Selecting Private Equity Funds Using Machine Learning

Reiner Braun[1], Borja Fernandez Tamayo[2], Florencio López-de-Silanes[3],
Ludovic Phalippou[4], and Natalia Sigrist[5]

Prospective investors in Private Equity (PE) funds receive a large amount of non-standardized and qualitative information about fund manager investment strategies. Using a novel and proprietary sample of 380 Private Placement Memoranda, we combine for the first time Natural Language Processing techniques and Machine Learning algorithms to forecast PE fund success based on investment approach descriptions. Our findings suggest that these novel statistical techniques help select PE fund managers. Their increased usage should lead to more efficient private markets.

Keywords: private market efficiency, private equity, fund performance, performance predictability, Natural Language Processing, machine learning

[1] Technische Universität München (TUM), TUM School of Management. Email at reiner.braun@tum.de
[2] Université Côte d'Azur (UCA), SKEMA Business School. Email at borja.fernandeztamayo@skema.edu
[3] Université Côte d'Azur (UCA), SKEMA Business School. 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

1. Introduction

The value of assets under management by Private Equity (PE) funds has increased tenfold over the last two decades, from $576 billion in 2000 to $4.5 trillion in 2019. These funds have a 10 to 12-year limited life and raise all the capital at inception. The illiquidity of these funds and the long-term lock-in period paired with substantial heterogeneity in fund performance highlight the importance of General Partner (GP) selection in PE[6]. Investors deploy substantial economic and human resources when conducting the due diligence of PE funds[7]. Despite those resources, investors focusing on Leverage Buy-outs (LBOs) are unable to choose funds that will outperform[8].

This paper considers a potential alternative approach to selecting PE funds: one that combines Natual Language Process (NLP) techniques and machine learning algorithms that rely on descriptions of investment approaches to predict GP quality. GPs communicate their investment approach to prospective investors through the Private Placement Memorandum (PPM). These documents are non-standardized, not regulated, and long (82 pages, 38956 words on average). Furthermore, there are no rating agencies or other mechanisms to reduce this information or make it comparable. The reasons why the combination of NLP and machine learning algorithms are promising in the context of PE are two-fold. First, NLP techniques, such as Term Frequency- Investment Document Frequency (TF-IDF), transform the substantial amount of unstructured, textual data contained in PPMs into numerical vectors seen as regressors. Second, unlike traditional econometrics, machine learning algorithms can be used as forecasting models in contexts where the number of features exceeds the number of observations. Moreover, these algorithms make out-of-sample accurate predictions due to their ability to handle multiple complex non-linear interactions. We rely on machine learning algorithms fed with TF-IDF-produced regressors for these two reasons to identify GP quality.

We construct a database of 380 PPM sent between 1996 and 2014 to a Fund-of-Fund focusing on European LBOs. This unique dataset of PPMs is provided by a large institutional investor and contains both funds it invested into and those it rejected. Performance information is from a public dataset maintained by Preqin (as in Chung et al., 2012; Barber and Yasuda, 2017)

[6] See Metric and Yasuda (2011) and Kaplan and Sensoy (2015) for surveys about fund performance in Private Equity.
[7] See Da Rin and Phalippou (2017) for a survey about due diligence practices of limited partners.
[8] Harris et al., (2018) examine the returns of funds of funds investing in Private Equity. Their results suggest that investors focusing on LBOs are unable to identify and access superior performing funds.

and internal documents gathered by our data provider. We employ several machine learning algorithms designed to predict GP quality. We identify GP quality as the ability to outperform the median Total Value to Paid-In (TVPI) of funds part of Preqin and raised in the same vintage year, investing in the same type of investment and geographic location. We complement this binary indicator of GP quality with a continuous variable calculated as the difference between a fund's TVPI and the median TVPI of its peers.

We employ several machine learning algorithms to generate an out-of-sample performance forecast: they are trained on PE funds raised between 1996 and 2011 and tested on funds raised between 2012 and 2014 (with performance measured as of June 2020).

We find that these algorithms make accurate out-of-sample predictions of GP quality. Algorithms document Accuracy rates above 0.5 (e.g., pure randomness), highlighting the selective power of these algorithms. To understand the economic implications of selecting funds with our machine learning-based approach, we compute and compare the size-weighted mean TVPI of funds predicted to succeed and fail. We find that funds predicted to succeed (fail) deliver, on average, high (low) returns.

Because investors might not have enough capital to allocate to all funds predicted to succeed, we construct portfolios composed of the top and bottom five funds per year according to the *Predicted Probability of Success*. Then, to benchmark the performance of these portfolios, we simulate a distribution of 1000 portfolios investing in the same number of funds per vintage year from our test set. We find that the size-weighted mean TVPI of machine learning-selected portfolios investing in the top five funds lies above the 90^{th} percentile for all algorithms. On the other hand, the machine learning-selected portfolios lie below the 20^{th} percentile when investing in the bottom five funds per year. These findings suggest that machine learning algorithms are slightly more suitable to select than to deselect funds.

We ensure that the partition of the sample, in terms of vintage years, does not drive the out-of-sample Accuracy of the machine learning algorithms. Because investors at the beginning of 2012 do not observe a reliable performance indicator for those funds raised in recent years, we restrict the training sample to funds raised between 1996 and 2007. The backtest aims at presenting the consequences of using machine learning to predict performance in real life.

We find that the predictive power of machine learning algorithms is slightly reduced but remains economically significant. All algorithms document a size-weighted mean TVPI higher for funds predicted to succeed than for those predicted to fail.

Because our sample is mainly composed of funds focusing on LBOs, we evaluate the predictive power of the algorithms when restricting the sample to LBO funds. We find that that the out-of-sample predictive power of machine learning algorithms is robust to the investment type. Furthermore, we find that LBO funds predicted to succeed (fail) deliver, on average, high (low) returns.

We test whether the machine learning-generated *Predicted Probability of Success* contributes to explaining cross-sectional variations in fund performance. Regardless of fund performance metrics used as the independent variable, including machine learning-generated *Predicted Probability of Success* as a regressor leads to a better explanation of cross-sectional variation in fund performance. Moreover, the coefficient of the *Predicted Probability of Success* is statistically and economically significant across all model specifications. For example, a 1% increase in the *Predicted Probability of Success* is associated with a 42 basis point increase in TVPI.

Machine learning algorithms are often considered black-boxes because of the difficulty of understanding how variables are combined to make predictions. However, an emerging strand of the machine learning literature is evolving to develop techniques that improve model interpretability (e.g., Lundberg and Lee 2017). We employ methods from this literature to identify which combinations of words are, on average, more relevant to predict GP quality. Among others, we find that investment approaches that include "potential buyers" and "best practices" are positively associated with GP quality.

The benefits of machine learning techniques in finance and accounting have been examined for different contexts. These techniques are shown to help predict stock price movements (Ke, Kelly, and Xiu, 2019), corporate fraud (Purda and Skillicorn, 2015), select corporate directors (Erel et al., 2021), and measure corporate culture (Li et al., 2020). Our paper explores the benefits of using machine learning techniques to contribute to the literature on the selection of PE funds (Kaplan and Schoar, 2005; Robinson and Sensoy, 2013; Harris, Jenkinson, and Kaplan, 2014; and Barber and Yasuda, 2017 among others).

The paper is organized as follows. After this introduction, we describe the data collection and sample characteristics in Section 2. Section 3 discusses the methodology applied to use PPM text to predict GP quality. Section 4 presents the statistical and economic power of our machine learning approach to select funds. A series of robustness checks are presented in Section 5. Section 6 explores the relation between the probability of success predicted by the

algorithms and GP quality. The combinations of words that most contribute to predicting GP quality are shown in Section 7. Finally, section 8 summarizes the paper and discusses the benefits of machine learning to select PE funds.

2. Constructing a Sample on which Algorithms Can Select Funds

2.1. The sample

When PE firms raise funds, they provide a private placement memorandum (PPM) to potential investors (e.g., limited partners, LPs) to provide them with all the information deemed relevant for their investment decision. PPMs are not publicly available. We source them from a large global institutional investor based in Europe and known for focusing on European Leveraged Buy-Out (LBO) funds. This proprietary database consists of 941 PPMs submitted to the investor between 1996 and 2019. Panel A of Table 1 presents the filters applied to our initial sample.

Table 1: Sample construction

The table describes the sample decomposition from the initial sample to the sample used in the empirical analysis. Panel A describes the filters applied to our initial sample to attain the final Private Placement Memoranda (PPM) sample with performance available. Panel B presents the sources used to collect the EUR-denominated performance of the funds included in the sample. Finally, panel C shows the amount of PPMs containing the following sections: Market Opportunity, Investment Highlights, and Investment Strategy and Processes.

	Number of Private Placement Memoranda (PPM)
Panel A: Sample Decomposition	
Initial sample of funds raised between 1996 and 2020	941 (100%)
Funds raised in the year 2014 or earlier	646 (68.65%)
Standard private equity funds raised in the year 2014 or earlier	488 (51.86%)
Standard private equity funds raised in the year 2014 or earlier and investing in Europe, North America, or Asia	486 (51.65%)
Funds with Total Value to Paid-In (TVPI) and/or Internal Rate of Return (IRR) at least six years later than the vintage year	380 (40.38%)

Panel B: Performance in USD (out of 380)	**TVPI or IRR (% over 380)**	**TVPI**	**IRR**
Data provider performance sample	151 (39.74%)	151	151
Preqin Cash Flows sample	47 (12.37%)	47	47
Preqin Performance sample	123 (32.36%)	118	107
Internal Sources Performance	59 (15.53%)	56	38
TVPI to IRR Formula		8	37

Panel C: PPM sections (out of 380)	**Market Opportunity**	**Investment Highlights**	**Investment Strategy & Processes**
PPM containing corresponding section	307 (80.79%)	375 (98.68%)	377 (99.21%)

As we study the relationship between the PPM content of a fund and its eventual performance, we restrict the sample to funds raised in the year 2014 or earlier. From this sample of 646 PPMs, we exclude not standard private equity funds and a few funds that focus on emerging markets.[9] Of the remaining 486 funds, our data provider invested in 151 funds. As our data provider invests in those funds, we have the complete time-series of cash flows and Net Asset Values of these funds.

Panel B of Table 1 shows the different sources used to collect fund performance. Of the 335 funds our data provider did not invest into, 93 are present in the Preqin dataset. However,

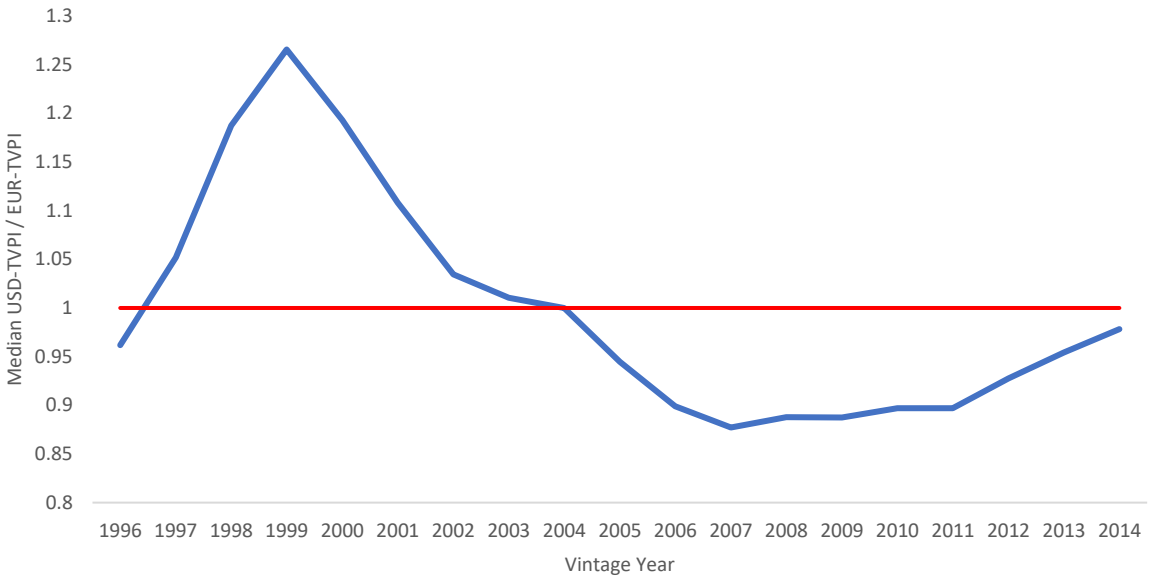
[9] Using Preqin classification, excluded funds belong to one of the following categories: Natural Resources, Special Situations, Secondaries, Distressed Debt, Co-Investment, Mezzanine, Infrastructure, Direct Secondaries, Venture Debt, Fund of Funds, Real Estate. Included funds are: Buy-Out, Balanced, Venture Capital, Turnaround, and Growth Capital.

only 47 funds have a complete time-series of cash flows and Net Asset Values. Of the remaining funds, 130 funds are present in the Preqin Performance summary dataset. Of the remaining 185 funds, 63 funds have a performance summary reported in some other internal documents of our data provider (e.g., PPMs of subsequent funds). The rest of the funds (N=96) have no performance information available. Finally, we remove 12 funds for which we do not observe performance, at least six years later than the vintage year (e.g., immature funds).

Preqin does not report performance metrics of all funds in a single currency but uses the currency reported by the source without making any conversion. Performance thus is available in different currencies[10].

Figure 1. Comparative USD-denominated and EUR-denominated Total Value to Paid-In (TVPI)

The figure compares the median Total Value to Paid-In (TVPI) denominated in USD and EUR across vintage years. The TVPI is computed for each fund with the complete time-series of cash flows, and Net Asset Values in Preqin Fund Cash Flows dataset.



To understand whether having performance available in different currencies prevents us from fairly compare ultimate performance across funds, we compute and compare the median TVPI achieved by a USD-denominated investor with that achieved by an investor operating in EUR. First, we calculate cash flows and unrealized values in USD and EUR for each fund with the entire history of cash flows available in Preqin. Then, we compute the TVPI in both currencies

[10] For details on how Preqin collects, validates, and documents fund performance: https://docs.preqin.com/pro/Private-Capital-Performance-Guide.pdf.

and calculate the USD-denominated TVPI over the EUR-denominated TVPI. Next, we compute the median of that ratio for funds raised in the same vintage year.

Figure 1 shows the median USD-denominated TVPI over the median EUR-denominated TVPI from 1996 to 2014[11]. The graph highlights substantial differences between the median USD-denominated TVPI and the median EUR-denominated TVPI across vintage years. For example, in 2007, the median USD-denominated TVPI to the median EUR-denominated TVPI is approximately 0.88. The magnitude of these ratios outlines the necessity of having performance data in a single currency to compare performance across funds fairly. Because most of the funds in our sample are Europe-focused, we use EUR-denominated fund performance in the analyses presented below.

First, we compute the TVPI and IRR in EUR for the 198 funds for which we have the entire history of cash flows since inception. Preqin Performance summary contains the TVPI and IRR denominated in EUR for 59 and 52 funds, respectively. 42 and 28 funds out of the funds whose performance recovered through internal documents have TVPI, and IRR, respectively, reported in EUR. The remaining 81 and 102 have no TVPI and IRR, respectively, available in EUR.

Out of the 81 funds without TVPI available in USD, we have 73 funds with TVPI denominated in any of the following nine currencies: USD, GBP, NOK, SEK, DKK, CAD, NZD, JPY, or INR. First, for every fund with the entire history of cash flows in Preqin, we calculate the TVPI in EUR and each of those nine currencies as previously done with USD in Figure 1 (see above). Then, we compute the ratio EUR-TVPI over the TVPI calculated using each of the nine currencies. Finally, we use the median of the ratio per vintage year to convert the TVPI of the 73 funds into EUR.

We apply the same procedure to the 102 funds with IRR denominated in other currencies than EUR. To maximize our sample and have TVPI and IRR in EUR for the 380 funds, we apply a formula to estimate the TVPI from the IRR and vice-versa. The TVPI of a fund with a known IRR can be approximated using the following formula:

$$LN(TVPI) = 4 * LN(1 + IRR)$$

Thus, our final sample consists of 380 funds for which we have both the TVPI and IRR in EUR.

[11] Appendix Figure A1 shows the analogous graph to Figure 1 but using the mean TVPI of the vintage year rather than the median TVPI.

Our study focuses on the qualitative information given about the fund strategy: investment approach in terms of sourcing deals, monitoring and adding value, and approach to exiting deals. This information is contained mainly in the Investment Strategies and Processes section. In addition, most PPMs also include an Investment Highlights section on which the GP summarizes the key reasons why her offer is attractive (e.g., investment approach, track record, management team, and market outlook). Furthermore, 80% of the PPMs complement those sections with another section describing the market outlook and the fitness with its investment approach.

The rest of the content of a PPM consists of quantitative information (e.g., past performance, value creation decomposition), other qualitative information (e.g., biographies of fund managers, selected case studies), and a part that is similar across funds (broadly speaking, legal disclaimers).

Panel C of Table 1 presents the number of PPM containing different sections. 377 out of the 380 PPMs contain the Investment Strategy and Process section, while 375 and 375 PPMs include the Investment Highlights and Market Opportunity section, respectively.

The main body of text shows the analyses using Investment Strategy and Process section because this section documents the best results and is the most common across the three sections. Nevertheless, we present the results of analogous analyses using the Investment Highlights and Market Opportunity in Sections 1 and 2, respectively, in the Appendix.

2.2. Measuring GP quality

We transform the TVPI of the 380 funds to binary indicators of GP quality. We identify GP quality as the ability of the fund to outperform a particular benchmark. To compute that benchmark, we rely on the Preqin Performance summary. Even though Preqin provides benchmarks updated every quarter, we compute our customized benchmarks. As mentioned above, Preqin reports performance in different currencies, so benchmarks are computed aggregating funds with performance denominated in different currencies. Following the same approach as above, we approximate the EUR-denominated TVPI for all funds with non-EUR-denominated TVPI available in the Preqin Performance summary. As there is variation in fund performance across vintage years, investment types, and geographic focus, we classify a fund as successful if its TVPI is equal to or above the median TVPI of its Preqin peers satisfying the

following conditions: raised in the same vintage year; invest in the same of companies (LBO, VC, or other PE types); and the exact geographic location (Europe, US, or Asia), and as failure otherwise. A key concern of using a binary indicator to determine GP quality is that funds with observed TVPI close to the threshold receive the same category as those in the distribution's tails. We, therefore, define *Benchmark Distance* as the difference between the observed TVPI and the median TVPI of its Preqin peers[12]. We use the binary indicator of GP quality and *Benchmark Distance* to evaluate the predictive power of machine learning algorithms.

2.3. Summary statistics

In Table 2, we report the average performance of our sample of 380 funds by vintage year and geographic focus. Average performance is weighted by the capital committed to each fund. Columns 1-4 present the performance of all funds; Columns 5-8 show performance of funds investing in Leveraged Buy-Outs (LBOs); Columns 9-12 document performance of funds investing in Venture Capital (VC); and, Columns 13-16 present performance of other PE funds. We show the average TVPI and IRR and the proportion of funds with *Benchmark Distance* equal to or above zero (see above). The size-weighted average TVPI and IRR over the sample period are 1.73 and 13.95%, respectively. Around 52% of the funds document a negative *Benchmark Distance*.

Most of the funds (74%) focus on LBOs, with the rest of the funds split between VC and Other Private Equity. Regarding fund performance, LBO funds show the highest size-weighted average TVPI and IRR (e.g., 1.74 and 14.19%) than VC and other PE funds.

In terms of vintage years distributions, the number of funds tends to increase over time in the beginning. For example, we have more than ten funds in 2003 for the first time. Thirty funds were raised between 1996 and 2002. As expected, the number of funds peaks in 2006-2008 and then stabilizes post-crisis. In terms of performance, it peaks in 2004-2005 and right after the 2008 crisis: vintage years 2009-2011 have a similar TVPI at about 1.77. These patterns are consistent with what is observed in large datasets and reported in the literature (Harris et al., 2018).

[12] Please note the *Benchmark distance* can take negative and positive values.

In terms of geographic focus, our data provider, being Europe-based, receive more European PPMs. 5.5% of the funds focus on the UK, and 62.6% on the rest of Europe, including Scandinavia (which we denote Europe). 22.4% focus on the US, and 9.5% on Asia.

Table 2. Private Equity Returns by Vintage Year and Geographic Focus

The table presents basic statistics for the entire sample of 380 funds and subsamples classifying funds by investment types (Panel A), fund investment geographies (Panel B), and vintage years (Panel C). The basic statistics are presented for the following two performance metrics: Total Value to Paid-In (TVPI) and Internal Rate of Return (IRR). The statistics include the number of funds, mean, median, standard deviation (SD), and proportion of funds with a negative *Benchmark Distance* defined as the difference between the observed TVPI and the median TVPI of funds sharing the vintage year, investment type, and geographic focus.

	All funds				Buy-out Funds				VC Funds				Other PE Funds			
	Obs	TVPI	IRR	Under-performing Ratio	Obs	TVPI	IRR	Under-performing Ratio	Obs	TVPI	IRR	Under-performing Ratio	Obs	TVPI	IRR	Under-performing Ratio
All funds	380	1.73	13.95	0.52	282	1.74	14.19	0.51	31	1.41	5.26	0.61	57	1.63	13.32	0.47
Panel A: Performance by vintage year																
1996	1	1.36	7.99	1.00	1	1.36	7.99	1.00	0	-	-	-	0	-	-	-
1997	1	0.86	-0.02	1.00	1	0.86	-0.02	1.00	0	-	-	-	0	-	-	-
1998	1	0.36	-9.58	1.00	0	-	-	-	1	0.36	-9.58	1.00	0	-	-	-
1999	4	1.57	12.88	0.25	3	1.70	14.30	0.00	1	0.50	-6.88	1.00	0	-	-	-
2000	7	1.80	15.96	0.29	5	1.81	15.31	0.40	0	-	-	-	2	1.71	9.78	0.00
2001	7	2.42	30.28	0.43	5	2.51	38.85	0.40	2	0.50	-11.14	0.50	0	-	-	-
2002	9	1.83	21.89	0.44	6	1.94	29.85	0.50	3	0.86	-3.07	0.33	0	-	-	-
2003	15	1.67	22.89	0.60	11	1.76	27.99	0.55	4	0.78	-8.61	0.75	0	-	-	-
2004	15	1.87	19.17	0.27	12	1.88	21.76	0.33	2	1.34	4.85	0.00	1	3.29	34.68	0.00
2005	22	1.73	11.48	0.50	16	1.74	11.28	0.56	2	3.84	14.80	0.50	4	1.10	1.81	0.25
2006	43	1.70	9.03	0.44	36	1.69	7.47	0.50	4	1.36	3.83	0.25	3	2.03	11.72	0.00
2007	54	1.65	11.33	0.56	40	1.66	9.50	0.53	6	1.35	5.41	0.50	8	1.57	7.91	0.75
2008	42	1.83	13.51	0.64	33	1.85	10.76	0.67	5	1.34	-1.07	0.80	4	1.49	7.16	0.25
2009	27	1.74	13.31	0.52	21	1.75	11.43	0.52	2	1.58	7.06	1.00	4	1.70	10.81	0.25
2010	23	1.80	14.24	0.48	14	1.83	12.03	0.43	3	1.56	9.56	0.67	6	1.76	13.22	0.50
2011	28	1.77	14.92	0.61	20	1.78	12.16	0.55	2	1.01	0.15	1.00	6	1.89	18.31	0.67
2012	27	1.70	16.94	0.70	19	1.71	15.26	0.74	1	1.74	27.61	1.00	7	1.67	15.05	0.57
2013	33	1.79	22.51	0.39	24	1.81	20.05	0.33	3	2.04	22.53	0.33	6	1.59	23.71	0.67
2014	21	1.57	16.03	0.48	15	1.58	14.36	0.47	0	-	-	-	6	1.38	15.14	0.50
Panel B: Performance by graographic focus																
Europe	259	1.71	13.59	0.49	213	1.71	13.77	0.52	17	1.55	7.69	0.47	29	1.49	10.30	0.31
Rest of Europe	238	1.72	13.78	0.48	192	1.73	14.03	0.51	17	1.55	7.69	0.47	29	1.49	10.30	0.31
United Kingdom	21	1.52	11.42	0.57	21	1.52	11.42	0.57	0	-	-	-	0	-	-	-
US	85	1.76	14.52	0.60	49	1.78	14.64	0.53	21	1.31	3.45	0.67	15	1.71	15.77	0.73
Asia	36	1.58	12.38	0.53	20	1.59	13.23	0.50	3	1.27	3.66	0.67	13	1.57	9.83	0.54

3. Characterizing Investment Strategies with Natural Language Processing Techniques

We find 15326 different words across the 377 Investment Strategy and Process sections, and only 8810 words appear in more than one document. Table 3 shows the thirty most common words, bigrams, and trigrams in the Investment Strategy and Processes section[13]. We show the frequency in the corpus and the number of documents containing it for each word and combination of words. *Investment*, the most common word in our corpus, appears 13025 times and is present in all PPMs. *Portfolio companies* and *due diligence process* are the most common bigram and trigram, respectively. Table 3 summarizes the main topics discussed by GPs when presenting their Investment Strategy and processes. We observe terms associated with the investment cycle: Deal sourcing (e.g., *deal flow, deal team,* and *proprietary deal flow)*, Value Creation (e.g., *value creation, add acquisitions,* and *buy build strategies)*, and Exit (e.g., *exit)*. Terms related to Market Segments and Investment Strategies are also present (e.g., *investment criteria, mid market,* and *medium sized companies)*. We also find terms associated with GP Behaviour - Characteristics (e.g., *decision making process,* and *track record)*.[14]

Table 3 documents variation in investment approaches within the sample. For example, *add acquisitions* and *buy build* only appear in 122 and 101, respectively, documents. These two last terms are helpful to differentiate GPs pursuing a buy-and-build strategy from those implementing other value creation initiatives. Analogously, terms that appear in most documents (e.g., high document frequency), such as *portfolio companies*, are not helpful to discriminate between investment approaches. Following this logic, the most standard method available is that developed by (Salton and Buckley, 1988) called TF-IDF (Term Frequency-Inverse Document Frequency): [15]

$$TF - IDF\ (i,j) = TF\ (i,j)\ x\ IDF(j)$$

Where $TF(i,j)$ is the ratio of the number of times term i occurs in document j to the total number of terms in document j (e.g., the frequency of term i in document j).

[13] *Value creation* is an example of bigram (i.e., two adjacent words) and *Value creation strategy* is an example of trigram (i.e., three adjacent words).

[14] Because the vocabulary partially depends on the type of investment undertaken, we report the thirty most common bigrams in the Investment Strategy and Processes section for the three investment types (See Appendix Table A1)

[15] We use the TF-IDF vectorizer available in Scikit-Learn developed by Pedregosa et al. (2011). These authors modify the TF-IDF formula presented in the body of the text to produce more accurante results. They implement the natural logarithm to the IDF score to avoid high values for this score, preventing them from dominating the TF-IDF score. Furthermore, they normalize the TF-IDF score to make model training less sensitive to the scale of features:

$$TF - IDF\ (i,j) = \frac{TF\ (i,j) x\ LN(\ IDF(j))}{SQRT\ (Sum\ of\ Squares\ of\ the\ product\ of\ TF\ (i,j) x\ LN(\ IDF(j)))}$$

$IDF(j)$ is the ratio of the number of documents in the sample to the number of documents containing the term i at least once. It measures the frequency of term i across documents.

TF-IDF scores characterize investment approaches. Intuitively, the closer TF-IDF scores are between two documents, the more similar the investment approach is.

Table 3: Describing investment approaches

The table shows the thirty most common words, bigrams, and trigrams in our sample. The first three columns show the thirty most common words, the frequency the word appears in the corpus, and the percentage of documents containing the word, respectively. The following three columns show the same descriptives for the thirty most common bigrams (e.g., combinations of two adjacent words). Finally, the last three columns show the same descriptives for the thirty most common trigrams (e.g., combinations of three adjacent words).

	Words			Bigrams			Trigrams		
	Word	**Word Frequency**	**Document Frequency (%)**	**Bigram**	**Bigram Frequency**	**Document Frequency (%)**	**Trigram**	**Trigram Frequency**	**Document Frequency (%)**
1	investment	13025	100%	portfolio companies	2299	91%	due diligence process	348	45%
2	companies	7633	99%	due diligence	1858	85%	portfolio company management	182	27%
3	management	7171	100%	portfolio company	1414	72%	attractive investment opportunities	147	25%
4	team	6030	96%	management team	1321	81%	value creation plan	144	12%
5	company	5912	98%	private equity	1257	78%	proprietary deal flow	143	25%
6	fund	4872	87%	value creation	1202	65%	private equity firms	133	22%
7	portfolio	4775	98%	investment opportunities	1174	81%	middle market companies	110	14%
8	capital	4313	95%	management teams	1054	76%	value portfolio companies	101	21%
9	market	4241	98%	investment team	863	39%	mid market companies	101	13%
10	value	4112	97%	investment committee	841	45%	decision making process	97	17%
11	business	4074	97%	deal flow	759	67%	lower mid market	95	8%
12	investments	3989	98%	investment strategy	728	79%	private equity funds	88	15%
13	growth	3413	93%	investment process	549	62%	fund portfolio companies	76	12%
14	opportunities	3373	98%	long term	462	56%	potential investment opportunities	76	16%
15	equity	2946	92%	deal team	462	27%	company management team	75	14%
16	deal	2867	90%	business plan	420	48%	management portfolio companies	75	18%
17	process	2847	93%	company management	414	53%	buy build strategies	71	10%
18	potential	2725	95%	mid market	408	32%	private equity investment	68	13%
19	strategy	2655	97%	general partner	407	19%	buy build strategy	67	11%
20	exit	2601	90%	diligence process	376	47%	private equity market	66	13%
21	financial	2480	94%	track record	372	51%	value creation strategy	66	9%
22	partners	2459	77%	investment professionals	354	34%	three five years	65	12%
23	industry	2410	90%	add acquisitions	354	32%	fund investment strategy	64	12%
24	due	2244	90%	middle market	336	25%	private equity investors	63	13%
25	strategic	2126	89%	target company	331	40%	value creation potential	63	11%
26	diligence	2039	86%	cash flow	324	44%	company management teams	62	12%
27	also	1941	93%	investment criteria	319	46%	investment strategy fund	60	15%
28	experience	1916	90%	target companies	318	41%	private equity investments	59	12%
29	focus	1820	93%	buy build	309	27%	non core assets	56	11%
30	key	1736	87%	fund investment	307	38%	medium sized companies	55	11%

See Appendix Section A3 for discussion of the implementation details.

4. Evaluation Machine Learning Predictions of Fund Performance

4.1. Model specification

We split our sample of PPMs into two samples by vintage year. Although the nature of the data does not allow for pure (non-overlapping) out-of-sample tests, we nonetheless use funds raised between 1996 and 2011 as a training sample and use funds raised between 2012 and 2014 for the out-of-sample tests.

As shown above, we have 15326 different words and many more bigrams and trigrams. In this case, standard regression techniques cannot be used. The Statistics literature has proposed four main methods in this context: *Naïve Bayes, Lasso, Random forest,* and *Gradient Boosting* (Hastie, Tibshirani, and Friedman, 2009)[16].

An intuitive way to evaluate the quality of a model predicting two classes is to calculate the proportion of correct instances the model forecasts (e.g., Accuracy). However, since our sample is not perfectly balanced (e.g., the proportion of classes is not precisely 0.5), we also calculate the Balanced Accuracy. Balance Accuracy is the average of the proportion corrects of each class individually:

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

Where true positive (TP) are correct predictions of success, true negative (TN) are correct predictions of failure, false negatives (FN) are incorrect predictions of success, and false positives (FP) are inaccurate classifications of failures.

We complement the Accuracy scores with the Area under the Receiver Operating Characteristic curve (ROC AUC). This metric indicates the probability that the model ranks a random positive example more highly than a random negative example. For example, a model whose predictions are 100% wrong has an AUC of 0, while one whose predictions are 100% correct has an AUC of 1. For each fund in the test set, our method produces a probability of outperforming the benchmark threshold computed using Preqin (hereafter, *Predicted Probability of Success*). As mentioned above, we classify a fund as successful if the predicted probability is higher than 50%. Otherwise, we classify the fund as a predicted failure.

[16] The algortihms are defined in Appendix Section A4

We use stratified K-fold cross-validation for each of our four models, a resampling procedure to evaluate machine learning models, to estimate the unknown tuning parameters. Cross-validation is the most widely used method for estimating prediction error. We set K=5 as results do not suffer from much bias for samples with at least 200 observations (Hastie, Tibshirani, and Friedman, 2009)[17].

4.2. Predictions of fund performance

Table 4 summarizes the statistical ability of the machine learning models, once trained on the earlier portion of the sample, to classify funds in the later part (e.g., funds raised in 2012-2014).

Table 4. Statistical assessment of machine learning to predict fund performance

This table reports the out-of-sample statistical performance of several algorithms in the 2012-2014 test set. Algorithms are trained on the 1996-2011 training set. Area Under the Receiver Operating Characteristic curve (AUC ROC) is the probability that a random positive instance is ranked higher than a random negative instance. Accuracy is the proportion of correctly predicted instances. Finally, Balanced Accuracy is the average of the proportions corrects of each class individually.

	Lasso	Random Forest	Gradient Boosting	Naïve Bayesian
Accuracy	0.543	0.568	0.630	0.543
Balanced Accuracy	0.543	0.566	0.629	0.540
AUC ROC	0.564	0.622	0.593	0.534

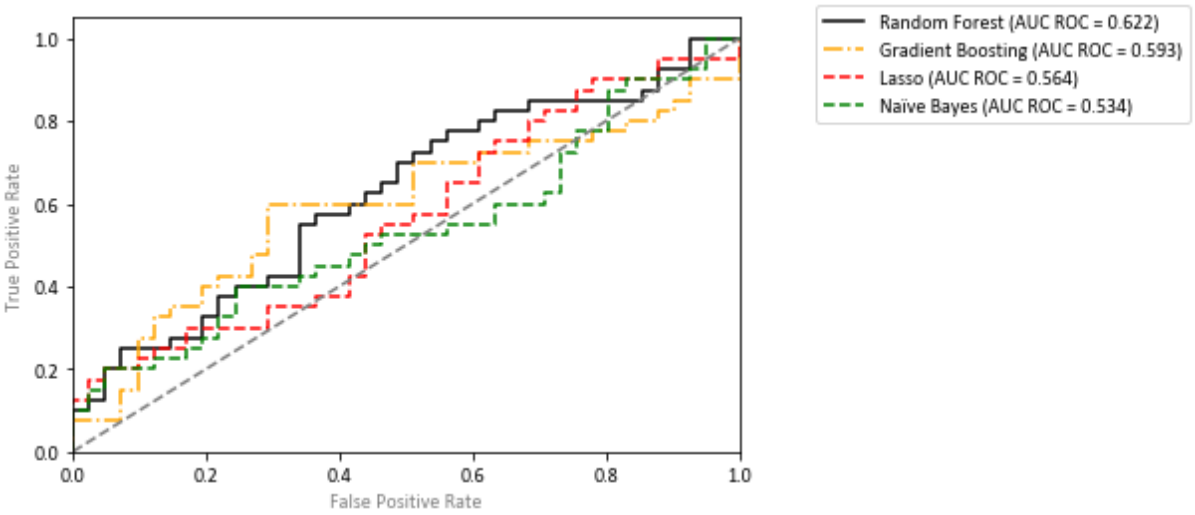
Table 4 indicates that *Gradient Boosting* achieves the best Accuracy among the machine learning algorithms, whereas *Random Forest* yields the best AUC ROC. Nevertheless, the four machine learning algorithms outperform the 0.5 thresholds, representing a higher predictive power than pure randomness.

Figure 2 shows the ROC for the different algorithms. The ROC plots True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings in the 2012-2014 test set. Thus, the figure depicts how machine learning algorithms predict classes across different thresholds. For example, the black line, representing the ROC of Random Forest, suggests that the algorithm presents a higher TPR than the FPR across all thresholds.

[17] See Appendix Figure A2 depicts the implementation process of fivefold cross-validation

Figure 2. Evaluating machine learning algorithms across different thresholds

The figure shows the Receiving Operating Characteristic curve (ROC) that plots True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings in the 2012-2014 test set. The diagonal red line shows random guesses for different thresholds. The blue line represents the algorithm discrimination ability for different thresholds.



We complement Table 4 with a comparison of *Benchmark Distance* between predicted classes[18]. Figure 3 shows the mean *Benchmark Distance* across predicted successful and failure funds for the machine learning algorithms in the 2012-2014 test period[19]. Intuitively, the higher the value of *Benchmark Distance*, the better is the fund performance relative to its peers (e.g., funds raised in the same vintage year and investing in the same types of investment and geographic focus).

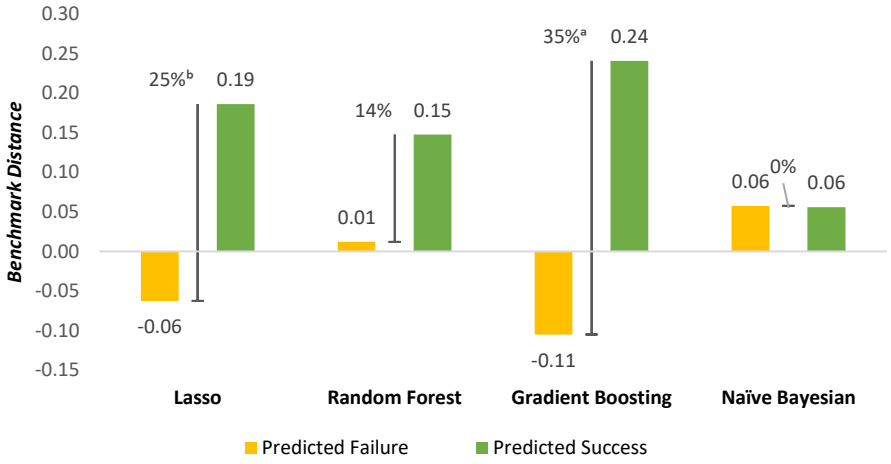
All algorithms show a mean *Benchmark Distance* lower for funds predicted to fail than for funds predicted to succeed. Except for Naïve Bayesian, the Predicted Failure class shows a negative or close to zero mean *Benchmark Distance*. In contrast, the Predicted Success class presents a positive, relatively high mean *Benchmark Distance*. For *Gradient Boosting*, the difference in the mean TVPI between the two predicted classes is statistically significant at the 0.01 level.

[18] Our results remain practically unchanged when we use the median predicted probability as the threshold to classify funds between the two classes (see, e.g., Appendix Figure A3)

[19] To minimize the influence of outliers in our sample, we winsorize the top and bottom 1% of Benchmark Distance and fund performance metrics.

Figure 3. Mean Benchmark Distance for predicted classes

The figure shows the mean *Benchmark Distance* of predicted classes for ML models in the 2012-2014 test set. Algorithms are trained on the 1996-2011 training set. *Benchmark Distance* is defined as the difference between observed Total Value to Paid-In (TVPI) and the median TVPI of Preqin funds sharing the vintage year, investment type, and geographic focus. TVPI is the ratio of all capital distributions plus the last reported Net Asset Value to the total amount of capital invested (including fees). Predicted Success (Predicted Failure) comprises the funds with a *Predicted Probability of Success* equal to or above (below) 0.5. *Benchmark Distance* is winsorized at 1% and 99%. [a], [b], and [c] indicate significance at the 1%, 5%, and 10% levels, respectively, in a two-sample t-test comparing mean *Benchmark Distance* across predicted classes.

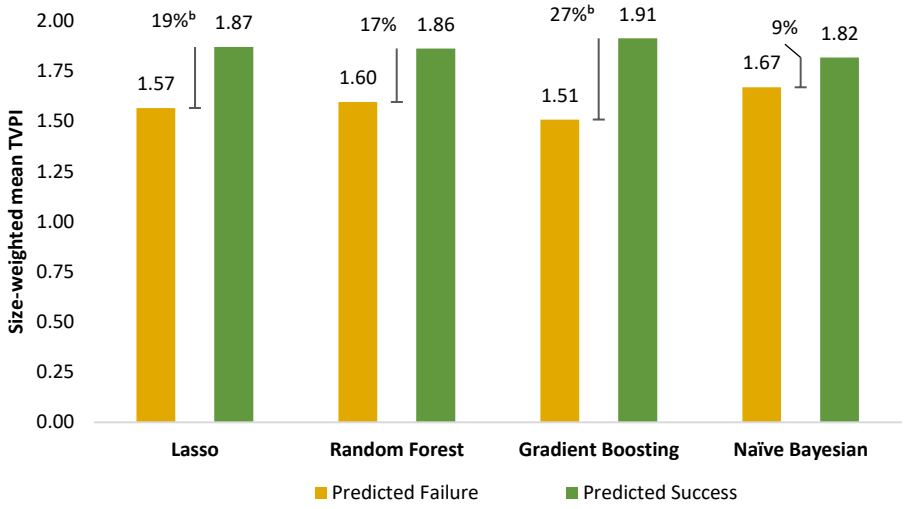


4.3. The economic implications of employing machine learning algorithms

This subsection assesses the economic implications of using machine learning algorithms to select and deselect funds. To this end, we compute and compare the size-weighted average TVPI of funds Predicted Failures and Predicted Success classes. Figure 4 presents the average weighted-size average TVPI across the two predicted classes for the machine learning algorithms in the 2012-2014 test period. In line with Figure 3, the figure documents that the weighted-size average TVPI is higher for the Predicted Success class than for the Predicted Failures class. *Gradient Boosting* yields the largest delta between the successful and failure funds, where delta is the difference in average TVPIs between the two predicted classes as a percentage of the average TVPIs of one predicted class. Funds predicted to success by *Gradient Boosting* have a size-weighted average observed TVPI of 1.91, which is 27% higher than the size-weighted average observed TVPI for funds predicted to fail.

Figure 4. Observed size-weighted mean TVPI of predicted failures and successes

The figure shows the size-weighted mean Total Value to Paid-In (TVPI) of predicted classes for ML models in the 2012-2014 test set. Algorithms are trained on the 1996-2011 training set. TVPI is the ratio of all capital distributions plus the last reported Net Asset Value to the total amount of capital invested (including fees). Predicted Success (Predicted Failure) comprises the funds with a *Predicted Probability of Success* equal to or above (below) 0.5. TVPI is winsorized at 1% and 99%. [a], [b], and [c] indicate significance at the 1%, 5%, and 10% levels, respectively, in a two-sample t-test comparing mean TVPIs across predicted classes.



4.4. Comparing the Performance of Machine Learning-Selected Portfolios to that of Potential Alternative Choices

Our results document that funds identified by machine learning algorithms as likely to succeed (fail) are more likely to deliver high (low) TVPI returns. One concern with Figure 4 is that investors might not have enough capital to commit to all funds with a *Predicted Probability of Success* higher than 0.5. Harris et al. (2018) empirically analyze funds-of-funds (FoF) returns in PE. FoFs are sophisticated investors that raise a fund to invest in several PE funds. In their sample, the average FoF invests in 25 PE funds. Assuming an FoF investment period of five years, this corresponds to five fund investments per year. We, therefore, build portfolios composed of the five funds presenting the highest *Predicted Probability of Success* per vintage year in our test sample (2012-2014). Then, we compute the size-weighted average observed TVPI return to this portfolio of fifteen funds. Next, we randomly draw five funds from the same initial sample with PPMs closed each year of the test set and compute the size-weighted average TVPI return to such a portfolio of fifteen funds. We repeat this last step one thousand times (with replacement), which produces a distribution of fund portfolio returns. Table 5 exhibits the results from this procedure using the top and bottom five funds predicted by machine learning algorithms in the test sample.

Table 5. Evaluating the relative performance of the top and bottom five funds per year

The table shows how machine learning algorithms-selected portfolios rank in a randomly computed distribution of portfolios in the test set. Algorithms are trained on the 1996-2011 training set. For the vintage years in our test sample (e.g., from 2012 until 2014), we pick the top and bottom five funds per year according to the *Predicted Probability of Success* and compute the size-weighted average Total Value to Paid-In (TVPI) of this fund portfolio (e.g., 15 funds in total). Next, we simulate benchmark portfolios by drawing bootstrap samples, e.g., random samples with replacement, from the same test sample of funds closed in the vintage year in our test sample. We run one thousand simulations (for each portfolio size scenario) and compute the size-weighted mean TVPIs of the portfolio. TVPI is winsorized at 1% and 99%.

	Lasso	Random Forest	Gradient Boosting	Naïve Bayesian
Bottom five funds per year	14th	10th	18th	17th
Top five funds per year	99th	97th	91st	91st

Table 5 documents that machine learning algorithms perform slightly better at selecting than deselecting funds. For example, the *Lasso*-selected portfolio of fifteen funds with the highest *Predicted Probability of Success* lies on the 99th percentile. In contrast, the portfolio formed by the fifteen funds with the lowest *Predicted Probability of Success* lies on the 29th.

5. Alternative Subsamples

5.1. Backtesting Machine Learning algorithms

A key concern is that the predictive power shown above is achieved by the algorithms trained on funds raised between 1996 and 2011 and evaluated on funds raised between 2012 and 2014. A realistic assumption in this exercise is that at the end of each year in the test sample, say 2012. A decision-maker can observe, rank, and select based on predictive probabilities of success for funds raising at this point. However, the algorithm computing these predictions is partially trained on information not available as of 2012. For the later vintages in the training sample, no reliable signal on fund performance is available yet, because funds were still highly unrealized. To alleviate this concern, we restrict our training sample to funds raised in 2007 or earlier. Then, we test the predictive power on funds raised in 2012 and 2014 (e.g., the same test sample used in previous sections).

Table 6. Statistical assessment of machine learning to predict fund performance in the backtest

This table reports the out-of-sample statistical performance of several algorithms in the 2012-2014 test set. Algorithms are trained on the 1996-2007 training set. Area Under the Receiver Operating Characteristic curve (AUC ROC) is the probability that a random positive instance is ranked higher than a random negative instance. Accuracy is the proportion of correctly predicted instances. Finally, Balanced Accuracy is the average of the proportions corrects of each class individually.

Metric	Lasso	Random Forest	Gradient Boosting	Naïve Bayesian
Accuracy	0.556	0.580	0.568	0.531
Balanced Accuracy	0.555	0.582	0.567	0.532
AUC ROC	0.539	0.607	0.588	0.565

Figure 5. Mean Benchmark Distance for predicted classes

The figure shows the mean *Benchmark Distance* of predicted classes for ML models in the 2012-2014 test set. Algorithms are trained on the 1996-2007 training set. *Benchmark Distance* is defined as the difference between observed Total Value to Paid-In (TVPI) and the median TVPI of Preqin funds sharing the vintage year, investment type, and geographic focus. TVPI is the ratio of all capital distributions plus the last reported Net Asset Value to the total amount of capital invested (including fees). Predicted Success (Predicted Failure) comprises the funds with a *Predicted Probability of Success* equal to or above (below) 0.5. *Benchmark Distance* is winsorized at 1% and 99%. [a], [b], and [c] indicate significance at the 1%, 5%, and 10% levels, respectively, in a two-sample t-test comparing mean *Benchmark Distance* across predicted classes.

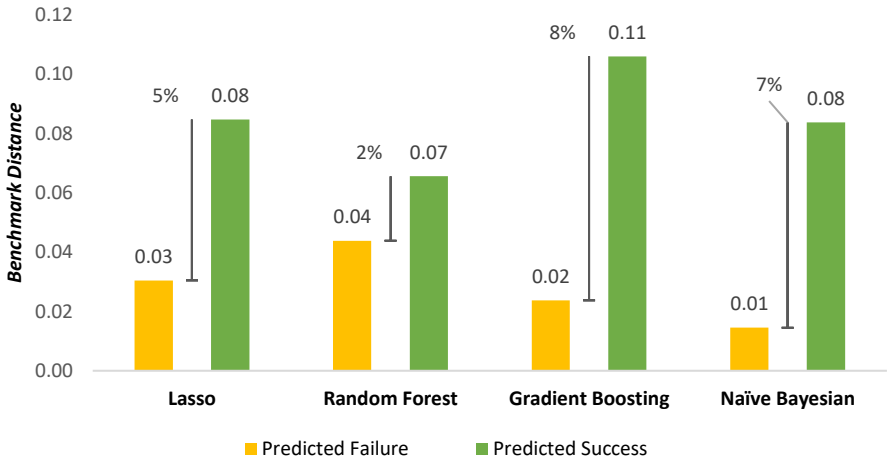


Table 6 reports the statistical ability of the machine learning models, once trained on funds raised in 1996-2007, to classify funds in the later part. Again, all algorithms document metrics above the 0.5 thresholds. *Random Forest* presents the best Accuracy and AUC across all algorithms. Figures 5 and 6 show the mean *Benchmark Distance* and size-weighted mean TVPI, respectively, of the predicted classes for the machine learning algorithms. The Predicted Success class shows a higher mean *Benchmark Distance* and size-weighted mean TVPI than the Predicted Failure class. While the difference in the size-weighted TVPI between predicted classes is not significant for any of the algorithms, the results are economically significant. For

example, the size-weighted mean TVPI of the Predicted Class is 11% (e.g., 1.81/1.63 - 1) higher than that of the Predicted Failure Class.

Figure 6. Observed size-weighted mean TVPI of predicted failures and successes in the backtest

The figure shows the size-weighted mean Total Value to Paid-In (TVPI) of predicted classes for ML models in the 2012-2014 test set. Algorithms are trained on the 1996-2007 training set. TVPI is the ratio of all capital distributions plus the last reported Net Asset Value to the total amount of capital invested (including fees). Predicted Success (Predicted Failure) comprises the funds with a *Predicted Probability of Success* equal to or above (below) 0.5. TVPI is winsorized at 1% and 99%. [a], [b], and [c] indicate significance at the 1%, 5%, and 10% levels, respectively, in a two-sample t-test comparing mean TVPIs across predicted classes.

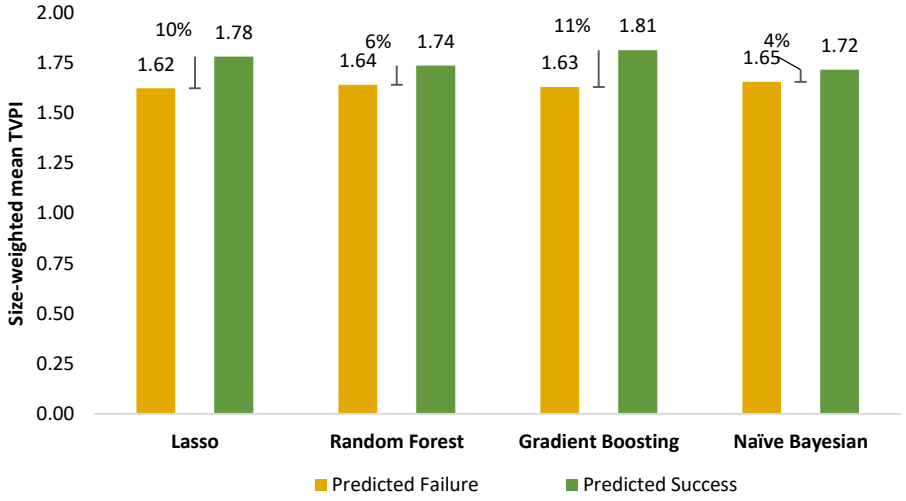


Table 7. Statistical assessment of machine learning to predict fund performance in the backtest

The table shows how machine learning algorithms-selected portfolios rank in a randomly computed distribution of portfolios in the test set. Algorithms are trained on the 1996-2007 training set. For the vintage years in our test sample (e.g., from 2012 until 2014), we pick the top and bottom five funds per year according to the *Predicted Probability of Success* and compute the size-weighted average Total Value to Paid-In (TVPI) of this fund portfolio (e.g., 15 funds in total). Next, we simulate benchmark portfolios by drawing bootstrap samples, e.g., random samples with replacement, from the same test sample of funds closed in the vintage year in our test sample. We run one thousand simulations (for each portfolio size scenario) and compute the size-weighted mean TVPIs of the portfolio. TVPI is winsorized at 1% and 99%.

	Lasso	Random Forest	Gradient Boosting	Naïve Bayesian
Bottom five funds per year	64th	56th	55th	56th
Top five funds per year	69th	93rd	94rd	69th

Finally, we conduct the same type of analysis presented in Table 5 to assess the relative performance of portfolios investing in the top and bottom five funds per year. Table 7 presents

the percentiles on which the machine learning-selected portfolios lie. We observe that the portfolios formed by the best funds selected by *Random Forest* and *Gradient Boosting* lie above the 90th percentile. However, any of the algorithms perform well at deselecting funds.

5.2. Predictions of Leverage Buy-Outs funds

The sample used in previous sections includes LBOs, VC, and other types of PE funds (see Table 2). However, as the vocabulary varies across investment types and certain words are investment type-specific (see Table 4), the algorithms might be capturing investment types instead of investment approaches. To rule out this possibility, we examine the statistical and economic performance of the algorithms when restricting the sample to LBOs. Thus, our training sample consists of LBO funds raised between 1996 and 2011, and our out-of-sample is composed of LBO funds raised between 2012 and 2014. Table 8 presents the statistical power of the machine learning algorithms when restricting the sample to LBO funds.

Table 8. Statistical assessment of machine learning to predict LBO fund performance

This table reports the out-of-sample statistical performance of several algorithms using only LBO funds in the 2012-2014 test set. Algorithms are trained on the 1996-2011 training set of LBO funds. Area Under the Receiver Operating Characteristic curve (AUC ROC) is the probability that a random positive instance is ranked higher than a random negative instance. Accuracy is the proportion of correctly predicted instances. Finally, Balanced Accuracy is the average of the proportions corrects of each class individually.

Metric	Lasso	Random Forest	Gradient Boosting	Naïve Bayesian
Accuracy	0.586	0.621	0.569	0.534
Balanced Accuracy	0.586	0.625	0.575	0.544
AUC ROC	0.554	0.579	0.615	0.506

Figure 7. Mean Benchmark Distance for predicted classes using LBO funds

The figure shows the mean *Benchmark Distance* of predicted classes for ML models in the 2012-2014 test set using only LBO funds. Algorithms are trained on the 1996-2011 training set of LBO funds. *Benchmark Distance* is defined as the difference between observed Total Value to Paid-In (TVPI) and the median TVPI of Preqin funds sharing the vintage year, investment type, and geographic focus. TVPI is the ratio of all capital distributions plus the last reported Net Asset Value to the total amount of capital invested (including fees). Predicted Success (Predicted Failure) comprises the funds with a Predicted Probability of Success equal to or above (below) 0.5. *Benchmark Distance* is winsorized at 1% and 99%. [a], [b], and [c] indicate significance at the 1%, 5%, and 10% levels, respectively, in a two-sample t-test comparing mean Benchmark Distance across predicted classes.

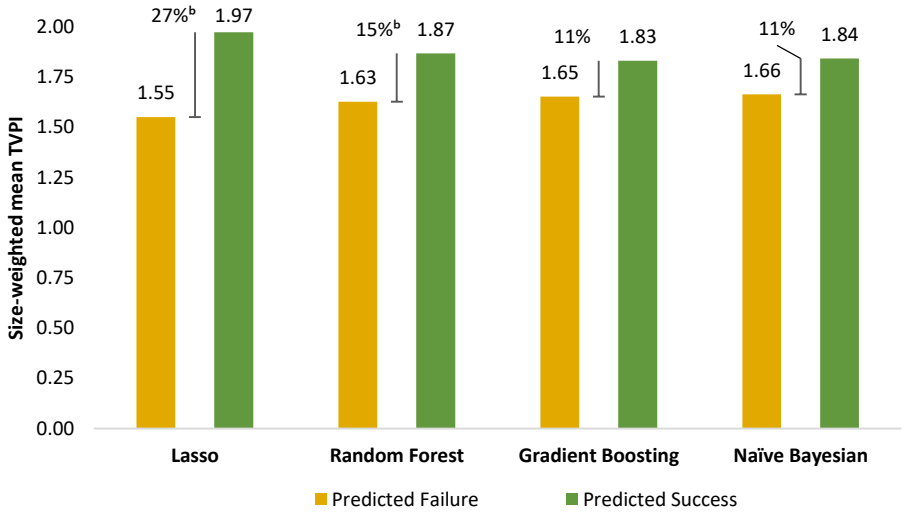
All statistical metrics are above the 0.5 thresholds. *Gradient Boosting* and *Random Forest* document the highest AUC ROC and Accuracy, respectively. Figure 8 presents the comparison between the mean TVPI across predicted classes when restricting the sample to LBO funds.

Except for Naïve Bayesian, all algorithms document a higher mean *Benchmark Distance* for the Predicted Success class than for the Predicted Failure class. For example, in the case of *Random Forest*, the *Benchmark Distance* of the Predicted Success class is 34% above the mean *Benchmark Distance*. This difference in means is statistically significant at the 0.01 level.

The economic implications of using machine learning algorithms to select and deselect LBO funds align with those presented above. Figure 8 presents the average weighted-size average TVPI across the two predicted classes for the machine learning algorithms in LBO funds' 2012-2014 test period. The delta between predicted classes is economically significant for all algorithms. For example, the *Lasso*-selected group of funds likely to succeed documents a size-weighted average TVPI 27% higher than the size-weighted average TVPI of the other predicted class.

Figure 8. Observed size-weighted mean TVPI of predicted failures and successes in the LBO sample

The figure shows the size-weighted mean Total Value to Paid-In (TVPI) of predicted classes for ML models in the 2012-2014 test set using only LBO funds. Algorithms trained on the 1996-2011 training set of LBO funds. TVPI is the ratio of all capital distributions plus the last reported Net Asset Value to the total amount of capital invested (including fees). Predicted Success (Predicted Failure) comprises the funds with a *Predicted Probability of Success* equal to or above (below) 0.5. TVPI is winsorized at 1% and 99%. [a], [b], and [c] indicate significance at the 1%, 5%, and 10% levels, respectively, in a two-sample t-test comparing mean TVPIs across predicted classes.



Finally, we evaluate the relative performance of portfolios formed by the top and bottom five funds per vintage year according to the *Predicted Probability of Success*. In line with previous analogous analyses in this paper, we find that machine learning algorithms perform better at selecting rather than deselecting funds. Funds composed by the top five funds per vintage year lie above the 90th percentile for all the algorithms. For example, the *Lasso*-based portfolio of the top five funds lies on the 99th percentile. In contrast, the portfolio of the bottom five funds predicted by the same algorithm lies on the 15th percentile.

Table 9. Statistical assessment of machine learning to predict fund performance in the LBO sample

The table shows how machine learning algorithms-selected portfolios rank in a randomly computed distribution of portfolios in the test set using only LBO funds. Algorithms trained on the 1996-2011 training set of LBO funds. For the vintage years in our test sample (e.g., from 2012 until 2014), we pick the top and bottom five funds per year according to the *Predicted Probability of Success* and compute the size-weighted average Total Value to Paid-In (TVPI) of this fund portfolio (e.g., 15 funds in total). We simulate benchmark portfolios by drawing bootstrap samples, e.g., random samples with replacement, from the same test sample of funds closed in the vintage year in our test sample. We run one thousand simulations (for each portfolio size scenario) and compute the size-weighted mean TVPIs of the portfolio. TVPI is winsorized at 1% and 99%.

	Lasso	Random Forest	Gradient Boosting	Naïve Bayesian
Bottom five funds per year	15th	42nd	38th	16th
Top five funds per year	99th	94th	92nd	94th

6. The predictive power of machine learning algorithms from an econometric perspective

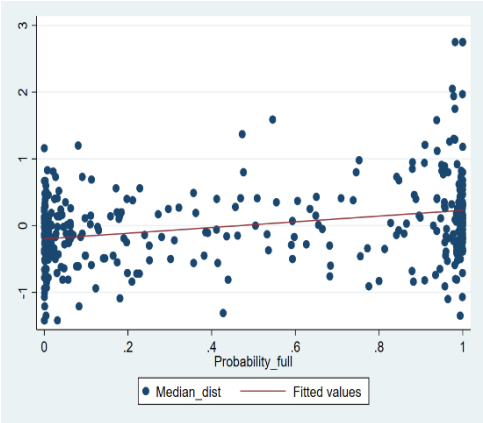
This subsection examines how *Benchmark Distance* and metrics of fund performance correlate with predicted probabilities and fund characteristics. To illustrate this, we use the probabilities predicted by *Gradient Boosting* using the training and test sample. Because we conduct cross-validation with replacement to train the model (see above), we have an out-of-sample *Predicted Probability of Success* for the 377 funds with the investment strategy and processes section.

Figure 9 plots *Benchmark Distance* and metrics of fund performance against *the Predicted Probability of Success* predicted by *Gradient Boosting*. Panel A depicts a significant and positive correlation between *Benchmark Distance* and *Predicted Probability of Success* (e.g., coefficient= 0.29 and p-value: 0.00). Panel B and C show that the *Predicted Probability of Success* is positively associated with TVPI and IRR. Panel B shows a correlation coefficient of 0.27 (p-value: 0.00), and Panel C depicts a correlation coefficient of 0.22 (p-value: 0.00).

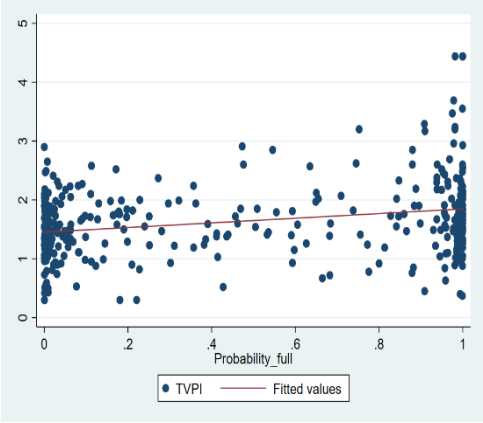
Figure 9. Fund performance and Predicted Probability of Success

The figure presents the relation between GP quality indicators and the probability of success predicted by *Gradient Boosting*. Panel A uses the *Benchmark Distance*; Panel B uses the observed Total Value to Paid-In (TVPI); and Panel C uses the observed Internal Rate of Return (IRR). The three GP quality indicators TVPI are winsorized at 1% and 99%.

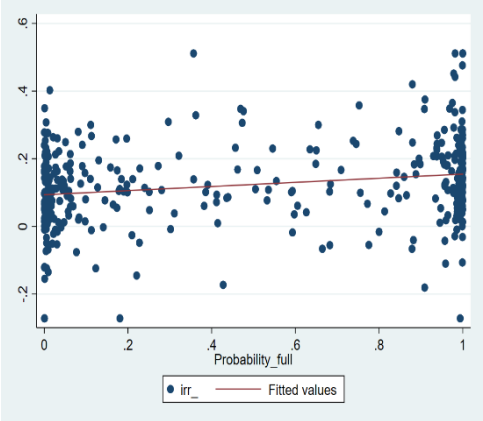
Panel A. Benchmark Distance



Panel B. Total Value to Paid-In



Panel C. Internal Rate of Return



We compare model specifications, including and excluding *Predicted Probability of Success* as an independent variable. Finally, we use Akaike Information Criterion (AIC) and adjusted R^2 as

the criteria to assess whether *Predicted Probability of Success* contributes to a better explanation of cross-sectional observed TVPIs.

Table 10 shows a series of model specifications regressing *Benchmark Distance* and metrics observed performance on fund characteristics and *Predicted Probability of Success*. Columns 1-2 show the cross-sectional relations between *Benchmark Distance* and fund characteristics, with Column 2 including *Predicted Probability of Success* as an independent variable. Columns 3-4 and 5-6 replicate Columns 1-2 using observed TVPI and IRR, respectively, rather than *Benchmark Distance* as the dependent variable.

Table 10. The econometric interpretation of Predicted Probability of Success

The table reports the results from an OLS regression where the dependent variable is Benchmark Distance (Columns 1-2), Total Value to Paid-In (Columns 3-4), or Internal Rate of Return (Columns 5-6). *Benchmark Distance* is the difference between observed TVPI and the median TVPI of Preqin funds sharing the vintage year, investment type, and geographic focus. TVPI is the ratio of all capital distributions plus the last reported Net Asset Value to the total amount of capital invested (including fees). We include funds used in the training and test set. Fund Size is the amount of capital a fund has under management in EUR. First Dummy is equal to 1 if the fund is a first-time fund and 0 otherwise. *Predicted Probability* is the probability of outperforming the median TVPI of the vintage year predicted by *Gradient Boosting*. Standard errors are in parentheses and are adjusted for serial correlation and heteroskedasticity. The three dependent variables are winsorized at 1% and 99%. [a], [b], and [c] indicate significance at the 1%, 5%, and 10% levels, respectively.

	Benchmark Distance		Total Value to Paid-In		Internal Rate of Return	
	(1)	(2)	(3)	(4)	(5)	(6)
Predicted Probability		0.433[a]		0.423[a]		0.059[a]
		(0.073)		(0.074)		(0.016)
Log(Fund Size)	0.751[b]	0.731[b]	0.752[b]	0.732[b]	0.099	0.097
	(0.319)	(0.289)	(0.339)	(0.311)	(0.067)	(0.065)
Log(Fund Size)^2	-3.942[b]	-3.856[b]	-3.824[b]	-3.740[b]	-0.476	-0.465
	(1.667)	(1.506)	(1.786)	(1.638)	(0.357)	(0.349)
First Dummy	-0.012	-0.022	-0.014	-0.024	0.001	0.000
	(0.111)	(0.107)	(0.110)	(0.108)	(0.021)	(0.021)
VC Dummy	-0.193[b]	-0.173[b]	-0.387[b]	-0.368[b]	-0.098[a]	-0.095[a]
	(0.136)	(0.133)	(0.151)	(0.147)	(0.024)	(0.024)
Other PE Dummy	0.162	0.177	-0.038	-0.024	-0.004	-0.002
	(0.104)	(0.094)	(0.102)	(0.093)	(0.020)	(0.019)
Asia Dummy	-0.231[b]	-0.144	-0.016	0.070	-0.021	-0.009
	(0.112)	(0.107)	(0.101)	(0.097)	(0.019)	(0.019)
US Dummy	-0.195[b]	-0.103	-0.002	0.088	-0.006	0.007
	(0.092)	(0.084)	(0.094)	(0.087)	(0.016)	(0.021)
Year F.E.	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.108	0.187	0.103	0.176	0.164	0.203
IC	723.158	690.078	737.911	707.759	-535.878	-551.986
No. Of Observations	377	377	377	377	377	377

Standard errors are in parentheses and are adjusted for serial correlation and heteroskedasticity

Consistent with Kaplan and Schoar (2005), we find a significant concave relation between *Fund Size* and TVPI in Columns 1-4. We also find a concave relation between *Fund Size* and IRR, although not significantly so. The estimates also confirm that VC funds in our sample have

worse performance than LBO and other PE funds. The point estimate on the VC dummy is negative and significant across the six model specifications. Regarding geographic focus dummies, we find that funds investing in Asia and the US present a negative, significant coefficient when using *Benchmark Distance* as the independent variable and not including *Predicted Probability of Success* as an independent variable. The effect of these two dummies capturing geographic focus is not significant in the other model specifications.

Model specifications, including *Predicted Probability of Success* as an independent variable, document better AIC scores and adjusted R^2 than their analogous excluding that variable. Moreover, the estimates in Columns 2, 4, and 6 indicate that funds with a higher *Predicted Probability of Success* have significantly higher returns and higher *Benchmark Distance*. For example, the point estimate of *Predicted Probability of Success* in Column 4 is 0.42 with a standard error of 0.07. The coefficient implies that a fund with a 1% increase in the *Predicted Probability of Success* is associated with a 42 basis point increase in TVPI.

Overall, the evidence we present suggests that out-of-sample probability predicted by *Gradient Boosting* is a powerful feature to explain cross-sectional differences in fund performance and, consequently, contributes to price more efficiently investment proposal (e.g., more efficient markets)

7. Characteristics that Affect Fund Performance

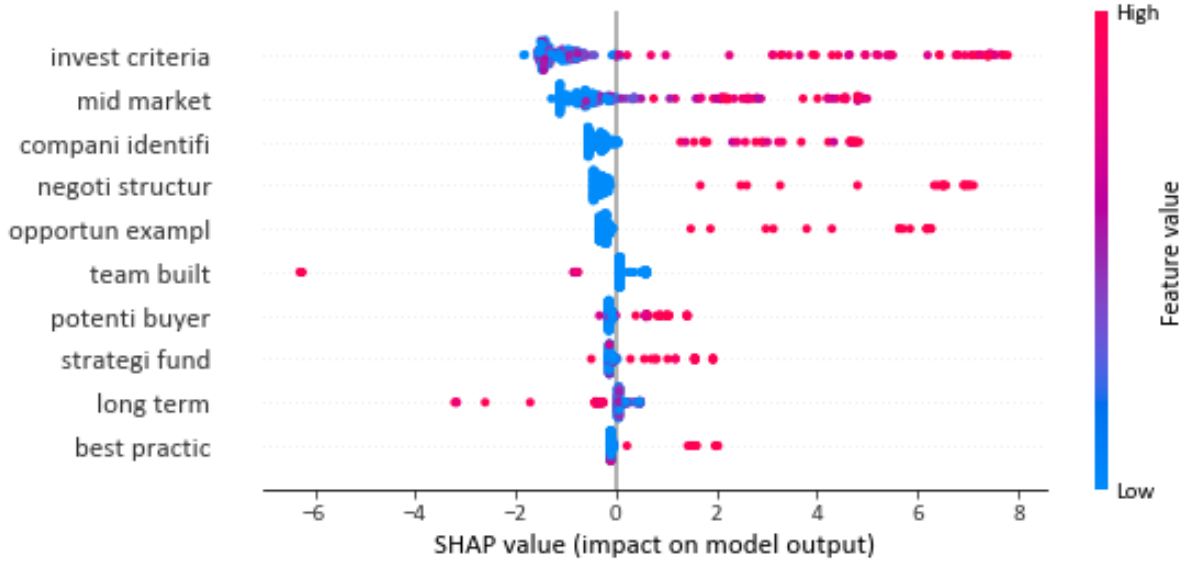
We argue that our approach can capture differences in the investment approaches of GPs and link these to performance. In order to shed some light on the sources of the predictive capability of our machine learning algorithms, we employ methods from the machine learning literature focused on developing methods targeted to increase the interpretability of the underlying patterns governing the predictions of machine learning algorithms (e.g., Lundberg and Lee 2017). We implement SHAP (Shapley Additive exPlanations) method developed by Lundberg and Lee (2017). This approach relies on Shapley Values from coalitional game theory to explain the output of machine learning models. The idea behind Shapley values is to assume that each feature value of the instance is a "player" in a game where the prediction is the payout. Features collaborate with each other to explain the prediction and, based on the effects of such collaborations in the prediction, each feature is assigned a Shapley value that represents its marginal contribution across all possible coalitions. We identify combinations

of words that are relevant in the training process of *Gradient Boosting* with funds raised between 1996 and 2011.

To identify the most relevant word combinations, we first average SHAP values for each feature across observations. Then, we rank features by average SHAP values. Figure 10 shows the ten combinations of words that contribute the most to *GradientBoosting's* predictions of GP quality, measured by average SHAP values. Combinations of words are ordered from top to down according to their feature importance. The SHAP value determines the position on the x-axis, and the feature determines the y-axis. Points represent observations. Reddish points indicate high values, while bluish points indicate low values. *Invest criteria* is the combination that contributes the most to *GradientBoosting's* predictions, and it is positively associated with GP quality. Other combinations of words such as *invest citeria, mid market, negoti structure, opportun exampl, potential buyer, best practi,* and *build strategi* are also positively associated with GP quality. On the other side of the coin, combinations of words such as *team built* and *long term* are indirectly associated with GP quality.

Figure 10. Most relevant combinations of words to make predictions

The figure presents the SHAP values for the top-10 characteristics in terms of variable importance in GP quality. We use the *Gradient Boosting* algorithm in predictions. Combinations of words are ranked in decreasing order according to their importance. The Shapley value determines the position on the x-axis, and the feature determines the y-axis. Points represent observations. The color represents the value of the feature from low (blue) to high (red).

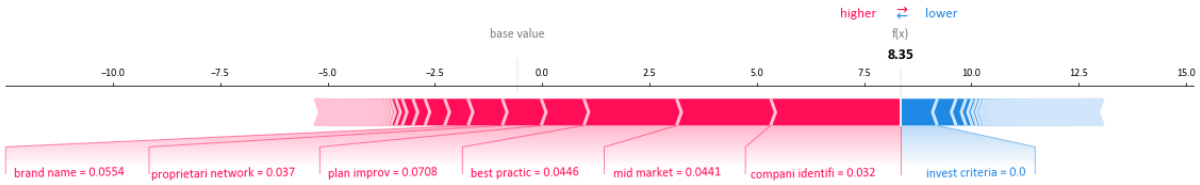


To understand how SHAP values work in an individual observation, we randomly select an investment approach from our sample and compute the SHAP values of the features for that investment approach. Figure 11 shows the features that most contribute to the randomly

selected investment approach description. Combinations of words in red (blue) directly (inversely) influence the prediction of GP quality for this observation. The base value represents the mean of all output values on the model on the training. The model output is the model prediction for this randomly selected example. The arrow's length for each combination of words reflects the SHAP value.

Figure 11. Individual SHAP values for a randomly-selected investment approach

The figure presents the combinations of words that contribute the most to the *Gradient Boosting*-generated predictions for a randomly selected investment approach. For each of the most contributing combinations of words, we show the SHAP value. The color represents the sign of the combination of words' contribution to *GradientBoosting's* output. Reddish (bluish) combination of words directly (inversely) influence the output.



The randomly selected example document that certain combinations of words such as *compani identifi, mid market, best practic, plan improv, proprietary network,* and *brand name* directly affect *GradientBoosting's* output. Conversely, zero or relatively small values of *invest criteria* inversely influence the output.

We emphasize that machine learning algorithms use non-linear interactions among multiple word combinations to make predictions. As a result, we cannot state that a fund will perform well or badly because its investment approach's description includes a specific combination of words. Here is where the beauty of machine learning algorithms lies. The ability to make sense of complex, non-linear relationships among various features makes machine learning algorithms suitable for identifying patterns that humans cannot observe.

8. Summary and outlook

We combine Natural Language Processing and machine learning algorithms to predict fund performance. We show that algorithms could help investors identify fund managers and contribute to a more efficient market. Our analysis contributes to our understanding of the fund manager selection in three ways. First, unlike previous literature that focuses on structured data such as past performance included in fundraising documents, we use Natural

Language Processing techniques to characterize investment approaches utilizing the textual data of PPM's Investment Strategy and Process section. Second, we evaluate the possibility of constructing machine learning algorithms trained uniquely on textual data available during fundraising to classify funds as successful or failures. Third, we identify combinations of words that tend to be associated with success and failure funds. Fourth, we show that probabilities of success predicted by machine learning algorithms have a statistically and economically significant relationship with observed fund performance.

We conduct numerous analyses to rule out the possibility that the predictive power of the algorithms is driven by the funds included in the training set in terms of the vintage year and investment type.

We assess the economic implications of using our machine learning approach the select and deselect funds. First, we use probabilities of success predicted by the algorithms to build portfolios of funds. Then, we compare the performance of those portfolios with that of the funds not selected by the algorithms. We find that algorithms-based portfolios outperform deliver attractive returns in comparison with portfolios of randomly selected funds.

Machine learning algorithms combine multiple non-linear interactions to make predictions of GP's quality. While we cannot identify all those interactions, we identify the individual combinations of words that most contribute to generating those predictions.

This paper shows how investors could potentially benefit from combining NLP and machine learning while conducting due diligence of PE funds. While investors conduct rigorous analyses and look at a wide range of characteristics when evaluating the suitability of committing to a PE fund, the final investment decision is partially influenced by the gut feeling built on previous experience. Thus, a key advantage of our machine learning-based approach to selecting funds relies on the objectivity applied by the algorithms in the selecting process.

The main implication of our results is that machine learning techniques contribute to a more accurate expectation of GP quality and, in turn, towards a more efficient market.

While this paper opens the path to understanding how machine learning techniques contribute to a more efficient PE market, future research should consider examining additional potential applications of these techniques leveraging the increased availability of commercial databases and the surge of more advanced algorithms.

9. Bibliography

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

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

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.

Chung, J. W., Sensoy, B. A., Stern, L., & Weisbach, M. S. (2012). Pay for performance from future fund flows: the case of private equity. The Review of Financial Studies, 25(11), 3259-3304.

Harris, R. S., Jenkinson, T., & Kaplan, S. N. (2014). Private equity performance: What do we know?. The Journal of Finance, 69(5), 1851-1882.

Harris, R. S., Jenkinson, T., Kaplan, S. N., & Stucke, R. (2018). Financial intermediation in private equity: How well do funds of funds perform?. Journal of Financial Economics, 129(2), 287-305.

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

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

Kaplan, S. N., & Sensoy, B. A. (2015). Private equity performance: A survey. Annual Review of Financial Economics, 7, 597-614.

Ke, Z. T., Kelly, B. T., & Xiu, D. (2019). Predicting returns with text data (No. w26186). National Bureau of Economic Research.

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

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

Lundberg, S. M., & Lee, S. I. (2017, December). A unified approach to interpreting model predictions. In Proceedings of the 31st international conference on neural information processing systems (pp. 4768-4777).

Metrick, A., & Yasuda, A. (2011). Venture capital and other private equity: a survey. European Financial Management, 17(4), 619-654.

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.

Purda, L., & Skillicorn, D. (2015). Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. Contemporary Accounting Research, 32(3), 1193-1223.

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.

Appendix for «*Selecting Private Equity Funds Using Machine Learning*»

This Appendix includes the following additional sections:

1. Section A1 presents the predictive power of the algorithms when using the Investment Highlights section
2. Section A2 shows the predictive power of the algorithms when using the Market Opportunity section
3. Section A3 discusses the implementation details of Term Frequency-Inverse Document Frequency
4. Section A4 describes the machine learning algorithms
5. Table A1 shows the most common combinations of two adjacent words across investment types
6. Figure A1 presents the mean USD-denominated TVPI over the mean EUR-denominated TVPI from 1996 to 2014
7. Figure A2 depicts the implementation process of fivefold cross-validation
8. Figure A3 shows the analogous to Figure 2 but using the median *Predicted Probability of Success* rather than 0.5 as the threshold to classify funds as successful or failure.

Section A1. Investment Highlights sections results

Figure A11. Statistical assessment of machine learning algorithms to predict performance using the Investment Highlights section

This table reports the out-of-sample statistical performance of several algorithms using funds in the 2012-2014 test set. Algorithms are trained on the 1996-2011 training set of funds describing their Investment Highlights. Area Under the Receiver Operating Characteristic curve (AUC ROC) is the probability that a random positive instance is ranked higher than a random negative instance. Accuracy is the proportion of correctly predicted instances. Finally, Balanced Accuracy is the average of the proportions corrects of each class individually.

	Lasso	Random Forest	Gradient Boosting	Naïve Bayesian
Accuracy	0.568	0.543	0.543	0.568
Balanced Accuracy	0.566	0.541	0.541	0.566
ROC AUC	0.573	0.565	0.522	0.592

Figure A12. Mean Benchmark Distance for predicted classes using the Investment Highlights section

The figure shows the mean *Benchmark Distance* of predicted classes for ML models in the 2012-2014 test set. Algorithms are trained on the 1996-2011 training set describing their Investment Highlights. *Benchmark Distance* is defined as the difference between observed Total Value to Paid-In (TVPI) and the median TVPI of Preqin funds sharing the vintage year, investment type, and geographic focus. TVPI is the ratio of all capital distributions plus the last reported Net Asset Value to the total amount of capital invested (including fees). Predicted Success (Predicted Failure) comprises the funds with a *Predicted Probability of Success* equal to or above (below) 0.5. *Benchmark Distance* is winsorized at 1% and 99%. [a], [b], and [c] indicate significance at the 1%, 5%, and 10% levels, respectively, in a two-sample t-test comparing mean *Benchmark Distance* across predicted classes.

Section A2. Market Opportunity sections results

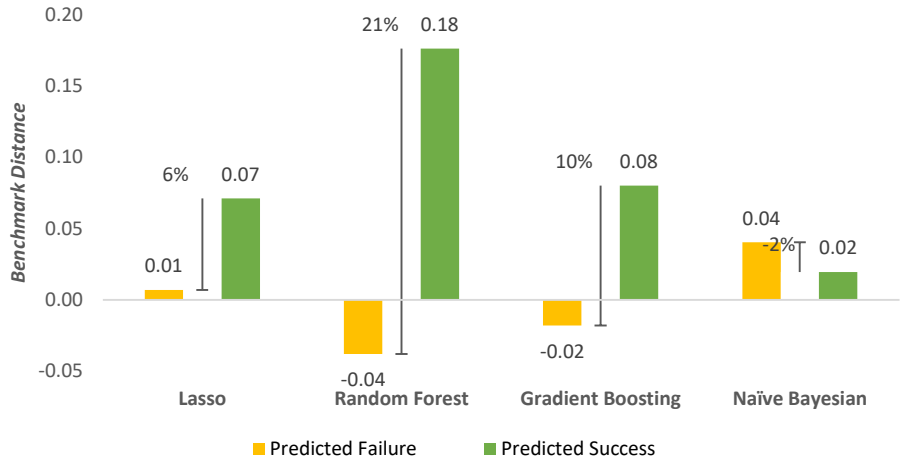
Figure A21. Statistical assessment of machine learning algorithms to predict performance using the Investment Highlights section

This table reports the out-of-sample statistical performance of several algorithms using funds in the 2012-2014 test set. Algorithms are trained on the 1996-2011 training set of funds describing their Market Opportunity. Area Under the Receiver Operating Characteristic curve (AUC ROC) is the probability that a random positive instance is ranked higher than a random negative instance. Accuracy is the proportion of correctly predicted instances. Finally, Balanced Accuracy is the average of the proportions corrects of each class individually.

	Lasso	Random Forest	Gradient Boosting	Naïve Bayesian
Accuracy	0.508	0.556	0.619	0.524
Balanced Accuracy	0.507	0.553	0.619	0.521
ROC AUC	0.555	0.585	0.619	0.560

Figure A22. Mean Benchmark Distance for predicted classes using the Market Opportunity section

The figure shows the mean *Benchmark Distance* of predicted classes for ML models in the 2012-2014 test set. Algorithms are trained on the 1996-2011 training set describing their Market Opportunity. *Benchmark Distance* is defined as the difference between observed Total Value to Paid-In (TVPI) and the median TVPI of Preqin funds sharing the vintage year, investment type, and geographic focus. TVPI is the ratio of all capital distributions plus the last reported Net Asset Value to the total amount of capital invested (including fees). Predicted Success (Predicted Failure) comprises the funds with a *Predicted Probability of Success* equal to or above (below) 0.5. *Benchmark Distance* is winsorized at 1% and 99%. [a], [b], and [c] indicate significance at the 1%, 5%, and 10% levels, respectively, in a two-sample t-test comparing mean *Benchmark Distance* across predicted classes.



Section A3. Implementation of the Term Frequency-Inverse Document Frequency vectorizer

We first tokenize each document and convert all characters to lower case. Then, we remove the punctuation, numbers, and most common words (e.g., "the" so-called stop words) [20]. Next, we apply the Porter stemming algorithm. This algorithm removes the commoner morphological and inflexional endings from words (e.g., we transform "companies" into "compani").

The fund invests in companies with upside potential, a strong management team, and sustainable cash flows.

↓

fund invest compani upsid potenti strong manag team sustain cash flow.

[20] We filter stop words out using the Natural Language Toolkit (NLTK) in python.

We use the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer developed by Pedregosa et al. (2011) to transform the pre-processed text into numerical variables. The informativeness of the resulting numerical variables to predict performance depends on some parameters subject to be calibrated: type of terms to include in the model (e.g., words, bigrams, or trigrams); the minimum number of documents that contain a term to include it in the model; and the maximum number of features to include in the model.

Words, bigrams, and trigrams

The features included in our model can represent a single word or a combination of two or three adjacent words. While combinations of words provide more context than words, the document frequency of the former is below that of the latter so is the ability to generalize patterns. A priori, the optimal type of feature to include in the model is unknown. Thus, we try models with only words or words and combinations of two or three adjacent words. We select the type of combination that yields the highest predictive power.

Minimum document frequency

Terms with low document frequency (e.g., appear in few documents) do not help us generalize patterns and, consequently, discriminate investment approaches. For example, our corpus contains the investment strategy and process of three consecutive funds belonging to the same firm. However, the firm's name has no informative power to discriminate investment approaches of funds raised by other firms. Because of this, we try different thresholds and select the one documenting the best forecasting power.

Maximum number of features

Terms with a low term frequency in the entire corpus might not help us to generalize patterns. We thus different thresholds for the maximum number of features to include in the model.

We explore several combinations of parameters and select the one that provides the highest discriminative power.

Section A4. Machine learning algorithms

This table describes the algorithms used in our analyses. All algorithms are implemented using the Sklearn package.

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:

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

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

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

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

The overall cost is computed as follows:

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

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.

Naïve Bayesian

Probabilistic model that uses the probabilities of observed outcomes to return an estimate of an unknown outcome. The algorithm treats each feature as independent from the others.

Given an example i with a vector of features $x^{(i)}$ being j the total number of features, and an output $y^{(i)}$, the naïve Bayesian algorithm solves the following problem:

$$\hat{y}^{(i)} = \max_{y \in Y} \Pr\left(y^{(i)} \middle| x^{(i)}\right) \quad (1)$$

Where $\hat{y}$ is the category with the highest probability of having generated $x^{(i)}$. To calculate $\Pr\left(y^{(i)} \middle| x^{(i)}\right)$, known as posterior probability, we useBayes' rule:

$$\Pr\left(y^{(i)} \middle| x^{(i)}\right) = \frac{\Pr(y^{(i)}) \prod_{j=1}^{n} \Pr\left(x_j^{(i)} \middle| y^{(i)}\right)}{\prod_{j=1}^{n} \Pr(x_j^{(i)})} \quad (2)$$

Where $\Pr\left(x^{(i)}\middle|y^{(i)}\right)$ is the prior probability across all features n, $\Pr(x^{(i)})$ is the marginal likelihood across all features n, and $\Pr(y^{(i)})$ is the likelihood probability. Because the denominator stays constant for a given input, we can ignore the denominator. Therefore, we rewrite equation (1) as equation (3):

$$\hat{y} = \max_{c \in C} \Pr\left(y^{(i)}\right) \Pr\left(x^{(i)}\middle|y^{(i)}\right) \quad (3)$$

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.

Table A1: Describing investment approaches across investment types

The table shows the thirty most common bigrams across investment types. The first three columns show the thirty most common bigrams, their frequency, and the percentage of documents containing them, respectively, for the Leverage Buy-outs sample. The following three columns show the same descriptives for the thirty most common bigrams for the Venture Capital sample. Finally, the last three columns show the same descriptives for the thirty most common trigrams for Other Private Equity types.

	Leveraged Buy-Outs			Venture Capital			Other Private Equity		
	Bigram	**Bigram Frequency**	**Document Frequency**	**Bigram**	**Bigram Frequency**	**Document Frequency**	**Bigram**	**Bigram Frequency**	**Document Frequency**
1	portfolio companies	1730	90%	portfolio companies	208	98%	portfolio companies	361	91%
2	due diligence	1420	86%	management team	124	63%	due diligence	318	89%
3	portfolio company	1142	74%	due diligence	120	75%	investment opportunities	211	82%
4	value creation	1047	73%	early stage	99	75%	investment team	190	47%
5	management team	1042	85%	deal flow	85	80%	portfolio company	190	67%
6	private equity	1040	85%	portfolio company	82	70%	private equity	187	74%
7	investment opportunities	884	83%	investment opportunities	79	65%	management teams	157	67%
8	management teams	851	80%	venture capital	77	63%	management team	155	72%
9	investment committee	662	45%	investment manager	76	5%	deal flow	140	58%
10	investment team	627	39%	general partner	63	23%	investment committee	139	51%
11	investment strategy	539	79%	investment strategy	56	68%	value creation	139	58%
12	deal flow	534	67%	life science	50	28%	investment strategy	133	86%
13	investment process	413	63%	investment team	46	33%	investment manager	125	4%
14	long term	374	60%	management teams	46	60%	investment process	99	18%
15	deal team	366	31%	investment committee	40	3%	general partner	95	2%
16	company management	347	3%	stage companies	38	30%	deal team	81	67%
17	mid market	346	56%	investment process	37	35%	board directors	76	25%
18	add acquisitions	339	39%	advisory board	37	48%	long term	73	25%
19	business plan	337	41%	life sciences	35	35%	diligence process	71	40%
20	buy build	294	51%	business model	34	28%	business plan	68	65%
21	investment professionals	291	33%	later stage	33	40%	fund investment	60	56%
22	track record	288	38%	investment professionals	32	30%	mid market	59	44%
23	middle market	285	52%	private equity	30	20%	track record	57	4%
24	diligence process	279	29%	start ups	29	33%	business model	53	54%
25	cash flow	273	46%	limited partners	28	18%	target company	51	21%
26	target companies	269	50%	intellectual property	28	38%	company management	48	51%
27	investment criteria	261	46%	business models	28	38%	investment opportunity	48	2%
28	target company	261	50%	track record	27	43%	middle market	47	35%
29	competitive advantage	252	44%	diligence process	26	43%	growth capital	46	4%
30	general partner	249	50%	technology companies	25	3%	corporate governance	45	40%

Figure A1. Comparative of Total Value to Paid-In (TVPI) in USD and EUR

The figure compares the mean Total Value to Paid-In (TVPI) measured in USD and EUR across vintage years. The TVPI is computed for each fund with the complete time-series of cash flows, and Net Asset Values in Preqin Fund Cash Flows dataset.

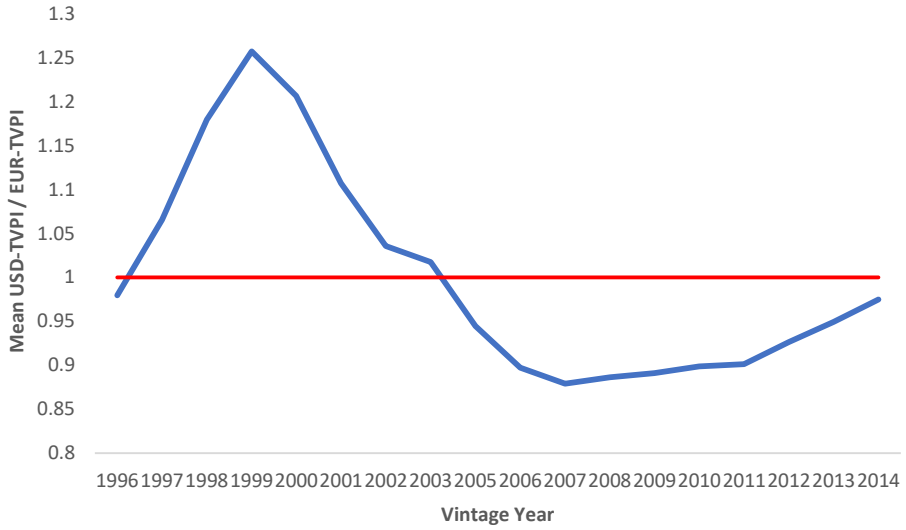


Figure A2. Fivefold cross-validation implementation

The figure depicts the implementation process of fivefold cross-validation, a resampling procedure to evaluate machine learning models. This process delivers the unknown tuning parameters used to train the machine learning models on the training sample. Next, the trained model is used to classify the funds in the test sample. AUC stands from Area Under the Receiver Operating curve and is the metric used to select the best tuning parameters throughout the cross-validation process.

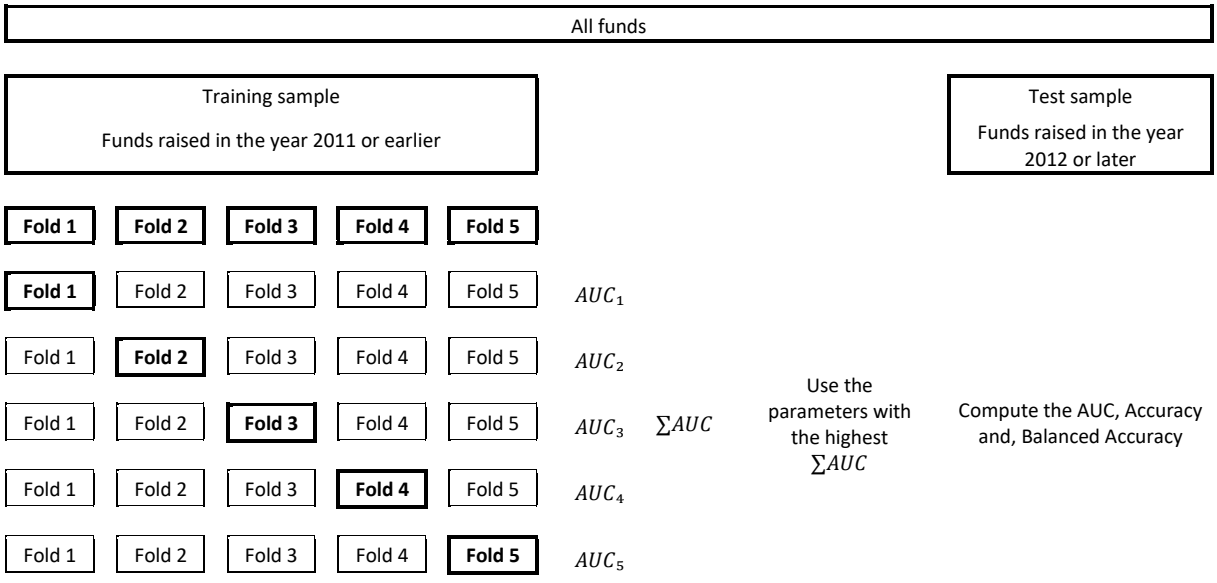


Figure A3. Mean Benchmark Distance for predicted classes

The figure shows the mean *Benchmark Distance* of predicted classes for ML models in the 2012-2014 test set. Algorithms are trained on funds raised between 1996 and 2011. *Benchmark Distance* is defined as the difference between observed Total Value to Paid-In (TVPI) and the median TVPI of Preqin funds sharing the vintage year, investment type, and geographic focus. TVPI is the ratio of all capital distributions plus the last reported Net Asset Value to the total amount of capital invested (including fees). Predicted Success (Predicted Failure) comprises funds with a *Predicted Probability of Success* equal to or above (below) the median *Predicted Probability of Success*. *Benchmark Distance* is winsorized at 1% and 99%. [a], [b], and [c] indicate significance at the 1%, 5%, and 10% levels, respectively, in a two-sample t-test comparing mean *Benchmark Distance* across predicted classes.