Balanced Accuracy	0.53	0.40
Correctly predicted fundraising speed < median	0.59	0.51
Correctly predicted fundraising speed < median	0.46	0.29

Table 13. Predicted Probability of Outperformance and Fund Characteristics

This table presents OLS regression results for the sample of 376 funds raised between 2003 and 2016 where dependent variable is the predicted probability of outperformance using the Gradient Boosting algorithm. Standard errors are in parentheses and clustered at the PE firm level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively. Definitions of all the variables are provided in Appendix Table A1.

Dependent Variable: Probability of Success predicted by the Gradient Boosting algorithm					
	(1)	(2)	(3)	(4)	(5)
Fund Size (ln)	0.008	0.022	0.022	0.022	0.026
	(0.025)	(0.028)	(0.028)	(0.029)	(0.032)
Fund Sequence (ln)		-0.056	-0.056	-0.056	0.008
		(0.048)	(0.049)	(0.049)	(0.073)
Number of words (strategy) (ln)			0.001	0.002	0.018
			(0.051)	(0.054)	(0.063)
Number of PPM Pages (ln)				-0.003	0.046
				(0.085)	(0.096)
Preceding Fund Gross TVPI					-0.025
					(0.043)
Other control variables					
Investment type	Yes	Yes	Yes	Yes	Yes
Targeted region	Yes	Yes	Yes	Yes	Yes
Vintage year	Yes	Yes	Yes	Yes	Yes
Observations	376	376	376	376	309
R-squared	0.039	0.043	0.043	0.043	0.069

Table 14. Fundraising Success and Predicted Probability of Success

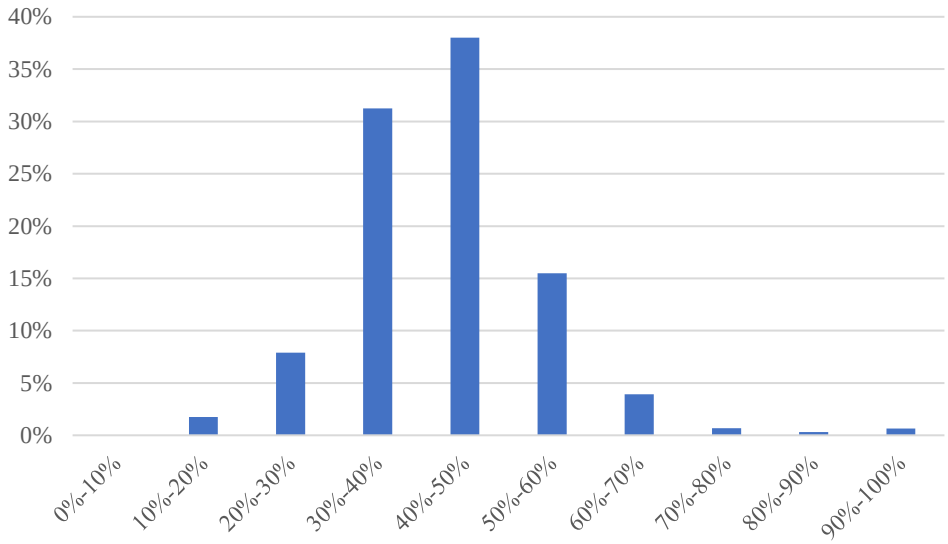
This table presents OLS regression results for the sample of 376 funds raised between 2003 and 2016. Standard errors are in parentheses and clustered at the PE firm level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively. Definitions of all the variables are provided in Appendix Table A1.

Dependent Variable: Natural logarithm of the number of months to raise the fund					
	(1)	(2)	(3)	(4)	(5)
Gradient Boosting - Probability	0.021	0.011	0.011	0.011	-0.077
	(0.101)	(0.101)	(0.100)	(0.098)	(0.109)
Fund Size (ln)	-0.102***	-0.065	-0.072*	-0.101**	-0.094**
	(0.038)	(0.043)	(0.043)	(0.042)	(0.044)
Fund Sequence (ln)		-0.144*	-0.127	-0.138*	-0.251*
		(0.086)	(0.084)	(0.083)	(0.130)
Number of words (strategy) (ln)			0.190**	0.107	0.061
			(0.075)	(0.081)	(0.086)
Number of PPM Pages (ln)				0.328**	0.374**
				(0.140)	(0.159)
Preceding Fund Gross TVPI					-0.148**
					(0.073)
Other control variables					
Investment type	Yes	Yes	Yes	Yes	Yes
Targeted region	Yes	Yes	Yes	Yes	Yes
Vintage year	Yes	Yes	Yes	Yes	Yes
Observations	376	376	376	376	309
R-squared	0.084	0.092	0.107	0.121	0.144

Figure 1. Similarity of PPM strategy sections of the same PE firm and of the same Law firm

The figures show histograms of cosine similarity between TF-IDF vectors obtained from the investment strategy section of a PPM. In Panel A, each observation is the average cosine similarity within the PPMs written by the same law firm. In Panel B, each observation is a pair of two consecutive funds raised by the same PE firm.

Panel A: Similarity within PPMs written by the same Law Firm



Panel B: Similarity across successive funds of the same PE firm

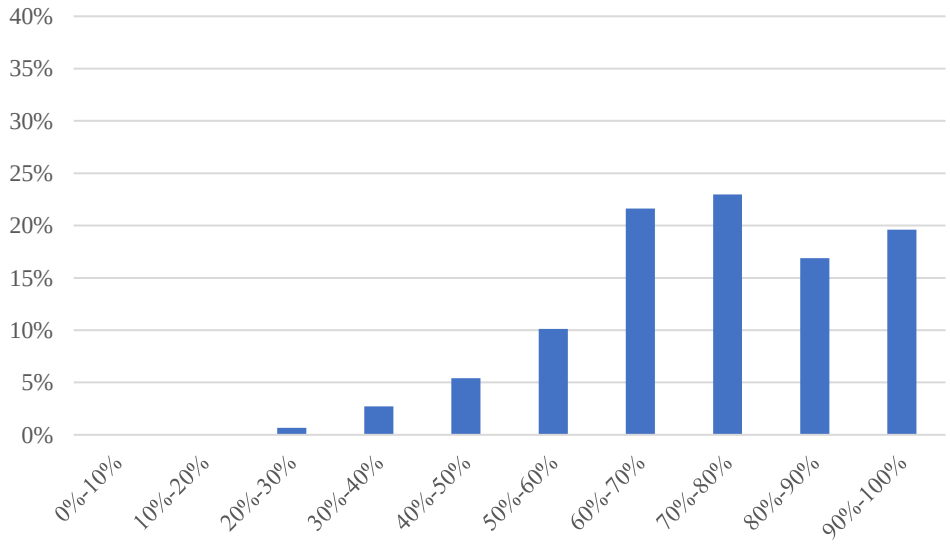


Figure 2. Receiver Operating Characteristic curve using Gradient Boosting

The figure shows the Receiving Operating Characteristic curve (ROC) using the Gradient Boosting algorithm. Gradient Boosting is trained using funds from 2003 to 2013. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at different threshold settings for the 2014-2016 test set of funds. The diagonal red line shows the results of random guesses for different thresholds. The blue line represents the discrimination power of the algorithm for different thresholds.

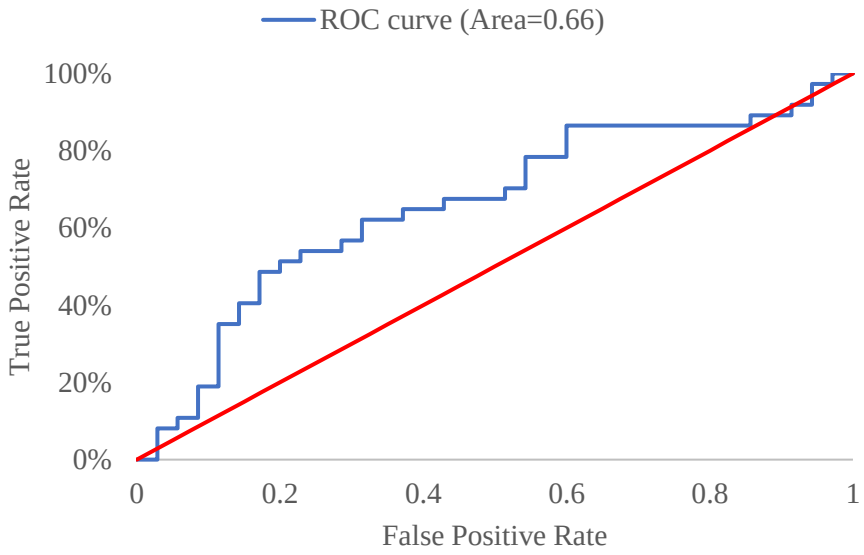


Figure 3. Performance of the Selected Sub-samples of Funds

This figure plots the average TVPI of sub-samples of funds raised between 2014 and 2016 according to three different selection criteria. The lines plot the average performance of the sub-samples of funds sorted by: the predicted probability of outperformance generated by Gradient Boosting (solid line); the inverse of the number of months to raise the fund (dotted line); and past performance (dashed line). The x-axis shows the fraction of funds eliminated from each sub-sample: it starts at 0% (i.e., all funds are included) and goes up to 75% (i.e., only the top quartile of funds in terms of each of the three selection criteria is included). The y-axis shows the average TVPI for each sub-sample of funds.

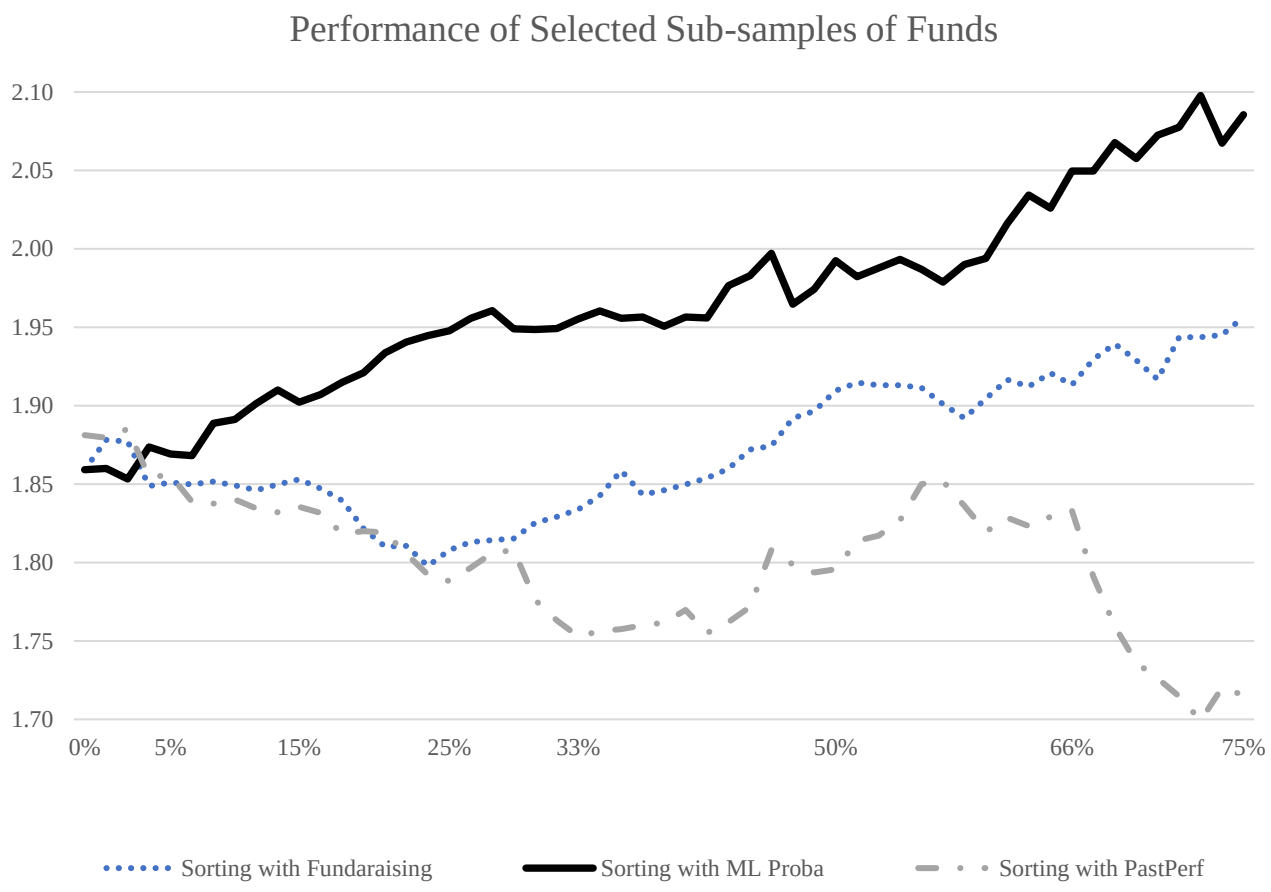
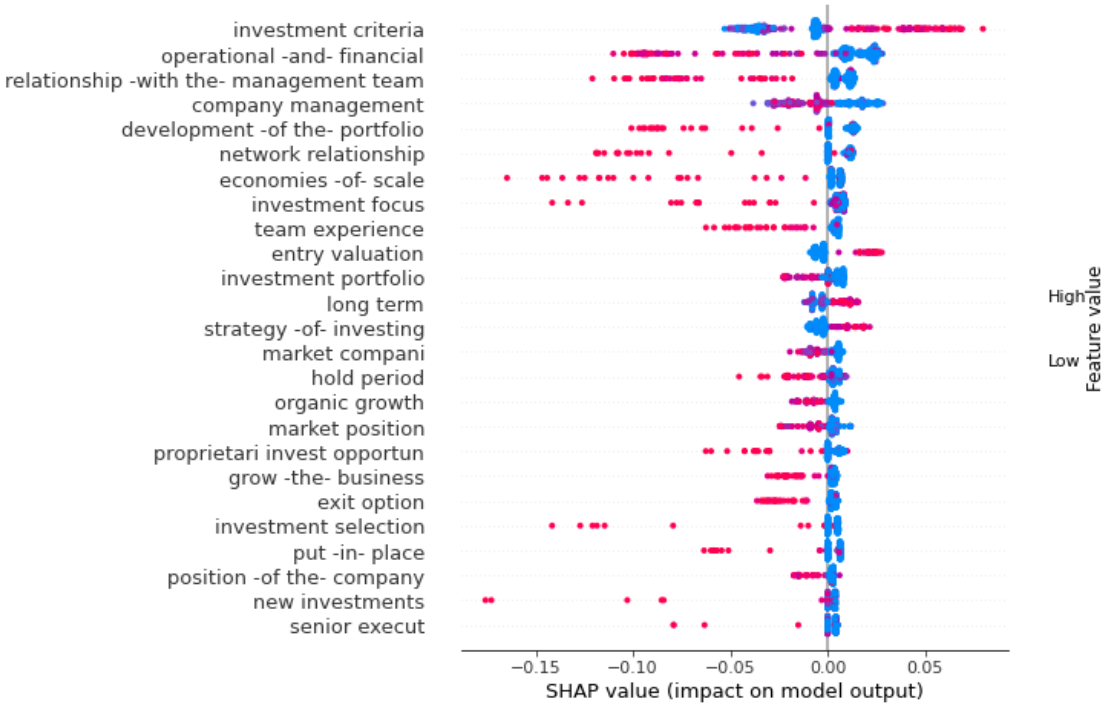


Figure 4. SHAP values of the most relevant terms

This figure reports the SHAP values for the most important terms (by median absolute value) predicting outperformance using the Gradient Boosting algorithm. Gradient Boosting is trained on funds from 2003 to 2013 with performance measures as of 2022. Panel A shows the importance of a term (value on x-axis) and the direction of the effect for each PPM in our training sample. Terms are ranked in decreasing order of importance. The sign (i.e., negative or positive) is captured by the x-axis. In Panel A, red (blue) coloring indicates a high (low) TF-IDF raw value. Panel B shows the median SHAP value for the same list of terms. The terms are ranked according to their median value. In Panel B, terms in blue (red) have a positive (negative) median value.

Panel A. Term importance and effect for each PPM.



Panel B. Absolute median term importance across all PPMs.

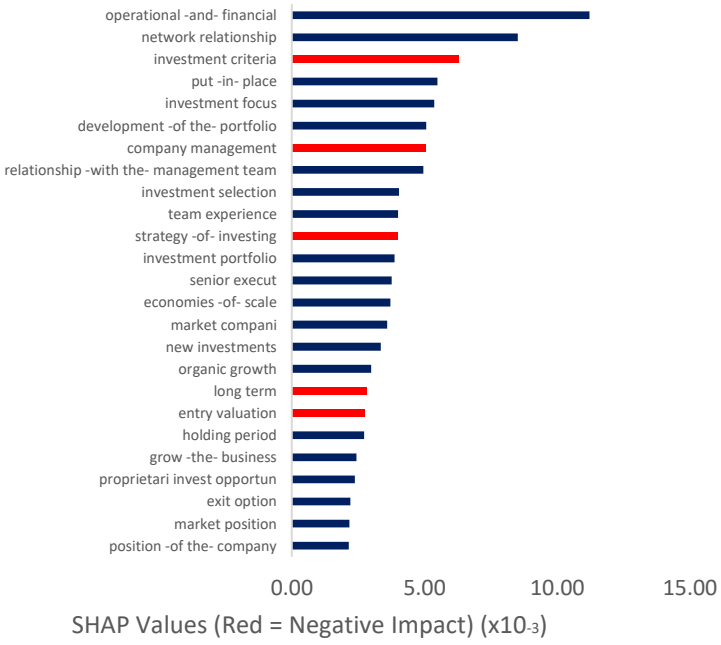
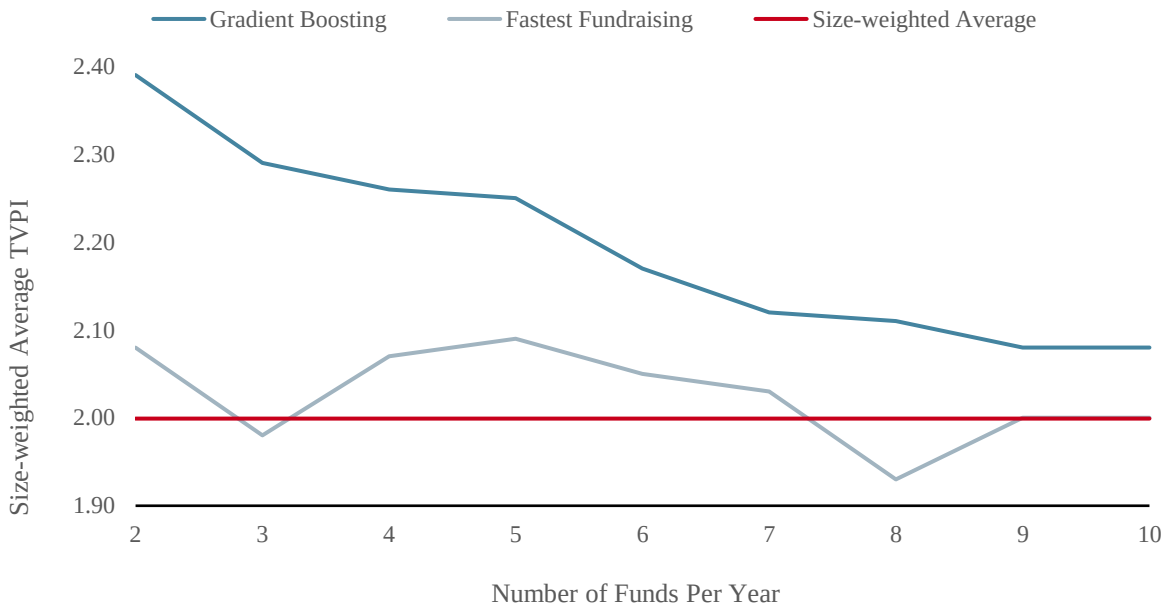


Figure 5. Top funds selected by Gradient Boosting versus Market selection

This figure plots the size-weighted average TVPI of portfolios composed of the top two to ten funds per year selected by the Gradient Boosting and by the number of fundraising months. The x-axis shows the numbers of funds included per year. The blue line depicts the size-weighted average TVPI when funds are sorted by their probability of outperformance generated by the Gradient Boosting algorithm. The algorithm is trained using the 122 funds from 2003 to 2007 with performance information available as of December 31st 2013. The gray line shows the size-weighted mean TVPI when funds are sorted by the inverse of the number of months it took to fundraise. The red line depicts the size-weighted average TVPI of the 72 funds raised between 2014 and 2016.



Appendix

Table A1 Variable definitions

Target Variables	Definition
Ultimate Total Value to Paid-In (TVPI)	Ratio of all capital distributions plus the last reported Net Asset Value to the total amount of capital invested (including fees).
Excess TVPI	Difference between the TVPI and the median TVPI of funds raised in the same year and with the same investment type with performance information available in Preqin
Oversubscription	Ratio of target size divided by actual fund size
Months in Fundraising (ln)	Natural log of the number of months that have passed between first and final closing of a fund
Soft Information Variables	**Definition**
Number of Strategy words (ln)	Natural log of the number of words in the PPM Strategy section
Number of PPM Pages (ln)	Natural log of the number of pages in the PPM
Salient Information Variables	**Definition**
Preceding Fund Performance	Lagged gross TVPI (excluding fees) of a given private equity firm's previous fund as of the fundraising date
Fund Size (ln)	Natural log of the amount of capital a fund has under management (in millions of Euros)
Fund Sequence (ln)	Natural log of the sequence number of a fund for a certain investment strategy by a PE firm
Probabilities	**Definition**
Gradient – Salient Information	Probability predicted by Gradient Boosting trained using the following variables: Fund Sequence (ln), and Fund Size (ln), including dummies capturing investment type and geographic focus fixed effects
Gradient – TF-IDF	Probability predicted by a Gradient Boosting trained using the TF-IDF scores of the PPM Strategy section to predict if TVPI>=Benchmark TVPI, where the benchmark is the median TVPI of funds raised in the same year and with the same investment type

Table A2. Converting all TVPIs into Euro

We compare TVPI in USD with TVPI in Euro. First, we calculate cash flows and unrealized values in USD and Euro for each fund with the entire history of cash flows available in Preqin. Then, we compute the TVPI in both currencies and calculate the ratio of the TVPI in USD to the TVPI in Euro. Next, we compute the median of that ratio for funds raised in the same vintage year.

We observe substantial differences over time. For example, for the 2007 vintage, the median ratio is 0.88. The magnitude of these ratios outlines the necessity of having performance data in a single currency to compare performance across funds. Because most of the funds in our sample are Europe-focused, we use the Euro.

For the funds that are not in Euro, we implement the following process. First, for every fund with the entire history of cash flows in Preqin, we can simply calculate the TVPI by converting each cash flow using the exchange rate prevailing at the time. When we do not have the detailed cash flows, we use the median ratio of USD to Euro TVPI in the same vintage year to obtain the TVPI in Euro.

Table A3. Details about Machine learning algorithms

Lasso Regression. Lasso Regression is an extension of logistic regression, a probabilistic linear model that uses a logistic sigmoid function to return a probability value. Lasso Regression, unlike logistic regression, includes a regularization penalty, the so-called L1 norm, to the loss function. Given an example i with a vector of features $x^{(i)}$ and an output $y^{(i)}$, the elastic regression solves the following equations:

$$z^{(i)} = w^T x^{(i)} + b \quad (1)$$

$$\hat{y}^{(i)} = sigmoid(z^{(i)}) \quad (2)$$

$$\text{Where, } sigmoid\left(z^{(i)}\right) = \frac{1}{1+e^{-z^{(i)}}}$$

$$\mathcal{L}(\hat{y}^{(i)}, y^{(i)}) = -y^{(i)} \log\left(\hat{y}^{(i)}\right) - (1 - y^{(i)}) \log\left(1 - \hat{y}^{(i)}\right) \quad (3)$$

The overall cost is computed as follows:

$$\mathrm{J} = \sum_{i=1}^{m} \mathcal{L}(\hat{y}^{(i)}, y^{(i)}) + \delta \left(\sum_{j=1}^{n} |w_j| \right)$$

Where j refers to the number of features, and δ denotes the amount of shrinkage. Note that for $\delta = 0$, Lasso Regression computes the cost function of logistic regression.

Random Forest. Tree-based method that randomly creates and merges multiple individual decision trees. To create these individual decision trees, we use bootstrapping. Each decision tree is implemented as a tree of binary decision nodes where each node compares one feature value of the sample to a threshold. The feature and the threshold are selected by comparing the Gini impurity of a random subset of features. The Gini impurity is calculated as follows:

$$G = \sum_{h=1}^{C} p(h) + (1 - p(h))$$

Where C is the total number of classes and $p(i)$ is the probability of picking a datapoint with class i. The final prediction is the most highly voted predicted variable. The random forest algorithm takes an average of predictions from all the decision trees.

Gradient Boosting. Like Random Forest, Gradient Boosting is a tree-based method that randomly creates and merges multiples decision trees. The key difference with the Random Forest is that the final prediction is a linear sum of all trees and the goal of each tree is to minimize the residual error of previous trees.

All algorithms are implemented using the Sklearn package.

Table A4. The determinants of predicted probability of success

This table presents results from OLS regressions on the sample of 376 funds raised between 2003 and 2016. In Panel A the dependent variable is the Probability of Success predicted by Lasso. In Panel it is the Probability of Success predicted by Random Forest We control for investment strategy, region, and vintage year effects in all regressions. Standard errors are in parentheses and clustered at the PE firm level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively. Definitions provided in the Appendix Table A1.

Panel A: Lasso

Dependent Variable: Probability of Success predicted by the Lasso algorithm						
	(1)	(2)	(3)	(4)	(5)	(6)
Fund Size (ln)	0.007	0.006	0.006	0.011	0.010	0.008
	(0.018)	(0.021)	(0.021)	(0.022)	(0.024)	(0.024)
Fund Sequence (ln)		0.006	0.006	0.008	0.021	0.018
		(0.038)	(0.038)	(0.038)	(0.051)	(0.050)
Number of words on strategy (ln)			0.000	0.015	0.004	-0.000
			(0.038)	(0.039)	(0.044)	(0.045)
Number of PPM Pages (ln)				-0.056	-0.068	-0.065
				(0.058)	(0.065)	(0.066)
Preceding Fund Gross TVPI					-0.026	-0.025
					(0.040)	(0.041)
Months in Fundraising (ln)						0.001
						(0.003)
Other control variables						
Investment type	Yes	Yes	Yes	Yes	Yes	Yes
Targeted region	Yes	Yes	Yes	Yes	Yes	Yes
Vintage year	Yes	Yes	Yes	Yes	Yes	Yes
Observations	376	376	376	376	309	309
R-squared	0.055	0.055	0.055	0.057	0.078	0.077

Panel B: Random Forest

Dependent Variable: Probability of Success predicted by the Random Forest algorithm						
	(1)	(2)	(3)	(4)	(5)	(6)
Fund Size (ln)	0.009*	0.010*	0.010*	0.011*	0.011*	0.011*
	(0.005)	(0.005)	(0.005)	(0.006)	(0.007)	(0.006)
Fund Sequence (ln)		-0.002	-0.002	-0.001	0.005	0.005
		(0.009)	(0.009)	(0.009)	(0.014)	(0.013)
Number of words on strategy (ln)			-0.006	-0.001	-0.003	-0.006
			(0.010)	(0.010)	(0.012)	(0.012)
Number of PPM Pages (ln)				-0.019	-0.014	-0.014
				(0.017)	(0.020)	(0.019)
Preceding Fund Gross TVPI					-0.005	-0.003
					(0.009)	(0.009)
Months in Fundraising (ln)						0.001
						(0.001)
Other control variables						
Investment type	Yes	Yes	Yes	Yes	Yes	Yes
Targeted region	Yes	Yes	Yes	Yes	Yes	Yes
Vintage year	Yes	Yes	Yes	Yes	Yes	Yes
Observations	376	376	376	376	309	309
R-squared	0.076	0.076	0.078	0.081	0.102	0.100

Table A5. Fundraising Success and Predicted Probability of Success

This table presents results from OLS regressions on the sample of 376 funds raised between 2003 and 2016. The dependent variable is fundraising success (in months needed to close a fund). In Panel A the independent variable of interest is the Probability of Success predicted by Lasso. In Panel B it is the Probability of Success predicted by Random Forest. We control for investment strategy, region, and vintage year effects in all regressions. Standard errors are in parentheses and clustered at the PE firm level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively. Definitions provided in the Appendix Table A1.

Panel A: Lasso

Dependent Variable: Natural logarithm of the number of months it took to raise the fund					
	(1)	(2)	(3)	(4)	(5)
Lasso – Probability	-0.016	-0.014	-0.014	-0.002	0.004
	(0.134)	(0.132)	(0.132)	(0.131)	(0.144)
Fund Size (ln)	-0.101***	-0.065	-0.071*	-0.101**	-0.096**
	(0.038)	(0.043)	(0.043)	(0.042)	(0.044)
Fund Sequence (ln)		-0.144*	-0.128	-0.139*	-0.252*
		(0.086)	(0.084)	(0.083)	(0.130)
Number of words (strategy) (ln)			0.190**	0.107	0.059
			(0.075)	(0.081)	(0.085)
Number of PPM Pages (ln)				0.328**	0.371**
				(0.140)	(0.157)
Preceding Fund Gross TVPI					-0.146**
					(0.073)
Other control variables					
Investment type	Yes	Yes	Yes	Yes	Yes
Targeted region	Yes	Yes	Yes	Yes	Yes
Vintage year	Yes	Yes	Yes	Yes	Yes
Observations	376	376	376	376	309
R-squared	0.084	0.092	0.107	0.121	0.142

Panel B: Random Forest

Dependent Variable: Natural logarithm of the number of months it took to raise the fund					
	(1)	(2)	(3)	(4)	(5)
Random Forest – Probability	0.484	0.477	0.517	0.585	0.437
	(0.757)	(0.740)	(0.738)	(0.718)	(0.731)
Fund Size (ln)	-0.106***	-0.069	-0.076*	-0.107**	-0.101**
	(0.039)	(0.043)	(0.044)	(0.043)	(0.045)
Fund Sequence (ln)		-0.144*	-0.127	-0.138*	-0.254*
		(0.086)	(0.084)	(0.083)	(0.131)
Number of words (strategy) (ln)			0.193**	0.108	0.061
			(0.075)	(0.080)	(0.085)
Number of PPM Pages (ln)				0.339**	0.377**
				(0.141)	(0.158)
Preceding Fund Gross TVPI					-0.144**
					(0.073)
Other control variables					
Investment type	Yes	Yes	Yes	Yes	Yes
Targeted region	Yes	Yes	Yes	Yes	Yes
Vintage year	Yes	Yes	Yes	Yes	Yes
Observations	376	376	376	376	309
R-squared	0.087	0.095	0.111	0.125	0.145

Random Forest Appendix Tables

Table A6. Predicting Fund Performance

This table presents the goodness of fit of Random Forest. In-sample fit refers to the goodness of fit of the training sample. The test sample consists of funds raised between 2014 and 2016. In Panel A, fund performance is measured by TVPI as of June 2022 for the funds in the training and test sample. In Panel B, fund performance is measured by TVPI as of June 2022 for the funds in the test sample, and is measured by TVPI as of December 2013 for the funds in the training sample.

Panel A: Pseudo out of sample training on

	1999-2013 funds	2003-2013 funds
In-sample Fit:		
Area Under Curve	0.61	0.62
Balanced Accuracy	0.57	0.58
Pseudo Out-of-sample Fit:		
Area Under Curve	0.57	0.63
Balanced Accuracy	0.54	0.62

Panel B: Pure out of sample training on

	1999-2007 funds	2003-2007 funds
In-sample Fit:		
Area Under Curve	0.58	0.59
Balanced Accuracy	0.55	0.52
Pseudo Out-of-sample Fit:		
Area Under Curve	0.55	0.56
Balanced Accuracy	0.52	0.51

Table A7. Predicted Probability of Success by terciles

This table shows the pseudo out-of-sample accuracy of Random Forest by terciles. The test sample consists of funds raised between 2014 and 2016. In Panel A, algorithms are trained using fund performance measured by the TVPI as of June 2022. In Panel B, algorithms are trained using fund performance measured by the TVPI as of December 2013. For each tercile, the table shows the average probability of success, the average TVPI, the mean Excess TVPI, and the percentage of outperforming funds. Fund performance is measured by TVPI as of June 2022 for all funds. We perform statistical tests to see whether funds in the high predicted probability of success tercile ultimately outperformed those in the low predicted success tercile: T-tests for mean TVPI and Excess TVPI, respectively, Wilcoxon rank-sum tests for median TVPIs, and Chi-square tests for shares of under- and outperformers, respectively. *, **, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

Panel A: Pseudo out of sample

	Predicted Probability of Success			TVPI				Excess TVPI	Outperformers
	Mean	Min	Max	Mean	Median	SD	Bottom 25% (Perc.)	Mean	Percentage
Total	0.52	0.32	0.79	1.86	1.77	0.52	25%	0.06	51%
Low	0.44	0.32	0.48	1.68	1.69	0.43	42%	-0.12	38%
Medium	0.51	0.49	0.53	1.89	1.76	0.44	25%	0.10	50%
High	0.60	0.54	0.79	2.01**	1.92**	0.62	8%***	0.21**	67%**

Panel B: Pure out of sample

	Predicted Probability of Success			TVPI				Excess TVPI	Outperformers
	Mean	Min	Max	Mean	Median	SD	Bottom 25% (Perc.)	Mean	Percentage
Total	0.47	0.23	0.63	1.86	1.77	0.52	25%	0.06	51%
Low	0.39	0.23	0.44	1.82	1.73	0.73	29%	0.01	42%
Medium	0.46	0.44	0.49	1.86	1.83	0.40	25%	0.06	54%
High	0.54	0.50	0.63	1.90	1.79	0.35	21%	0.11	58%

Table A8. Other ML Predictions

The table shows the goodness of fit of Random Forest. In-sample fit refers to the goodness of fit of the training sample. The test sample consists of the funds raised between 2014 and 2016. Fund performance is measured by TVPI as of June 2022 for all funds. Fundraising Speed is measured by the number of months needed to reach the final closing of the fund. In the second column, algorithms are trained on 295 funds using quantitative (salient) information as features to predict outperformance (i.e., past performance, fund size, and fund sequence). In the third column, algorithms are trained on 303 funds using TF-IDF vectors as features to predict Fundraising Speed.

	Predicting Fund Performance with Quantitative Information	Predicting Fundraising Speed using Qualitative Information
In-sample Fit:		
Area Under Curve	0.55	0.64
Balanced Accuracy	0.52	0.58
Pseudo Out-of-sample Fit:		
Area Under Curve	0.52	0.57
Balanced Accuracy	0.47	0.54

Figure A1. Fivefold cross-validation

The figure depicts the implementation process of fivefold cross-validation, a resampling procedure to evaluate machine learning models. First, the model is trained with observations included in Folds 2 to 5 and tested on Fold 1. The process is repeated until the five folds are used in the test sample. The tuning parameters that correspond to the highest average of the five AUC scores is retained for the test sample.

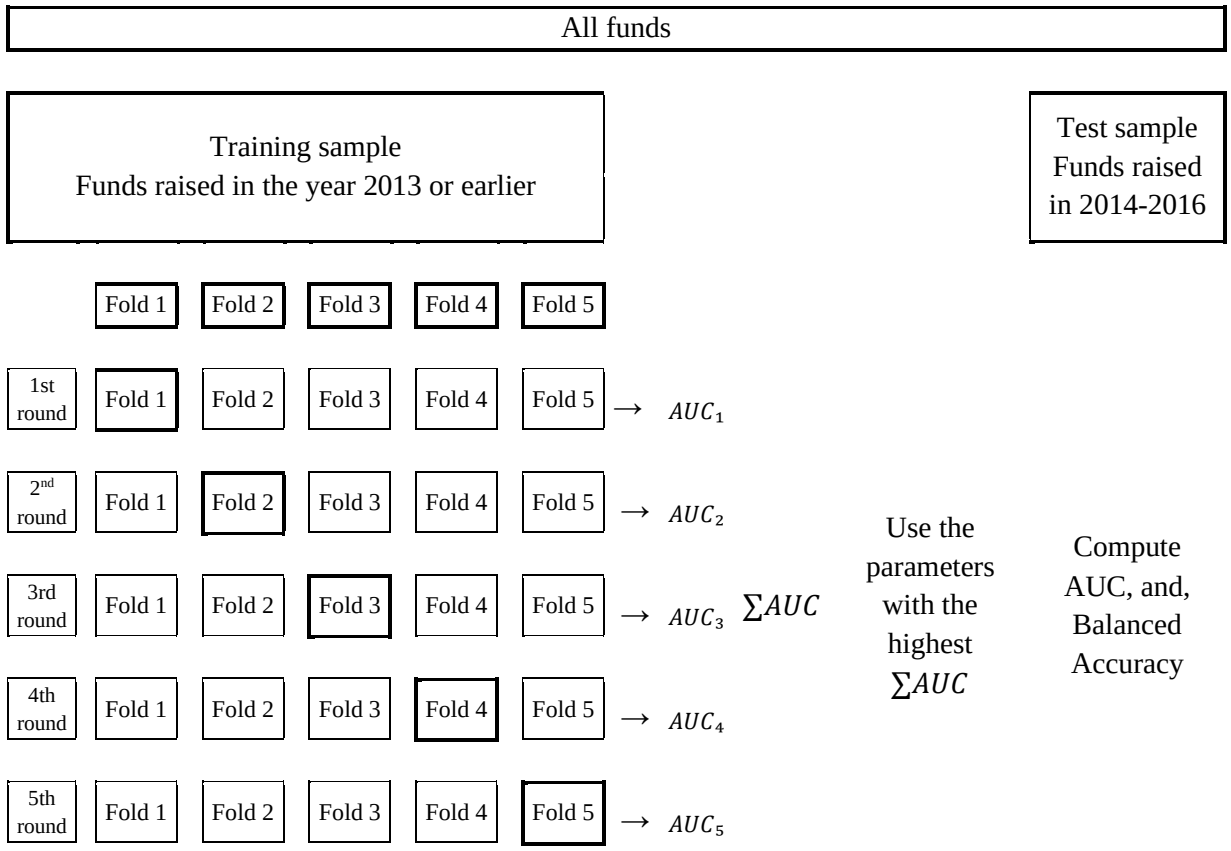


Figure A2. Receiver Operating Characteristic curve using Gradient Boosting

The figure shows the Receiving Operating Characteristic curve (ROC) using the Gradient Boosting algorithm. Gradient Boosting is trained using funds from 2003-2007 with performance information available as of December 31st 2013. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at different threshold settings for the 2014-2016 test set of funds. The diagonal red line shows the results of random guesses for different thresholds. The blue line represents the discrimination power of the algorithm for different thresholds. The figure is different from Figure 3 in the paper since this figure is for the back-test.