Predictability and Persistence in Deal Selection: Evidence from Venture Capital*

Luiz Bissoto†

May 29, 2026

Abstract

I study whether predictive technologies such as machine learning (ML) methods can improve deal selection in venture capital (VC) and whether investors can systematically translate predictive advantages into economic gains. Using a large panel of U.S. financing rounds, I show that portfolios constructed using ML signals derived exclusively from past and publicly available information to select out-of-sample deals outperform most investors in the U.S. VC market. I find that these gains are larger when models are trained to predict rare outcomes, such as IPOs and acquisitions, and when applied to early-stage deals. Overall, these ML-based portfolios tend to outperform a large share of investors—often those outside the top quartile in terms of success rates. Despite this potential incremental performance ("benefit"), I find that its persistence is weak within investors: once investor and time fixed effects are accounted for, benefit is strongly mean reverting, both in existence and magnitude. Exploiting a plausibly exogenous shock to the implementation cost of these technologies, I show that investors with higher benefit ex ante tend to have it quickly eroded post-shock, suggesting that the existence of exploitable predictive gains is structural rather than behavioral.

Keywords: Venture Capital, Institutional Investors, Investment Decisions, Machine Learning, Performance Prediction

JEL Codes: G11, G17, G23, G24

*I am grateful to my advisor, Ruediger Fahlenbrach, as well as Andreas Fuster and Francesco Celentano, for their continuous support and guidance. I also thank Alexandre Azoulay, Purba Bhattacherjee, Luise Eisfeld, Benjamin Gabay, Jean Herelle, Federico Baldi Lanfranchi, Semyon Malamud, Giuseppe Matera, Darius Nik Nejad, Florian Perusset, Mohammad Pourmohammadi, and the participants of the 9th Annual Cambridge Conference on Alternative Finance for their feedback and helpful comments.

†EPFL and Swiss Finance Institute, Extranef (Unil-Dorigny), Lausanne 1015. Email: luiz.bissoto@epfl.ch

1 Introduction

Venture capital (VC) investments are typically understood as high-risk and high-return investments. With failure rates frequently exceeding 90 percent[1], investor returns tend to be tightly linked to the realization of successful investment exits (Cumming and MacIntosh (2003), Cochrane (2005), Phalippou and Gottschalg (2009)). Over the past two decades, technological advances, particularly in machine learning (ML), have facilitated the development and diffusion of tools that have expanded the scope of predictive analysis in many forecasting tasks. These tools have increasingly been applied to VC investment screening and classification (Bonelli (2025)), and a growing literature documents predictability of startup success using historical data (Krishna et al. (2016); Retterath (2020); Ross et al. (2021); Kiss (2022); Razaghzadeh Bidgoli et al. (2024); Kalbande and Karmore (2024)).

In this paper, I study whether these ML signals can reliably generate out-of-sample predictions that translate into realized investment gains, and what explains the existence of such predictability in a highly competitive market with relatively free entry (Cochrane (2005)). I address these questions through a comprehensive predictability study built around four complementary hypotheses. First, I examine whether these predictive signals are stable over time. The underlying idea is that, if such signals cannot be expected to improve investment decisions ex ante, investors may rationally refrain from relying on them. Second, I investigate whether predictive signals imply distinct economic gains. For instance, Bonelli (2025) finds that VC investors who rely on historical data tend to invest in startups that subsequently receive financing, but fail to identify major breakthrough startups. Third, I study whether potential gains from predictive signals are persistent within investors. That is, whether investors can know ex ante that these signals are likely to benefit them specifically, rather than investors on average. Finally, I examine the extent to which investors are constrained in applying these signals in practice. Such constraints may arise from limited access to predictive technology (DeMiguel et al. (2023)), insufficient or unavailable data to

[1] In my sample, fewer than 10% of startups survive and eventually become public companies.

learn meaningful patterns (Bonelli and Foucault (2023); Agarwal et al. (2024)), or systematic biases in decision-making (Lyonnet and Stern (2022); Davenport (2022)).

I find that ML-based predictive models are able to learn economically meaningful patterns from past and publicly available data. In particular, the gap between success rates from model-based predictions and those realized by most investors is larger for high-return outcomes, such as initial public offerings (IPOs) and acquisitions. These results are especially strong for predictions based on information available at early investment stages, such as seed and Series A rounds, when investor bargaining power relative to entrepreneurs is higher (Hsu (2004)) and access to deal flow (Nanda et al. (2020)) is a lesser concern. At the investor level, I find that the potential gains from relying on ML predictions are widespread, but strongly mean-reverting over time. Importantly, I find little evidence that differences in investment constraints, such as geographic location, sectoral focus, or investor network structure, explain the gap between ML-implied and realized performance. By contrast, I find evidence that access to predictive technology plays a central role. Following the release of major ML libraries and frameworks, the gap between counterfactual and realized investor success rates systematically shrinks, with the largest effects observed for investors who would have benefited the most ex ante, consistent with a technology diffusion mechanism.

I work with a dataset covering financing rounds of U.S.-based startups from 2001 to 2021. For each financing round, I construct a set of predictive variables using only information available at the time of the transaction. These variables include investor experience, startup sector, and capital raised in prior rounds. I define success outcomes based on post-investment events, distinguishing between subsequent financing, acquisition, and initial public offering (IPO). Using these data, I train ML models—including logit and extreme gradient boosting (XGBoost; see, e.g., Breiman (2001); Friedman (2001))[2]—to generate predicted success

[2]I focus on tree-based methods, which are well suited for structured, tabular data such as VC investment records and have been shown to perform competitively in similar empirical settings (see, e.g., Couronné et al. (2018); Hollmann et al. (2023)). Neural-network–based approaches typically require substantially richer feature representations and larger sample sizes to perform well in this context, and I therefore treat them as less natural benchmarks for the present application.

probabilities. Models are estimated separately by success outcome and investment stage and are trained using rolling windows to generate strictly out-of-sample predictions. This design prevents results from being driven by differences across success definitions, investment stages, or specific time periods. I then compare success rates implied by ML-based portfolios with those realized by investors' actual investment choices across outcomes and stages. From this comparison, I construct a benefit measure defined as the gap between the counterfactual ML-implied success rate and the realized success rate, which I use to study the determinants and persistence of gains from relying on predictive signals.

I begin by evaluating whether ML predictive signals extract meaningful information from past and publicly available data. Across ML models, portfolios formed on predicted success probabilities outperform the realized success rates of most VC investors. This performance gap is largest for rare, high-value outcomes—IPOs, followed by acquisitions—and is especially pronounced when predictions are constructed using information available at early investment stages such as seed and Series A rounds. These settings are characterized by high uncertainty, strong outcome skewness, and limited scope for simple linear screening rules. Importantly, these findings indicate that rare outcomes are sufficiently frequent to avoid a "curse-of-rarity problem" and exhibit learnable structure, consistent with recent evidence in related settings (Agarwal et al. (2024); Peukert et al. (2024)). As a result, ML-based signals are particularly informative for predicting outcomes in the right tail of the returns distribution.

I then move to an across- and within-investor analysis. For each investor, I compute a counterfactual measure of performance by selecting the top n investments with the highest predicted success probabilities, where n equals the number of investments the investor actually made. I define the performance gap between this counterfactual portfolio and the investor's realized outcomes as *benefit*, and I measure it at the investor-year level. Three main results emerge from this analysis. First, the share of investors with a positive benefit is large, often exceeding 50% across outcomes and around 80–90% for IPOs. Second, the average benefit is sizable, particularly for IPOs, where it ranges from about 14% under logit

to nearly 40% under tree-based models. Third, benefit exhibits persistence across investors, in the sense that investors with a positive benefit in one period are more likely to exhibit a positive benefit in the next. However, this persistence is almost entirely explained by investor and time fixed effects. Once added, benefit is strongly mean reverting in both existence and magnitude. Moreover, differences in observable deal constraints such as geography, sector, or network proximity explain little of the gap between counterfactual and realized performance. Taken together, these findings indicate that predictive gains are widespread but not investor-specific, raising the question of whether benefit reflects systematic differences between startups predicted to succeed and those in which investors typically invest.

Next, I examine whether benefit can be explained by differences in startup characteristics between firms predicted to succeed and those in which investors typically invest. I find that neither geographic location, sector, nor the degree of sector "hotness", measured by the number of rounds occurring in the sector, explains the magnitude of benefit. In other words, investors do not miss successful startups because these firms are located far from their usual investment areas, operate in unfamiliar sectors, or compete in overly crowded segments.

In the final part of the analysis, I exploit a plausibly exogenous technological shock to study how access to predictive technology affects benefit. The shock is the staggered release of several major ML libraries and methods between 2015 and 2017, including TensorFlow, PyTorch, Microsoft Cognitive Toolkit, and CatBoost. The underlying hypothesis is that improved access to predictive technology should reduce benefit as investors become increasingly able to implement ML-based prediction tools. A key identifying assumption is therefore that an investor's exposure to the shock is proportional to average benefit prior to its occurrence. Rather than relying on observed adoption or usage of predictive technologies, which may be mismeasured or correlated with unobserved investor characteristics, I use pre-shock benefit as a revealed measure of exposure. Investors with higher benefit before the shock are those for whom access to ML tools should matter most. Consistent with this interpretation, I find that higher pre-shock exposure is associated with a substantial decline in post-shock benefit.

The estimated reduction ranges from 0.9 to 1.5% for early-stage investments and from 9.4 to 12.1% for mid-stage investments, defined as Series B and C rounds. The evidence indicates that advantages erode as access to ML tools becomes cheaper and more widespread.

My contribution is twofold. First, I provide an economic assessment of ML predictability in the context of VC by benchmarking ML-based counterfactual portfolios against the realized performance of actual investors and measuring the gap between the two. This assessment is conducted within a rigorous empirical setting that relies exclusively on backward-looking, publicly available data and ensures comparability across investments. Rather than simply showing that ML methods can improve performance, a result already established by a growing literature, I construct a measure targeted at quantifying these gains and study how they are distributed across and within investors. Second, I explain how such gains can arise in a competitive industry. I offer a structural, rather than behavioral, explanation. While benefit is widespread, it is also strongly mean reverting once investor and time fixed effects are accounted for, making its existence and magnitude difficult to assess ex ante at the investor level. Benefit is instead closely tied to access to predictive technology: investors with the largest ex ante gains experience the strongest subsequent erosion. This pattern is consistent with a limits-to-arbitrage framework (Shleifer and Vishny (1997)), in which predictive gains are difficult to arbitrage away because of constraints on identifying, accessing, and confidently exploiting these gains in practice.

I contribute primarily to two strands of literature. First, I add to the literature applying ML methods to evaluate whether and how they improve performance in settings where asset prices are not readily available and data are scarce, such as VC and private equity more broadly (Krishna et al. (2016); Retterath (2020); Ross et al. (2021); Kiss (2022); Razaghzadeh Bidgoli et al. (2024); Kalbande and Karmore (2024)). I revisit the central result of this literature, that ML-based gains are possible, using an empirical design that is simple to replicate and robust to forward-looking, selection, and data-availability biases. In doing so, I reinforce the broader idea that complex methods (Kelly et al. (2024); Chen et al. (2025)) can

deliver substantial improvements in predicting returns when applied in environments that closely mirror actual investment decisions. Second, I contribute to the literature studying the investor-level consequences of adopting ML-based decision tools and the constraints that limit their effectiveness. In particular, I add to Lyonnet and Stern (2022); Davenport (2022); Bonelli (2025) by providing evidence that investors face structural, rather than purely behavioral, reasons for failing to benefit more from these methods. Even if investors were able to mitigate biases in how predictive technologies are applied and expand data collection, they would still be constrained by access to these technologies and by uncertainty over whether the expected benefits are sufficiently large to justify the costs of adoption.

2 Hypotheses Development

I develop a set of hypotheses to test whether predictive gains exist in the data and whether, how, and for whom these gains can be realized in practice. The hypotheses focus on the stability, economic relevance, investor specificity, and feasibility of ML-based predictive gains, and together provide a framework for understanding why ML can appear powerful in aggregate while delivering uneven and fragile benefits at the investor level.

H1. Unstable Predictability. ML-based predictability of startup success exists on average but is unstable over time. The magnitude and usefulness of predictive signals vary sharply across years, even when models are trained and evaluated using strictly out-of-sample procedures. As a result, investors face uncertainty not only about which startups are likely to succeed, but also about whether ML signals are informative in a given year. This instability limits the practical value of algorithmic predictions, as investors cannot reliably anticipate when predictive gains will be sufficiently large or persistent. I evaluate this hypothesis by examining time-series variation in the predictive performance and investor-level benefit delivered by ML-based portfolios.

H2. Non-Value Predictability. Predictability of startup success does not necessarily imply economic value for investors. ML models may correctly identify startups that are more likely to succeed, but these successes may be associated with lower expected returns due to competitive pressures, valuation effects, syndication structures, or investment stage. As a consequence, improvements in success prediction need not translate into economically meaningful gains from an investor's perspective. This hypothesis motivates the distinction between aggregate predictive performance and investor-level benefits and cautions against interpreting success predictability as evidence of superior investment outcomes.

H3. Non-Investor-Specific Predictability. ML-based predictability is not investor-specific. Even when ML improves outcomes on average, individual investors cannot reliably determine ex ante whether they will benefit from using these predictions in a given year. Predictive gains may accrue to different investors at different times, without exhibiting stable patterns at the investor level. This lack of investor-specific predictability weakens incentives to adopt ML tools, as investors cannot expect consistent improvements in their own performance. I examine this hypothesis by studying the distribution and persistence of investor-level benefits and by analyzing whether positive gains tend to repeat for the same investors over time.

H4. Constrained Predictability. Investor-level gains from ML signals are limited because predicted successful investments lie outside investors' feasible investment domains. Differences in specialization, geographic focus, deal access, and portfolio constraints prevent investors from acting on ML signals, even when those signals are informative. As a result, predictive gains exist in principle but are difficult to realize in practice because predicted successful investments are distant from investors' typical activity. I test this hypothesis by constructing a distance measure between an investor's realized portfolio and the set of startups predicted to succeed by the model and by relating this distance to the magnitude and persistence of investor-level benefits.

Together, these hypotheses provide a framework for understanding why ML-based predictability can coexist with limited and uneven investor-level gains. The empirical analysis evaluates each hypothesis within a common methodological setting, allowing competing mechanisms to be distinguished while recognizing that several may operate simultaneously.

3 Data

3.1 Sample

My main data source is Crunchbase, from which I obtain a comprehensive dataset of U.S. VC financing rounds with detailed deal-level information on startups, investors, and financing events. Two features of this data provider are particularly relevant for studying VC activity at scale. First, it offers distinctively comprehensive coverage of U.S. financing rounds in the post-2010 period, which aligns closely with the geography and time span of interest (see, e.g., Bonelli (2025)). Second, it is substantially more used by VC investors than close commercial alternatives such as PitchBook and Dealroom, supporting its reliability as a source of information on VC deals (Retterath (2024)).

I begin from the universe of all financing rounds recorded in the database between 2000 and 2021, and focus inference on the post-2010 period, where coverage is most complete. I restrict the sample to startups headquartered in the United States and to rounds denominated in U.S. dollars. Financing rounds are required to have a valid announcement date, a strictly positive amount raised, and at least one identifiable investor. I further restrict attention to standard venture financing rounds, focusing on seed and Series A through Series G investments. These restrictions exclude non-venture transactions and idiosyncratic deal types that are not comparable across investors or over time.

After applying these filters, the final sample consists of 55,858 financing rounds involving 30,373 unique startups. The total capital raised in the sample amounts to $823.5 billion. Financing rounds are typically syndicated, with an average of 3.58 investors per round. Panel

A in Appendix A reports summary statistics for the resulting sample. The unit of observation is a financing round. For each round, the dataset includes information on the investment stage, the announcement date, the set of participating investors, the amount of capital raised, pre-money valuation information when available, and detailed startup characteristics, including sector classifications and geographic location. Panel B provides a representative example of a financing-round observation.

3.2 Definition of Variables

I present here the outcome variables and the set of predictive features used in the empirical analysis. All variables are constructed at the financing-round level.

Success outcomes. I consider three alternative measures of startup success: follow-on financing, acquisition, and initial public offering (IPO). Each outcome is defined as a binary indicator at the level of a financing round.

A follow-on financing indicator equals one if the startup raises at least one additional financing round after the focal investment and zero otherwise. An acquisition indicator equals one if the startup is eventually acquired by another firm. An IPO indicator equals one if the startup eventually goes public. To allow sufficient time for outcomes to materialize, I use data observed through 2025 to classify success. In all cases, the indicator variable equals one if the success outcome occurs at any point in the future.

Each financing round is therefore associated with up to three distinct success labels, corresponding to the different definitions of success. These outcomes are analyzed separately, reflecting their different economic meanings and frequencies.

Predictive features. The predictive feature set is constructed to reflect the information available to investors at the time of the financing decision. All features are measured at the financing-round level and are based exclusively on information observed prior to the announcement date of the round.

The first group of features captures startup- and deal-level characteristics. These include the total amount of capital raised in the financing round, the size of the investment syndicate, indicators for the presence of at least one lead investor and the number of lead investors, and string-based characteristics of the startup name, namely name length, number of words, and number of unique characters. I also include binary indicators for startup sector classifications based on company category (or sector) tags, using the most frequent categories in the data.

The second group of features measures investor experience and historical activity. For each financing round, I construct cumulative measures of the participating investors' past investment activity, including the total number of prior investments and the total amount of capital previously invested. In addition, I compute rolling-window measures of investor activity over multiple horizons—12, 24, 36, 48, 60, and 120 months prior to the round—covering both investment counts and invested capital. For each horizon, I compute summary statistics across the investor syndicate, including sums, means, and maxima.

The third group of features captures investor–startup matching. These variables measure the extent to which the participating investors have prior experience in the sectors in which the startup operates. For each investor, I construct a historical sector exposure profile based on the investor's past investments, where exposure is measured by the number of prior investments in each sector. Each startup is characterized by a set of sector indicators derived from its category tags. I quantify investor–startup alignment using weighted Jaccard similarity measures[3] between the investor's historical sector exposure vector and the startup's sector indicator vector. Intuitively, this measure captures how much an investor's prior activity

[3]Let s index sectors. For investor i, let $w_{i,s} \geq 0$ denote the number of past investments made by investor i in sector s prior to the financing round. For startup j, let $x_{j,s} \in \{0, 1\}$ indicate whether startup j operates in sector s. The weighted Jaccard similarity between investor i and startup j is defined as

$$J_{i,j} = \frac{\sum_s \min\{w_{i,s},\, x_{j,s}\}}{\sum_s \max\{w_{i,s},\, x_{j,s}\}}.$$

This measure equals zero if there is no overlap between the investor's historical sector exposure and the startup's sectors, and equals one if the investor's historical activity is fully concentrated in the startup's sectors. At the financing-round level, I compute the mean and the maximum of $J_{i,j}$ across all investors participating in the round.

overlaps with the startup's sectors, weighting overlap by the intensity of past investment. The resulting similarity index ranges from zero, indicating no overlap between investor experience and startup sectors, to one, indicating that the investor's historical activity is fully concentrated in the startup's sectors. I compute both the mean and the maximum overlap across investors in the syndicate. In addition, I include a measure of the startup's prior investor network depth, defined as the number of distinct investors that have previously invested in the startup before the focal financing round.

The fourth group of features reflects market conditions and competitive environment. These variables capture the intensity of financing activity in the VC market prior to the round. I compute counts of financing rounds of the same investment stage over rolling windows of 12, 24, 36, 48, 60, and 120 months. Analogous measures are constructed at the state level to capture local market conditions. I also include measures of financing activity in adjacent investment stages—both preceding and subsequent rounds—to capture congestion and momentum across the financing cycle.

The final group of features consists of time controls. These include calendar-month indicator variables to capture seasonality and a linear time index to capture long-run trends in VC activity. Appendix B.1 provides a complete list of all features used in the analysis, grouped by category and reported with their exact construction, while Appendix B.2 provides their summary statistics.[4]

4 Motivating Evidence: ML Models Applied to VC

In this section, I provide motivating evidence on the ability of ML models to predict startup success in VC settings. The goal is not to introduce new algorithms or to optimize predictive performance, but to illustrate what modern prediction methods can and cannot do when

[4]I do not conduct a feature-importance or model explainability analysis, as my objective is not to identify which specific variables drive predictive performance but to assess whether predictive gains exist, whether they are economically meaningful, and whether investors can systematically realize them in equilibrium. Upon request, I provide trained model specifications suitable for inference on similarly constructed datasets, the corresponding predicted success probabilities, and the generated feature-importance diagnostics.

applied strictly out-of-sample. This discussion motivates the empirical strategy used in the remainder of the paper.

4.1 Models

I consider three classes of models that differ in flexibility and functional form. The first is a standard logistic regression (logit), which serves as a familiar, transparent, and widely available methodological benchmark. Logit imposes a linear index structure on the predictors and is closely related to the classification models commonly used in empirical finance, providing a natural reference point against which more flexible methods can be evaluated.

The second class of models consists of tree-based ensemble methods, specifically gradient-boosted decision trees implemented using XGBoost (Friedman (2001), Chen and Guestrin (2016)) and CatBoost (Prokhorenkova et al. (2018)). These models have been widely used in applied ML for tabular data for over a decade and are known to perform particularly well in settings with heterogeneous predictors, nonlinear effects, and interactions (see, e.g., Hollmann et al. (2023)). Unlike linear models, tree-based methods do not require the researcher to specify nonlinearities or interaction terms ex ante; instead, these features are learned directly from the data. I provide a summary of these models and their properties in Panel A of Appendix C.

Despite their greater flexibility, these models are explicitly designed to control overfitting. First, they rely on *shrinkage*, which scales down the contribution of each additional tree so that predictions are adjusted gradually rather than aggressively, preventing the model from fitting noise in the training data. Second, they employ *early stopping*, whereby model training is halted once predictive performance on a held-out validation subset of the training data ceases to improve, avoiding unnecessary complexity. Third, model complexity is governed through *cross-validated hyperparameter tuning*, in which key parameters controlling tree depth, learning speed, and regularization strength are selected using only subsamples of the training data and are fixed prior to evaluation on future observations. Together, these

mechanisms make tree-based ensemble methods conservative prediction tools rather than unrestricted black boxes.

All models are trained using rolling windows and evaluated strictly out-of-sample. For each evaluation period, models are estimated using only information available prior to the prediction date and are tested on subsequent observations. At no point are outcomes from future periods used in training, tuning, or feature construction. Using rolling training windows ensures that predictive performance is evaluated under comparable information sets across time by maintaining a constant lookback period. This approach avoids conflating predictive improvements with changes in data availability, mitigates bias arising from slow-moving structural changes in VC markets, and allows model performance to be compared consistently across evaluation periods. Throughout the paper, the emphasis is therefore on out-of-sample predictive performance rather than in-sample fit.

4.2 Illustrative single-feature exercise

I provide an illustration of how different prediction models process information using a simple classification exercise. I rely on a single explanatory variable, the number of distinct investors that have previously invested in a startup prior to its Series A round, and use data observed prior to 2019 to predict whether startups raising a Series A round in 2019 or later eventually complete an IPO.

The models produce predicted probabilities, taking values in the unit interval. For the purpose of the figure, these probabilities are converted into binary predictions using a fixed cutoff of 0.40 for this illustration. This cutoff is intentionally chosen ex post to illustrate that there exist probability thresholds under which more flexible models can outperform a linear benchmark in classification. The choice is purely illustrative and is not meant to provide a fair or optimal comparison across models; it plays no role in the main analysis.

The results are reported in Figure 1. The horizontal axis reports the number of prior investors, while the vertical axis indicates the realized IPO outcome. Two summary statis-

tics are shown in the figure: precision and the F1 score. Precision measures the share of predicted IPOs that are ultimately correct and therefore captures how informative a positive signal is in economic terms. The F1 score combines precision and recall—the share of IPOs correctly identified by the model—and summarizes how well the model balances identifying rare successes while limiting false positives. Given the extreme rarity of IPOs, both statistics are naturally low; their role here is purely illustrative. For this particular cutoff, XGBoost attains higher precision and F1 scores than logit, demonstrating that nonlinear models can extract predictive structure even in very sparse settings.

The figure highlights a fundamental difference between linear and nonlinear prediction models. Logit maps the single predictor into a monotonic probability function, which induces a single implicit cutoff: above a given number of prior investors, predicted outcomes switch mechanically from success to failure. As a result, the model effectively delivers a pass–fail rule that partitions the data into two regions.

By contrast, tree-based models generate piecewise decision rules. Rather than imposing a single threshold, they identify multiple intervals of the predictor for which predictions differ. In the figure, this leads to clusters of correct predictions at non-monotonic values of the number of prior investors, reflecting patterns that are not captured by a linear specification.

This distinction has important conceptual implications. A linear model may suggest a simple narrative interpretation, such as fewer investors being associated with higher success, while a more flexible model reveals that predictive structure may depend on finer-grained patterns that are not monotonic and do not admit a single directional explanation. When extended to higher-dimensional settings with many features, these differences imply that flexible models can uncover economically meaningful patterns and generate questions that linear analyses are not able to formulate.

The purpose of this exercise is not to claim strong predictive power based on a single variable, but to illustrate how modern prediction methods can extract structured signals from noisy VC data. In the next section, I extend this approach to a comprehensive feature

set and study how such predictive signals perform at the investor level by comparing them with the realized outcomes of actual investors.

5 Main Results

5.1 Investment performance with ML signals

I start by evaluating whether and how predictive gains from ML models translate into economically meaningful investment performance. The objective of this analysis is to construct a transparent and replicable benchmark that mirrors the information and constraints faced by real investors. Performance is assessed at the portfolio level using outcome-based metrics based on startup success events—follow-on financing, acquisition, and initial public offering—which are standard proxies for value creation in the VC literature given the extreme skewness of returns and the fact that most investments yield either zero or a small number of outsized successes (see, e.g., Cumming and MacIntosh (2003), Cochrane (2005), Abuzov (2019), Nanda et al. (2020), Hellmann et al. (2021)).

5.1.1 Methodology

To ensure that the benchmark reflects the information environment faced by investors at the time of investment, I rely exclusively on past and publicly available data, thereby avoiding forward-looking and data-availability biases. The analysis is conducted from the perspective of an investor who allocates capital across a portfolio of startups in a given calendar year, and performance metrics are defined so as to be directly comparable across realized investor portfolios and model-based counterfactual portfolios.

Performance metric. The main performance metric is the *hit rate*. For a portfolio constructed in year t with investment budget k_t, the hit rate is defined as

$$\text{HitRate}_t \;=\; \frac{1}{k_t}\sum_{i=1}^{k_t} \mathbb{1}\{\text{startup } i \text{ is successful}\}, \tag{1}$$

where success is measured according to one of three outcomes: a follow-on financing round, an acquisition, or an initial public offering. The hit rate has a direct economic interpretation as the fraction of investments in the portfolio that ultimately achieve the specified success outcome.[5] The same definition is applied to realized investor portfolios and to counterfactual portfolios constructed using model predictions.

Investor benchmarks. For each calendar year, I compute the cross-sectional distribution of hit rates across all *active* investors, where active investors are defined as those making at least five investments in that year. From this distribution, I compute benchmark hit rates at selected percentiles (50th, 75th, and 90th). These percentiles summarize investor performance and serve as the primary benchmarks against which model-based portfolios are compared. Throughout, references to investor percentiles refer to percentiles of the hit-rate distribution, not to investment activity.

Fixing portfolio size. To ensure comparability in scale, model-based portfolios are constructed under an explicit investment budget. For each year t, I set the portfolio size k_t equal to the number of investments made by an investor at the 75th percentile of the *investment count* distribution in that year. This choice fixes the size of the counterfactual portfolio at a level typical of active professional investors. Importantly, investment-count percentiles are used only to determine portfolio size and do not enter the performance benchmarks.

[5] As emphasized in Strebulaev and Dang (2024), *"If you take out just one extremely successful startup, many of the funds that were outperforming nine out of every ten comparable VC funds suddenly become only a little better than the median."* This observation highlights why portfolio-level success rates, which capture the realization of rare extreme outcomes that drive VC returns, are a natural object for evaluating VC performance, rather than average returns, which are typically unobserved at the deal level in VC data.

Model-based portfolios. Given k_t, I construct a counterfactual ML portfolio for each year by ranking all startups that receive funding in year t by their predicted probability of success and selecting the top k_t startups. The resulting portfolio is evaluated using the same hit-rate metric as investor portfolios:

$$\text{HitRate}_t^{\text{ML}}(k_t) \;=\; \frac{1}{k_t} \sum_{i \in \text{Top-}k_t} \mathbb{1}\{\text{startup } i \text{ is successful}\}. \tag{2}$$

This construction relies exclusively on rankings of predicted probabilities rather than on fixed probability thresholds, ensuring that portfolio formation adapts endogenously to the distribution of model signals in each year.

Evaluation. I generate predictions using rolling ten-year training windows. Models are estimated on historical data to predict outcomes in the subsequent year, starting with the 2000–2009 window to predict 2010 and rolling forward annually through 2021. To ensure comparability in scale, all model-based portfolios assume an investment budget equal to the number of investments made by a 75th-percentile investor in the corresponding year.

Stage-specific analysis. In addition to aggregate results pooling all investments, I examine heterogeneity across the startup lifecycle by repeating the analysis separately for early-stage (pre-seed, seed, and Series A), mid-stage (Series B and Series C), and late-stage (Series D and later) investments. The same performance metric, benchmarking procedure, and portfolio-size normalization are applied within each stage. This allows for a consistent comparison of predictive gains across stages without changing the evaluation framework.

5.1.2 Performance Comparison

Figure 2 reports year-by-year hit rates for portfolios constructed using ML predictions (hereafter referred to as "ML-based portfolios") and benchmarks them against the 50th (p50), 75th (p75), and 90th (p90) percentiles of the investor hit-rate distribution in each year,

pooling investments across all financing stages. Each panel corresponds to a specific success definition and prediction model. The solid line reports the hit rate of the ML-based portfolio, while the dashed lines report the corresponding investor benchmark hit rates.

Across success definitions, the ML-based portfolios consistently achieve hit rates that exceed those of the median investor and, in many cases, approach or surpass the performance of top-quartile (75th percentile) investors. This pattern is especially pronounced for rare outcomes. For IPOs, ML portfolios constructed using flexible models frequently attain hit rates comparable to, and in several years exceeding, those of investors at the 90th percentile of the hit-rate distribution. For example, XGBoost achieves IPO hit rates ranging from roughly 40% in 2011 to nearly 80% in 2012, outperforming all investor benchmarks in those years. By contrast, performance differences are more muted for follow-on financing outcomes, which are substantially more common. In these cases, ML-based portfolios still outperform the median investor but exhibit smaller gains relative to top-performing investors. Taken together, the results indicate that predictive gains from ML are not uniform across outcomes, but instead increase sharply with outcome rarity.

A clear hierarchy also emerges across prediction models. Portfolios constructed using tree-based ensemble methods consistently outperform those based on logit, with XGBoost improving upon the linear benchmark and CatBoost delivering further, though more modest, gains. This ordering holds across success definitions and over time. The fact that more flexible models systematically dominate less flexible ones suggests that linear index models fail to capture important predictive structure in VC data. In particular, these gains arise from improved differentiation of rare, high-impact successes from the broader population of investments. Rather, gains appear to arise from the ability of ML methods to exploit nonlinearities and interactions in the information available at the time of investment.

Importantly, these patterns speak to where predictive gains arise rather than why they arise. The concentration of gains in rare outcomes and the superior performance of flexible models indicate that ML is particularly effective at distinguishing among investments with

low ex ante success probabilities. At the same time, these results leave open the question of whether such gains reflect genuinely superior selection within comparable opportunity sets, or instead arise mechanically from differences in the types of deals being selected—for example, across stages of the startup lifecycle. The next set of analyses addresses this distinction directly by examining how predictive performance varies across early-, mid-, and late-stage investments.

Figure 3 reports the performance of ML-based portfolios restricted to early-stage investments (seed and Series A). The results closely mirror those obtained in the full sample. ML-based portfolios consistently outperform the median investor across outcomes and years, and the hierarchy across models is preserved, with tree-based methods dominating logit and CatBoost generally outperforming XGBoost. As in the full sample, relative gains increase with outcome rarity, with the largest improvements observed for IPOs. In the later part of the sample, the performance gap between flexible ML models and logit becomes especially pronounced for early-stage IPO outcomes. In several years between 2017 and 2019, even 90th-percentile investors achieve near-zero IPO hit rates, while ML-based portfolios constructed using XGBoost or CatBoost attain hit rates between 30–50%. Overall, ML performance does not weaken when attention is restricted to earlier stages, consistent with the fact that early-stage investments involve greater uncertainty and less publicly observable information than later-stage deals.

These findings provide an important nuance to recent evidence on the use of historical data in VC decision-making. Bonelli (2025) shows that VCs who adopt data-related capabilities improve their ability to forecast relatively frequent outcomes such as follow-on rounds, but do not realize gains for major successes such as IPOs. The results here indicate that this lack of improvement is unlikely to reflect an absence of predictive signal in the data. To the contrary, mechanically applying ML signals to early-stage investments yields substantial improvements in IPO prediction, often outperforming even top-decile investors. The contrast between these counterfactual gains and the limited improvements observed in

practice points to behavioral or operational constraints in how investors adopt and act on data-driven signals, rather than to fundamental informational limits.

More broadly, my results also speak to the "curse of rarity" phenomenon, which argues that algorithms may struggle to predict rare outcomes due to the scarcity of relevant examples (Agarwal et al., 2024; Peukert et al., 2024). While such constraints are likely to be important in settings involving extreme tail events, they do not appear to bind in the VC context I study here. IPOs are rare but not exceedingly so—approximately 9% of financing rounds in the sample are associated with startups that eventually go public—and ML models trained on historical VC data are able to extract meaningful and stable predictive signals for these outcomes. Taken together, the evidence suggests that, in this setting, outcome rarity per se is not the primary obstacle to data-driven investment success.

Figure 4 reports results for mid-stage investments (Series B and C). The main patterns observed earlier remain in place: (i) ML-based portfolios continue to outperform the median investor across outcomes and years, particularly when using tree-based models, and (ii) the hierarchy across models is preserved, with CatBoost generally outperforming XGBoost and both dominating logit. As in previous subsamples, relative gains increase with outcome rarity, with the largest advantages observed for IPOs. At the same time, the magnitude of the performance differential between ML-based portfolios and investor benchmarks is smaller than at earlier stages. This attenuation is consistent with the greater availability of publicly observable information as investments progress toward later stages, as discussed above, which reduces the scope for extracting incremental predictive value from historical cross-sectional patterns. Nonetheless, for IPO outcomes, ML-based portfolios constructed using tree-based methods continue to outperform even the 90th-percentile investor in most years, whereas logit-based portfolios exhibit substantially noisier performance.

Finally, figure 5 reports results for late-stage investments (Series D and beyond). While the hierarchy across models and outcomes remains qualitatively intact, ML-based portfolios no longer exhibit a systematic advantage over top-performing investors. Tree-based models

typically outperform the median investor and often exceed the 75th-percentile benchmark, but they do not consistently outperform the 90th-percentile investor and in several years perform comparably to, or below, that benchmark. This pattern contrasts with the earlier-stage results and indicates a further narrowing of the performance gap as investments approach resolution. For IPO outcomes in particular, ML-based portfolios remain competitive but no longer dominate top investors in a systematic manner. Overall, the late-stage evidence suggests that the incremental advantage from data-driven prediction diminishes as uncertainty declines and outcomes become increasingly shaped by information and selection processes that are observable to market participants.

The late-stage results are also consistent with the importance of access to deal flow for VC investors. For instance, Nanda et al. (2020) show that performance persistence among VC firms operates largely through access: early success improves investors' ability to participate in later-stage rounds and larger syndicates, rather than reflecting persistent superior selection skill. In this context, the attenuation of ML-based performance at late stages is consistent with the idea that access constraints, rather than informational limitations, play a more prominent role in later rounds. Late-stage rounds are more competitive and syndicate-driven, and participation is often restricted to investors with established relationships and reputational capital. As a result, even when predictive signals are available, investors without access to the relevant deal flow may be unable to act on them. Importantly, this mechanism appears less binding at earlier stages, where access barriers are lower and where ML-based portfolios exhibit their strongest relative performance.

Taken together, these results highlight clear scope and limits for ML-based prediction in the context of VC. ML-based portfolios consistently outperform typical investors and deliver their largest gains at early stages, particularly for rare outcomes such as IPOs, where predictive uncertainty is highest and access constraints are weakest. These advantages attenuate at mid stages and largely disappear at late stages, where investments are closer to resolution and participation increasingly reflects access to competitive deal flow rather than informa-

tional advantages. The evidence therefore suggests that the effectiveness of ML in VC is not constrained by a lack of predictive signal in the data, but instead varies systematically with the informational environment and market frictions faced by investors. ML appears most valuable precisely in settings where uncertainty is high and traditional advantages tied to reputation and access play a smaller role.

In general, these findings reject Hypothesis **H2** and the strong implication of Hypothesis **H1** that predictive gains are too unstable to be economically meaningful. First, gains from ML are large in most years, especially relative to the median investor, across outcomes and model specifications. Second, ML-based gains are strongest for the most valuable and rare outcome, IPOs. As such, the limited ability of investors to materialize these potential gains is unlikely to reflect a lack of predictive scope in the data or the possibility that ML signals merely capture non-value predictability, such as the likelihood of follow-on funding, rather than the probability of high-value outcomes like IPOs.

5.2 Who benefits from ML signals

Having established that ML-based portfolios generate economically meaningful gains in aggregate, this section examines how these gains map into investor-level outcomes. Specifically, I test Hypothesis **H3**, which concerns whether predictive gains translate into systematic and persistent advantages for individual investors, before turning to Hypothesis **H4**, which examines whether investor-level constraints and distance—such as differences in investment scope, specialization, or preferences—help explain variation in these gains.

To quantify investor-level gains from ML signals, I define an investor–year measure of *benefit* that compares realized investment performance to a counterfactual ML-based benchmark. For each investor i and year t, benefit is defined as the arithmetic difference between the hit rate implied by the ML-based portfolio and the hit rate realized by the investor's actual portfolio:

$$Benefit_{i,t} = HitModel_{i,t} - HitActual_{i,t}. \tag{3}$$

The counterfactual ML-based portfolio is scaled to match the investor's realized level of investment activity in year t, so that benefit captures differences in selection quality rather than differences in investment scale. A positive value of $Benefit_{i,t}$ indicates that, holding fixed the number of investments undertaken, the investor would have achieved a higher success rate by following ML signals, while a negative value indicates that the investor's realized choices outperformed the ML-based benchmark. Because benefit is defined as an arithmetic difference in hit rates, a value of 0.20 corresponds, for example, to an increase in the hit rate from 0.10 to 0.30, rather than a 20% relative increase.

This construction implicitly assumes that an investor can invest in any startup funded in the market in a given year, an assumption that is standard in counterfactual portfolio exercises but not without limitations. In practice, investors may face frictions that prevent immediate participation in some deals identified by ML signals. For example, applying ML predictions to recently completed financings may identify startups with high long-run success potential, but access to these opportunities may only arise in subsequent rounds. Conversely, applying ML signals to deals in which an investor is already participating restricts the counterfactual to the investor's own opportunity set, limiting the scope for selection gains. To discipline the counterfactual exercise and avoid mechanically selecting negligible-probability outcomes, the ML-based portfolio is restricted to startups whose predicted success probability exceeds 10%. The benefit measure should therefore be interpreted as capturing the potential gains from improved selection in a frictionless benchmark environment, rather than as a literal prescription for immediate investment decisions in every identified deal.

Figure 6 reports the share of active investors whose counterfactual ML-based portfolio would have achieved a higher hit rate than their realized portfolio, separately by outcome and ML model. Across all outcomes, positive benefit is widespread. For acquisitions, the share of investors with positive benefit typically ranges between roughly 60% and 100% for tree-based models, and exceeds 50% in most years even under logit, with only a few isolated years falling materially below that threshold. For follow-on financings, positive benefit is

somewhat less pervasive but remains common: the share of investors with positive benefit generally lies between 50% and 80% for CatBoost and XGBoost, and between 40% and 70% for logit in most years. The prevalence of positive benefit is highest for IPO outcomes. For tree-based models, more than 80% of investors exhibit positive benefit in nearly every year, with several years approaching universal gains. Even under logit, the share of investors with positive benefit for IPOs typically exceeds 70%, aside from a small number of years characterized by sharp but transitory declines. Overall, the figure indicates that ML-based signals would have improved selection for a large majority of investors in most periods, particularly for high-value outcomes, while also exhibiting meaningful time variation and systematic differences across model classes.

Figure 7 complements the prevalence results by reporting the magnitude of investor-level benefits over time, measured by the median benefit and the interquartile range. Across all outcomes and models, median benefit is positive in most years, providing direct support for Hypothesis **H3** that ML-related predictive advantages translate into economically meaningful gains at the investor level. The magnitude of these gains varies systematically across outcomes. For IPOs, median benefit is large, typically ranging between 0.3 and 0.7 for tree-based models, and remains positive even under logit in most years. Median benefits are more moderate for acquisitions and smaller for follow-on financings, mirroring the ordering by outcome rarity observed in the portfolio-level analysis.

Importantly, the large median benefits observed for IPOs are unlikely to be driven mechanically by late-stage investments that are already close to exit. As shown in the previous section, ML-based gains are strongest at early stages and attenuate at later stages, ruling out an interpretation in which the results are driven by predicting IPOs among startups that have already progressed to advanced rounds. The interquartile ranges are also relatively tight for IPOs compared to other outcomes, indicating that positive benefits are broadly shared across investors in a given year rather than concentrated among a small subset.

Taken together, these results motivate a deeper analysis of Hypothesis **H3** by examining

whether investor-level benefits from ML signals are sufficiently stable over time to support sustained use in practice. To this end, I next describe the empirical methodology used to study the persistence and determinants of ML-related benefits at the investor level.

5.2.1 Methodology: Investor-Level Persistence and Constraints

This subsection describes the empirical framework used to study the persistence and determinants of investor-level benefits from ML signals. The analysis is conducted at the investor–year level and focuses on within-investor variation over time. All specifications include investor and year fixed effects. Investor fixed effects control for time-invariant differences across investors, such as baseline skill, reputation, network access, and long-run investment focus, while year fixed effects absorb common shocks to the VC environment, including changes in market conditions, exit opportunities, and aggregate predictability. Identification therefore comes from changes in benefit and related covariates within the same investor over time, net of aggregate fluctuations. I provide the precise definitions of all variables discussed in this subsection in Panel A of Appendix D.

Persistence at the extensive margin. I begin by examining whether benefiting from ML signals persists at the extensive margin by estimating

$$HasBenefit_{i,t} = \alpha_i + \delta_t + \beta\, HasBenefit_{i,t-1} + \varepsilon_{i,t}, \tag{4}$$

where $HasBenefit_{i,t}$ is an indicator equal to one if investor i exhibits positive benefit in year t, and zero otherwise. This specification asks whether an investor who benefits from ML signals in one year is more likely to benefit again in the following year, holding constant investor-specific characteristics and year-specific shocks. The coefficient β captures persistence in the probability of realizing any positive benefit from ML signals and provides a conservative test of Hypothesis **H3**, abstracting from variation in benefit magnitude.

Persistence at the intensive margin. I next study persistence in the magnitude of investor-level benefits by estimating

$$Benefit_{i,t} = \alpha_i + \delta_t + \beta\, Benefit_{i,t-1} + \varepsilon_{i,t}, \tag{5}$$

where $Benefit_{i,t}$ denotes the continuous benefit measure defined above. This specification tests whether investors who experience large benefits from ML signals in one year tend to experience similarly large benefits in subsequent years, conditional on investor and year fixed effects. Values of β substantially below one indicate mean reversion in benefits over time, even when ML-related gains are large on average. Together with the extensive-margin analysis, this specification provides a comprehensive test of Hypothesis **H3**.

Distance and investor constraints. To assess whether investor-level constraints help explain heterogeneity in ML benefits, I introduce a measure of distance between the set of startups predicted by ML to be successful and the investor's realized investment activity in a given year. Distance captures how far predicted successes lie from the investor's typical investment domain along several dimensions: stage, sector, geography, and market *hotness*. Hotness refers to the intensity of recent investment activity in a startup's segment, measured by the volume and frequency of financings in the corresponding stage–sector–geography cell in recent years. Intuitively, hotness captures whether ML-predicted successes are concentrated in market segments that are currently crowded or competitive relative to an investor's historical focus.

For each investor–year, distance is computed separately along each dimension by comparing the distribution of characteristics of ML-predicted successful startups to the distribution of the investor's realized investments in that year. Each component is standardized to have mean zero and unit variance, and the overall distance index is constructed as an equally weighted average of the standardized components. The exact definitions and mathematical expressions for each distance component, as well as for the composite distance index, are

reported in Panel B of Appendix D. I then estimate

$$Benefit_{i,t} = \alpha_i + \delta_t + \gamma\, Distance_{i,t} + \varepsilon_{i,t}, \tag{6}$$

which isolates the association between distance and investor-level benefit. With investor and year fixed effects, the coefficient γ measures whether deviations from an investor's usual activity domain are associated with systematically lower benefits from ML signals. This specification provides a first test of Hypothesis **H4**.

Distance versus persistence. I next combine persistence and distance by estimating

$$Benefit_{i,t} = \alpha_i + \delta_t + \beta\, Benefit_{i,t-1} + \gamma\, Distance_{i,t} + \varepsilon_{i,t}. \tag{7}$$

which assesses whether limited persistence in benefits reflects purely statistical mean reversion or is instead related to time-varying feasibility constraints faced by investors. Comparing the estimate of β in this specification to that obtained in the persistence-only regression reveals the extent to which distance accounts for the attenuation of benefits over time.

Decomposing distance. Finally, I decompose the aggregate distance measure into its underlying components and estimate

$$Benefit_{i,t} = \alpha_i + \delta_t + \beta\, Benefit_{i,t-1} + \sum_k \gamma_k\, Distance^k_{i,t} + \varepsilon_{i,t}, \tag{8}$$

where $Distance^k_{i,t}$ denotes distance along dimension k, including stage, sector, geography, and market hotness. This specification identifies which dimensions of distance are most strongly associated with lower investor-level benefits, shedding light on the specific constraints that limit the ability to exploit ML signals. Together, these specifications provide a systematic framework for evaluating the persistence of ML-related benefits and the role of investor constraints, directly addressing Hypotheses **H3** and **H4**.

5.2.2 Summary Statistics

Table 1 presents summary statistics for the investor–year variables, pooling across investment stages. Several features stand out. First, investor-level benefits from ML signals are large for IPO outcomes and positive both on average and at the median for acquisitions and follow-on financings. In Panel A, the mean IPO benefit equals 0.50 under XGBoost and 0.45 under CatBoost, compared to 0.21 under logit, with corresponding medians of 0.54, 0.48, and 0.20, respectively. For acquisitions, mean benefit is smaller but substantial at 0.31 (XGBoost) and 0.32 (CatBoost), versus 0.15 under logit, while for follow-on financings mean benefit ranges between 0.12 and 0.16 for tree-based models and 0.07 for logit. These differences reflect gaps between realized and model-implied hit rates, particularly for IPOs: in Panel A, the mean realized IPO hit rate is 0.084 (median 0.000), whereas the mean model-implied IPO hit rate is 0.58 under XGBoost and 0.53 under CatBoost. By contrast, realized follow-on hit rates are high on average (mean 0.81), leaving less scope for improvement.

Panels B through D of Table 1 decompose these patterns by investment stage and show that large IPO-related benefits are not driven by late-stage predictability. At the early stage (Panel B), the mean realized IPO hit rate is only 0.024 with a median of zero, yet mean benefit remains large at 0.40 under XGBoost and 0.38 under CatBoost. At the mid stage (Panel C), realized IPO hit rates increase to a mean of 0.144, with mean benefits of 0.54 and 0.53 under XGBoost and CatBoost, respectively. At the late stage (Panel D), realized IPO hit rates are substantially higher (mean 0.248), and average benefits are correspondingly smaller at 0.27 (XGBoost) and 0.29 (CatBoost). Overall, the composite distance index is centered close to zero at early and mid stages but shifts downward at the late stage, consistent with predicted successful opportunities lying further from investors' realized activity domains in later rounds. Together, these summary statistics provide a concrete sense of the magnitude, variation, and stage dependence of ML-related benefits at the investor level and motivate the persistence and distance analyses that follow.

5.2.3 Persistence of Investor-Level ML Benefits

Next, I examine whether investor-level gains from ML signals persist over time. If benefits are short-lived, investors may be unable to translate predictive signals into sustained improvements in their own deal selection. I first study persistence in the sign of benefit using descriptive evidence, before turning to the regression-based analyses described earlier.

Figure 8 reports the persistence of the sign of investor-level benefit over time. For each outcome and year, the figure plots three probabilities: the unconditional probability that an investor exhibits positive benefit; the probability of positive benefit conditional on having been positive in the previous year; and the probability of positive benefit conditional on having been non-positive in the previous year. Across all success outcomes, a clear ordering holds in every year. The probability of positive benefit is highest when conditioning on prior positive benefit, intermediate unconditionally, and lowest when conditioning on prior non-positive benefit. This pattern indicates meaningful state dependence in benefit realization.

The strength of sign persistence differs across outcomes. For IPOs, the probability of positive benefit conditional on prior positive benefit is typically between 0.8 and 1.0, with the unconditional probability also being consistently high. Acquisitions exhibit slightly weaker but still substantial persistence, with both unconditional and conditional probabilities generally exceeding 0.6 after the earliest sample year. Persistence is weakest for follow-on financings, where both probabilities fall below 0.6 in several years at the beginning and end of the sample, without a clear time pattern.

Economically, these patterns imply that while ML-based signals offer widespread opportunities for improved deal selection—as reflected in high unconditional probabilities—investors who fail to realize positive benefit in one year are significantly less likely to do so in subsequent years. Conversely, investors who do realize positive benefit are more likely to continue doing so. This asymmetry suggests that differences in benefit realization may reflect persistent investor-specific factors rather than purely transitory shocks.

Table 2 reports estimates from regressions of $HasBenefit_{i,t}$ on its lag, including investor

and year fixed effects. Across outcomes and models, the estimates reveal strong mean reversion at the extensive margin. In Panel A (all stages), the coefficient on $HasBenefit_{i,t-1}$ is negative and statistically significant in nearly all specifications. For example, for IPO outcomes, the estimated coefficient equals -0.102 under logit, -0.144 under XGBoost, and -0.085 under CatBoost. For follow-on financings, the corresponding coefficients are even larger in magnitude, ranging from -0.112 to -0.151 across models. These estimates imply that, holding fixed investor-specific characteristics and common year effects, an investor who experienced positive benefit in year $t-1$ is 8 to 15 percentage points less likely to experience positive benefit in year t.

The same pattern holds across investment stages. In the early-stage sample (Panel B), mean reversion is particularly pronounced: for IPOs, the coefficient on $HasBenefit_{i,t-1}$ is -0.136 under XGBoost and -0.118 under CatBoost, while follow-on outcomes exhibit coefficients below -0.15 across models. In the mid-stage sample (Panel C), negative coefficients remain statistically significant for most outcomes, though somewhat smaller in magnitude. By contrast, in the late-stage sample (Panel D), estimates are closer to zero and often imprecisely estimated, consistent with a narrower opportunity set and reduced within-investor variation at later stages. The negative sign of the coefficient appears under logit as well as under tree-based models, indicating that mean reversion at the extensive margin is not driven by a particular modeling approach.

Taken together, these results show that while many investors experience positive benefit in a given year, such states are not stable within investors over time once persistent investor heterogeneity and aggregate conditions are controlled for. Instead, positive benefit tends to reverse in subsequent years, helping to reconcile the large and widespread ML-related gains documented earlier with investors' difficulty in sustaining them through repeated reliance on ML-based signals. This extensive-margin mean reversion motivates a closer examination of whether a similar pattern holds for the magnitude of benefit, which I turn to next.

Figure 9 examines persistence in the magnitude of investor-level benefit by plotting av-

erage benefit in year t against quintiles of lagged benefit in year $t-1$, where the first quintile corresponds to the lowest lagged benefit and the fifth to the highest. Across all outcomes and models, higher lagged-benefit quintiles are associated with higher average benefit in the subsequent year, indicating that investors who benefited more from ML signals in the past tend to continue outperforming those who benefited less. This ordering is not purely mechanical: in a few cases, the profile is not strictly monotone across adjacent quintiles, underscoring that benefit realizations are noisy. Average benefit levels are highest for IPO outcomes, followed by acquisitions and then follow-on financings, and are systematically larger under tree-based models—particularly CatBoost—than under logit.

Despite this reversion, benefit differences across lagged-benefit quintiles remain significant. Investors in higher lagged-benefit quintiles exhibit substantially higher average benefit in year t than those in lower quintiles, suggesting incomplete mean reversion. In particular, while average benefit among top-quintile investors declines relative to its lagged level, it remains positive and sizeable, especially for IPO outcomes. Reversion is more pronounced for acquisitions and strongest for follow-on financings, where differences across quintiles narrow considerably. These patterns indicate that past ML-related gains retain predictive content for future relative performance.

Table 3 examines persistence at the intensive margin by regressing $\text{Benefit}_{i,t}$ on its lag. Across most outcomes, stages, and models, the coefficient on $Benefit_{i,t-1}$ is negative, indicating substantial mean reversion in benefit magnitudes. In Panel A (all stages), the estimates are economically large and statistically significant for follow-on financings, with coefficients ranging from -0.178 under logit to -0.137 under CatBoost.

There is no systematic hierarchy in the strength of reversion across success outcomes or prediction models. Instead, the magnitude and precision of the estimates vary primarily across investment stages. In the early-stage sample (Panel B), negative coefficients are large and precisely estimated across all outcomes and models—for example, for IPOs the coefficient ranges from -0.140 under logit to -0.110 under CatBoost. In the mid-stage

sample (Panel C), reversion remains present but is weaker and less uniformly significant, particularly for IPOs under tree-based models. In the late-stage sample (Panel D), estimates attenuate further and often lose statistical significance, especially for acquisitions and IPOs, consistent with reduced within-investor variation and a narrower investment opportunity set at later stages. Taken together, these results indicate that large ML-related benefits tend to dissipate within investor over time, even when relative performance differences persist, reinforcing the view that exploiting ML signals at scale is constrained by dynamic limits faced by investors rather than by a lack of predictive signal.

5.2.4 Distance and the Limits to Exploiting ML Signals

The preceding results show that predictive benefit from ML signals is widespread but strongly mean-reverting at the investor level. This subsection studies whether portfolio distance—a proxy for frictions in investors' ability to act on ML signals—is systematically related to predictive benefit and to its persistence. The analysis therefore distinguishes between two questions: whether ML gains tend to arise when predicted successful investments lie outside investors' typical investment domains, and whether such distance helps explain why these gains are difficult to sustain over time.

Table 4 relates investor-level benefit to portfolio distance, measured as a composite index capturing differences between realized and model-implied counterfactual portfolios. Across most specifications, greater distance is associated with higher predictive benefit, consistent with the idea that ML signals identify successful investments that fall outside investors' usual investment scope. Distance therefore helps explain where ML-related gains originate in the cross section of investors.

In Panel A (all stages), distance is positively and significantly related to benefit for acquisitions under XGBoost and CatBoost, with coefficients of 0.036 and 0.024, respectively, and for IPOs under XGBoost (0.015). The relationship is particularly pronounced at early stages (Panel B): for acquisitions, the distance coefficient ranges from 0.032 to 0.051 across

models. For IPOs, however, distance is negative and significant under CatBoost (−0.021), indicating that at very early stages predictive gains arise when ML-selected investments are closer to investors' existing activity domains. This pattern is consistent with extreme distance reflecting infeasible or inaccessible opportunities at early stages, so that benefits materialize only when ML signals point to startups that remain within an investor's effective opportunity set. At mid stages (Panel C), distance remains positively related to benefit for acquisitions and IPOs under tree-based models, while at late stages (Panel D) estimates are smaller and less precisely estimated, reflecting tighter opportunity sets and reduced variation in feasible investments.

Table 5 jointly relates benefit to both lagged benefit and portfolio distance, allowing an assessment of whether distance accounts for the observed persistence patterns. Across outcomes and stages, the coefficient on lagged benefit remains negative and statistically significant in most specifications, confirming strong mean reversion even after controlling for distance. For example, in Panel A, the coefficient on $Benefit_{i,t-1}$ for follow-on financings ranges from −0.178 under logit to −0.137 under CatBoost, closely mirroring earlier estimates.

At the same time, distance retains explanatory power in several cases. For acquisitions in Panel A, the distance coefficient remains positive and significant under XGBoost (0.028) and CatBoost (0.019), while in the early-stage sample (Panel B) distance is positively related to benefit for acquisitions across all models, with coefficients up to 0.042. These results indicate that portfolio distance explains an important component of cross-investor variation in the level of predictive benefit. However, distance does not materially attenuate the magnitude or significance of the lagged-benefit coefficient, implying that it does not account for the strong within-investor mean reversion documented earlier.

Table 6 decomposes portfolio distance into its constituent dimensions—market hotness, sectoral distance, and geographic distance—while simultaneously controlling for lagged benefit. Across outcomes and stages, the coefficient on lagged benefit remains negative and often statistically significant, confirming that mean reversion at the intensive margin persists even

after accounting for detailed distance measures. Turning to the distance components, no single dimension consistently explains the persistence of predictive benefit. Sectoral distance is occasionally positively related to benefit, particularly for acquisitions and IPOs at early and mid stages, while geographic distance and distance in market hotness exhibit limited and unstable explanatory power. Importantly, these effects do not materially weaken the role of lagged benefit.

Overall, the evidence indicates that portfolio distance helps explain where ML-related gains arise across investors, but does not explain why such gains dissipate within investors over time. Static differences in sectoral focus, geographic exposure, or market positioning therefore do not appear to be the primary constraint limiting the persistence of predictive benefit, as emphasized in Hypothesis **H4**. Instead, the results point to dynamic limits—such as adjustment frictions, learning, and changing opportunity sets—as the main forces preventing investors from sustaining ML-related advantages over time.

5.3 The Role of Technology Access

I conclude by studying whether changes in investors' access to ML technology impact the persistence of ML-related predictive benefits. This analysis is motivated by the preceding results showing that predictive benefit is widespread but strongly mean-reverting and not explained by static access constraints such as sectoral focus, geography, or market positioning. I examine whether changes in the availability and cost of ML infrastructure affect investors' ability to translate predictive signals into realized gains. Building on the idea that the economic value of ML depends on access to sophisticated prediction tools (DeMiguel et al. (2023); Kaniel et al. (2023)), I exploit the staggered release of widely adopted open-source ML libraries that substantially lowered the fixed costs of implementing advanced prediction methods as a source of plausibly exogenous variation in access to ML technology.

5.3.1 Methodology

The analysis exploits discrete reductions in the cost and complexity of implementing modern machine-learning methods induced by the public release of widely adopted open-source ML libraries between 2015 and 2017. These releases include scalable gradient-boosting and neural-network frameworks that made advanced prediction methods computationally feasible and substantially easier to deploy using standard programming environments. Collectively, these libraries lowered fixed costs related to model implementation, hyperparameter tuning, and large-scale training, thereby expanding access to sophisticated ML tools beyond specialized research teams. The set of libraries, their release dates, and the code used to construct event-time indicators are reported in Panel B of Appendix C.

Panel structure and exposure definition. The unit of observation is an investor–year. To capture heterogeneity in how investors are affected by reductions in ML implementation costs, I construct an investor-level measure of exposure to ML-related gains prior to the technology shocks. Specifically, for each investor i, exposure is defined as the average value of $Benefit_{i,t}$ over the pre-shock period, measured using the same portfolio construction as in the main analysis. This produces a single, time-invariant exposure measure for each investor, which is then assigned to all investor–year observations in the event-study window. Exposure is computed exclusively using pre-shock data and is held fixed throughout the analysis.

The measure captures how much an investor's realized portfolio decisions differed from ML-implied counterfactual portfolios before advanced ML tools became widely accessible. As such, exposure does not require that an investor actively used ML methods in the pre-shock period. Instead, it measures the extent to which an investor's historical investment choices were aligned with, or deviated from, the selections implied by ML-based predictions.

Event-study specification. I estimate the following event-study regression:

$$Benefit_{i,t} = \sum_{\tau \neq -1} \beta_\tau \, \mathbb{1}\{t - T = \tau\} \times Exposure_i + \alpha_i + \delta_t + \varepsilon_{i,t}, \tag{9}$$

where T denotes the year of the technology shock, $\mathbb{1}\{t - T = \tau\}$ are event-time indicators indexed relative to the year preceding the shock ($\tau = -1$), α_i are investor fixed effects, and δ_t are year fixed effects. The coefficients β_τ trace the evolution of predictive benefit around the technology shock as a function of pre-shock exposure.

Interpretation of coefficients. Each coefficient β_τ measures the differential change in predictive benefit at event time τ associated with a one-unit increase in pre-shock exposure. Because exposure is fixed prior to the shock and investor fixed effects absorb time-invariant differences across investors, identification comes from within-investor changes in benefit that coincide with the expansion of ML technology access. Negative post-shock coefficients indicate that investors who previously experienced larger ML-related benefits exhibit larger subsequent declines in benefit following the reduction in ML implementation costs.

Identification and assumptions. The identifying assumption is that, absent the technology shock, predictive benefit would have evolved similarly over time for investors with different levels of pre-shock exposure. This assumption is evaluated by examining the pre-shock event-time coefficients, which should be statistically indistinguishable from zero if no differential pre-trends are present. Year fixed effects absorb aggregate changes in investment conditions, exit markets, and startup quality that affect all investors simultaneously.

Scope and interpretation. This design does not require direct observation of ML adoption and does not assume that investors immediately adopt ML methods following the shock. Instead, it tests whether reductions in the cost and availability of ML technology are associated with systematic changes in investors' relative performance vis-à-vis ML-based bench-

marks. The estimates therefore capture the equilibrium implications of expanded technology access, rather than individual adoption decisions.

5.3.2 The Evolution of Predictive Benefit under Technology Access

Figure 10 reports event-study estimates of the evolution of predictive benefit around the release of open-source ML libraries, interacted with investors' pre-shock exposure to ML-related gains. The coefficients measure differential changes in benefit for investors with higher exposure relative to those with lower exposure, with event time normalized to the year preceding the technology shock. The estimates show no evidence of differential pre-trends: coefficients in the years leading up to the shock are close to zero and statistically indistinguishable from zero. Following the shock, the coefficients turn negative and increase in magnitude over time, indicating a sustained decline in predictive benefit for investors with higher pre-shock exposure. At longer horizons, the estimated effects reach values between approximately -0.05 and -0.10.

Because benefit is defined as an arithmetic difference in hit rates, these magnitudes correspond to declines of roughly 5 to 10 percentage points in the share of investments that realize the relevant success outcome, relative to the ML-implied counterfactual. Given baseline hit rates on the order of 5 to 15% for IPOs and 20 to 40% for acquisitions, these effects are economically large. The decline emerges gradually over the post-shock period rather than instantaneously, consistent with diffusion and learning in the adoption of ML tools rather than mechanical adjustment. Combined with the earlier evidence that portfolio distance does not explain persistence, the event-study results indicate that ML-related advantages erode as access to predictive technology expands.

Table 7 reports difference-in-differences estimates of how predictive benefit responds to the diffusion of modern ML technologies. The coefficients of interest are interactions between investor exposure and the post-shock indicator, which capture differential changes in benefit following the technology shock for investors with higher pre-shock exposure. In Panel A,

which pools all investment stages, the interaction terms are uniformly negative and statistically significant across success outcomes and prediction models. For follow-on financings, the estimates range from -0.039 to -0.043. For acquisitions, they range from -0.049 to -0.057. For IPOs, the estimates range from -0.022 under logit to -0.041 under CatBoost. Panels B and C show that these effects are largest and most precisely estimated at early and mid stages, while Panel D indicates substantially weaker and often statistically insignificant effects at late stages.

These estimates clarify the mechanism through which technology access affects predictive advantage. Investors whose realized portfolios diverged more from ML-implied selections prior to the shock experience systematically larger post-shock declines in benefit, even after conditioning on investor and year fixed effects. The fact that this pattern holds across outcomes and prediction models, and is concentrated in stages where predictive gains are largest ex ante, indicates that technology diffusion compresses cross-investor differences in selection quality rather than shifting the overall opportunity set. The post-shock adjustment therefore operates through a reduction in relative advantages previously held by a subset of investors, rather than through changes in investment scale, stage composition, or access to specific deals.

Overall, the results show that while ML-based signals provide substantial predictive power in VC, their translation into sustained investor-level gains is systematically limited. Predictive benefit is large and widespread, particularly for early-stage investments and rare success outcomes, but it is also strongly mean-reverting at the investor level. The evidence indicates that predictive advantages erode as access to ML technology expands, compressing differences in selection quality across investors. Taken together, these findings suggest that the limited ability of VC investors to materialize gains from data is better explained by frictions in technology access than by an absence of exploitable predictive signals in the data.

6 Conclusion

I study whether ML can systematically improve VC deal selection and, if so, whether and how investors can materialize the resulting estimated gains. Using a large panel of U.S. financing rounds and strictly out-of-sample predictions of follow-on, acquisition, and IPO outcomes, I show that portfolios constructed from ML signals consistently outperform the majority of VC investors at realistic investment scales in terms of success rates. These gains are particularly pronounced for early- and mid-stage investments and for rare outcomes such as IPOs, indicating that successful ventures exhibit learnable characteristics at the time of investment despite the rarity of extreme success. I further document that the benefits from ML are widespread across investors, positive on average, and persistent in relative terms, but strongly mean-reverting within investors over time.

Despite these large and systematic estimated gains, investors do not appear to fully arbitrage away these opportunities by avoiding predictably bad investments and attaining higher success rates at the individual level. I provide evidence that this gap is not primarily explained by investment stage risk, sectoral specialization, or geographic distance. Taken together, these results suggest that investment access—a primary driver of VC performance—is not the key obstacle to applying ML signals in deal selection. Instead, two structural forces limit sustained adoption. First, the predictive benefit from ML is mean-reverting: investors who stand to gain the most from algorithmic signals experience declining advantages over time. Second, I exploit a plausibly exogenous shock to the availability of ML technologies to show that reductions in the cost of implementing these tools lead to rapid erosion of cross-sectional differences in predictive advantage. These findings indicate that diffusion and competitive dynamics, rather than persistent misallocation or behavioral biases, play a central role in shaping the equilibrium use of data-driven decision tools in VC.

More broadly, my results contribute to the growing literature on the economic effects of data technologies by highlighting their limits in competitive investment environments. While ML can substantially improve prediction and selection, its benefits may be transitory and

unevenly realized once access becomes widespread. This perspective helps reconcile strong predictive performance with the limited long-run impact of algorithmic tools on investor outcomes. The framework I develop may also be informative beyond VC, including in settings where deal-level data dominate, such as leveraged buyout funds and M&A, and where prediction-driven allocation interacts with competition, access, and technological diffusion.

References

Rustam Abuzov. A little more monitoring, a little less screening: Busy venture capitalists and investment performance. *Working Paper*, 2019.

Nikhil Agarwal, Ruoxuan Huang, Andrew Moehring, Pranav Rajpurkar, Tobias Salz, and Frank Yu. Comparative advantage of humans versus ai in the long tail. *AEA Papers and Proceedings*, 114:618–622, 2024.

Maxime Bonelli. Data-driven investors. *The Review of Financial Studies*, page hhaf078, 10 2025. ISSN 0893-9454.

Maxime Bonelli and Thierry Foucault. Displaced by big data: Evidence from active fund managers. *Working Paper*, 2023.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322.

Zhimin Chen, Bryan T. Kelly, and Semyon Malamud. Limits to (machine) learning. *Swiss Finance Institute Research Paper Series*, 25-106, 2025.

John H. Cochrane. The risk and return of venture capital. *Journal of Financial Economics*, 75(1):3–52, 2005. ISSN 0304-405X.

Raphaël Couronné, Philipp Probst, and Anne-Laure Boulesteix. Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics*, 19(1):270, 2018.

Douglas J Cumming and Jeffrey G MacIntosh. A cross-country comparison of full and partial venture capital exits. *Journal of Banking & Finance*, 27(3):511–548, 2003. ISSN 0378-4266. Markets and Instututions: Global perspectives.

Diag Davenport. Predictably bad investments: Evidence from venture capitalists. *Chicago Booth Research Paper Series*, 2022. Working paper, Princeton School of Public and International Affairs; revised July 2022.

Victor DeMiguel, Javier Gil-Bazo, Francisco J. Nogales, and André A.P. Santos. Machine learning and fund characteristics help to select mutual funds with positive alpha. *Journal of Financial Economics*, 150(3):103737, 2023. ISSN 0304-405X.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232, 2001.

Thomas Hellmann, Paul Schure, and Dan H. Vo. Angels and venture capitalists: Substitutes or complements? *Journal of Financial Economics*, 141(2):454–478, 2021. ISSN 0304-405X.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2302.05668*, 2023.

David H. Hsu. What do entrepreneurs pay for venture capital affiliation? *The Journal of Finance*, 59(4):1805–1844, 2004. ISSN 00221082, 15406261.

Sheetal Kalbande and Rajvilas Karmore. Startup success prediction using machine learning. In *Proceedings of the Fifth Doctoral Symposium on Computational Intelligence (DoSCI 2024)*, pages 309–319. Springer, 2024.

Ron Kaniel, Zihan Lin, Markus Pelger, and Stijn Van Nieuwerburgh. Machine-learning the skill of mutual fund managers. *Journal of Financial Economics*, 150(1):94–138, 2023. ISSN 0304-405X.

Bryan Kelly, Semyon Malamud, and Kangying Zhou. The virtue of complexity in return prediction. *The Journal of Finance*, 79(1):459–503, 2024.

Barnabás Kiss. Using big data in startup selection: Exploring machine learning as a tool to predict successful startups in the age of social media. *Masters' Thesis, Católica Lisbon School of Business & Economics*, 2022.

Amar Krishna, Ankit Agrawal, and Alok Choudhary. Predicting the outcome of startups: Less failure, more success. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 798–805, 2016. doi: 10.1109/ICDMW.2016.0118.

Victor Lyonnet and Léa H. Stern. Machine learning about venture capital choices. *Fisher College of Business Working Paper Series*, 2022-03-002, 2022. Working paper, revised June 2022.

Ramana Nanda, Sampsa Samila, and Olav Sorenson. The persistent effect of initial success: Evidence from venture capital. *Journal of Financial Economics*, 137(1):231–248, 2020. ISSN 0304-405X.

Christian Peukert, Ananya Sen, and Jörg Claussen. The editor and the algorithm: Recommendation technology in online news. *Management Science*, 70(9):5816–5831, 2024.

Ludovic Phalippou and Oliver Gottschalg. The performance of private equity funds. *The Review of Financial Studies*, 22(4):1747–1776, 2009. ISSN 08939454, 14657368.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31, 2018.

Mona Razaghzadeh Bidgoli, Iman Raeesi Vanani, and Mehdi Goodarzi. Predicting the success of startups using a machine learning approach. *Journal of Innovation and Entrepreneurship*, 13:80, 2024.

Andre Retterath. Human versus computer: Benchmarking venture capitalists and machine learning algorithms for investment screening. SSRN Working Paper, 2020.

Andre Retterath. Data-Driven VC Landscape 2024. *https://datadrivenvc.io/*, 2024.

Greg Ross, Sanjiv Das, Daniel Sciro, and Hussain Raza. Capitalvx: A machine learning model for startup selection and exit prediction. *The Journal of Finance and Data Science*, 7:94–114, 2021. ISSN 2405-9188.

Andrei Shleifer and Robert W. Vishny. The limits of arbitrage. *The Journal of Finance*, 52 (1):35–55, 1997. ISSN 00221082, 15406261.

Ilya Strebulaev and Alex Dang. *The Venture Mindset: How to Make Smarter Bets and Achieve Extraordinary Growth*. Portfolio, New York, NY, May 2024. ISBN 9780593714232.

Pierre-François Verhulst. A note on the law of population growth. 1838.

Appendix A: Dataset Properties and Financing Round Example

Panel A reports descriptive statistics for the financing-rounds dataset used in the analysis, computed after applying the sample restrictions stated in the paper (e.g., currency and geography filters when enabled in the replication code). **Panel B** provides an illustrative financing-round observation (venture, round type, investors, date, and round characteristics) formatted to mirror the raw deal-level fields used to construct the investor-year panels.

Panel A: Sample Data Summary	
Unique Startups	30,373
Unique Financing Rounds	55,858
Funding Volume Total (USD Billions)	823.5
Period Covered	2000–2021
Country	United States
Currency	USD
Mean Investors per Round	3.58
Panel B: Financing Round Data Example	
Venture Name	Stripe
Investment Round	Series C
Investors	Allen & Company, Founders Fund, Khosla Ventures, Sequoia Capital
Date	22.01.2014
Funding Currency	USD
Location	San Francisco, California, United States, North America
Money Raised (USD millions)	80
Pre-Money Valuation (USD millions)	1,670
Sectors or Categories	Finance, FinTech, Mobile Payments, SaaS

Return to Section 3.1

Appendix B.1: Machine Learning Feature Set

This exhibit summarizes the variables used as inputs to the ML models. Features are grouped by economic category. **Panel A** describes startup- and deal-level characteristics observed at the time of the financing round. **Panel B** reports measures of investor experience and historical activity constructed from past investments. **Panel C** describes investor–startup matching variables capturing overlap between startup characteristics and investor histories. **Panel D** summarizes measures of market conditions and competitive environment. **Panel E** describes time controls and the out-of-sample estimation design. All features are constructed using only information available at or before the financing round announcement date.

Panel A: Startup- and Deal-Level Characteristics	
Capital Raised	Total amount of capital raised in the current financing round (USD).
Syndicate Size	Number of distinct investors participating in the current financing round.
Lead Investor Indicators	Indicator for the presence of at least one lead investor in the round and count of lead investors.
Startup Name Characteristics	String-based characteristics of the startup name, including name length, number of words, and number of unique characters.
Startup Categories	Binary indicators for the most frequent startup categories, constructed from company category tags.
Panel B: Investor Experience and Historical Activity	
Cumulative Investor Activity	Aggregated historical investment activity of the investor syndicate prior to the current round, including sums, means, and maxima of past investment counts and past invested capital.
Rolling-Window Investor Activity	Investor historical activity computed over rolling windows of 12, 24, 36, 48, 60, and 120 months prior to the current round. For each window, sums, means, and maxima of investment counts and invested capital are constructed using only past information.
Investor Category Exposure	Historical category exposure of investors based on prior investments, used to construct measures of investor specialization and overlap with startup categories.
Panel C: Investor–Startup Matching Features	
Category Overlap Measures	Weighted Jaccard similarity measures between the startup's category profile and the historical category exposure of the investor syndicate, computed using only pre-investment information.
Prior Investor Network Depth	Number of distinct investors that have invested in the startup prior to the current round, capturing the depth of the startup's existing investor network.
Panel D: Market Conditions and Competitive Environment	
Market Activity (Same Round)	Counts of financing rounds of the same type in the overall market over rolling windows of 12, 24, 36, 48, 60, and 120 months prior to the current round.
State-Level Market Activity	Counts of financing rounds of the same type within the startup's U.S. state over the same rolling windows.
Adjacent-Round Market Activity	Counts of financing rounds in the immediately preceding and subsequent financing stages over rolling windows, capturing congestion and momentum across adjacent stages.
Panel E: Time Controls and Estimation Design	
Calendar Time Controls	Month-of-year indicators and a continuous month index capturing slow-moving secular trends.
Out-of-Sample Construction	All features are constructed using only information available at or before the financing round date. Prediction models are trained using rolling 10-year windows and evaluated out of sample by predicting outcomes in the subsequent year.

Return to Section 3.2

Appendix B.2: Summary Statistics for ML Features and Outcomes

This exhibit reports summary statistics for the ML feature set and success outcomes used in the empirical analysis. The table presents the mean, standard deviation, and selected percentiles (25th, 50th, and 75th) computed on a common sample across all variables. Success outcomes correspond to follow-on financing, acquisition, and IPO indicators and are reported at the top of the table for reference. **Panel A** reports startup- and deal-level characteristics observed at the time of the financing round, including a subset of the most prevalent industry category indicators. **Panel B** reports measures of investor experience and historical activity, focusing on average prior investment activity and invested capital. **Panel C** reports investor–startup matching variables capturing overlap between startup characteristics and investor histories. **Panel D** summarizes measures of market conditions and competitive environment constructed over a rolling window prior to the financing round. **Panel E** reports time controls and indicators related to the out-of-sample estimation design. All summary statistics are computed using only information available at or before the financing round announcement date. For ease of interpretation, values are reported with two significant decimals and scaled units: quantities exceeding one thousand are reported in thousands (K), and quantities exceeding one million are reported in millions (M). Category and month indicators are restricted to the most frequently observed values to conserve space.

	Mean	SD	p25	p50	p75
Targets (Success Outcomes)					
Success: Follow-on	0.80	0.40	1.00	1.00	1.00
Success: Acquisition	0.45	0.50	0.00	0.00	1.00
Success: IPO	0.09	0.28	0.00	0.00	0.00
Panel A: Startup- and Deal-Level Characteristics					
Capital Raised (USD)	21.69 M	46.74 M	5.00 M	10.00 M	21.50 M
Syndicate Size	3.71	2.86	2.00	3.00	5.00
Lead Investor Indicator	0.76	0.43	1.00	1.00	1.00
Number of Lead Investors	0.92	0.67	1.00	1.00	1.00
Startup Name Length	11.13	5.63	7.00	9.00	14.00
Startup Name Words	1.45	0.63	1.00	1.00	2.00
Startup Name Unique Characters	8.47	3.18	6.00	8.00	11.00
Category Dummy: software	0.29	0.45	0.00	0.00	1.00
Category Dummy: health care	0.18	0.38	0.00	0.00	0.00
Category Dummy: information technology	0.11	0.32	0.00	0.00	0.00
Category Dummy: saas	0.11	0.32	0.00	0.00	0.00
Category Dummy: enterprise software	0.13	0.33	0.00	0.00	0.00
Panel B: Investor Experience and Historical Activity					
Investor Prior Investment Count (Mean)	71.95	96.70	10.50	36.50	97.00
Investor Prior Invested Capital (Mean, USD)	1.58 B	2.63 B	128.41 M	616.69 M	1.90 B
Investor Activity: Count Mean (12m)	10.18	11.48	2.25	6.75	14.20
Investor Activity: Capital Mean (12m, USD)	277.47 M	505.13 M	25.45 M	113.33 M	311.52 M
Panel C: Investor–Startup Matching Features					
Category Overlap (Weighted Jaccard, Mean)	0.12	0.20	0.01	0.04	0.14
Category Overlap (Weighted Jaccard, Max)	0.26	0.35	0.02	0.08	0.36
Prior Investor Network Depth	3.46	5.04	0.00	2.00	5.00
Panel D: Market Conditions and Competitive Environment					
Market Activity (Same Round, 12m)	1.42 K	1.10 K	537.00	1.11 K	1.92 K
State Market Activity (Same Round, 12m)	138.18	154.54	15.00	75.00	211.00
Adjacent-Round Activity (Prev, 12m)	3.51 K	3.58 K	696.00	1.59 K	7.08 K
Adjacent-Round Activity (Next, 12m)	685.61	490.61	282.00	579.00	985.00
Panel E: Time Controls and Estimation Design					
Month Index	24.16 K	66.72	24.10 K	24.17 K	24.22 K
Month Dummy: 01	0.10	0.30	0.00	0.00	0.00
Month Dummy: 03	0.09	0.29	0.00	0.00	0.00
Month Dummy: 06	0.09	0.29	0.00	0.00	0.00

Return to Section 3.2

Appendix C: Prediction Algorithms and Machine Learning Infrastructure

This exhibit summarizes the prediction algorithms used in the analysis and documents the timing of major open-source ML framework releases. **Panel A** describes the three classification algorithms employed, spanning linear parametric models and flexible nonlinear tree-based methods. **Panel B** lists major ML libraries released between 2015 and 2017 that expanded access to scalable prediction tools.

Panel A: Prediction Algorithms	
Logit	Linear parametric classification model with a logistic link function. The model assumes a linear relationship between predictors and the log-odds of the outcome and does not capture nonlinearities or interactions unless explicitly specified. Logit serves as a transparent and interpretable benchmark in the analysis. Originally introduced in the context of population growth modeling by Verhulst (1838).
XGBoost	Nonlinear, nonparametric ensemble learning method based on gradient-boosted decision trees. XGBoost constructs additive tree models optimized via gradient boosting and is designed to efficiently handle high-dimensional feature spaces and complex nonlinear interactions. The algorithm builds on the gradient boosting framework of Friedman (2001) and was introduced as a scalable implementation by Chen and Guestrin (2016).
CatBoost	Nonlinear gradient-boosted decision tree algorithm with native support for categorical features. CatBoost incorporates ordered boosting and specialized encoding schemes to reduce overfitting and improve performance when categorical variables are prevalent. The algorithm was introduced by Prokhorenkova et al. (2018) and represents a state-of-the-art tree boosting method for structured data.
Panel B: Machine Learning Infrastructure Shock	
TensorFlow (2015)	Open-source machine learning framework released by Google, providing scalable tools for training and deploying machine learning models. TensorFlow substantially lowered the fixed costs of implementing advanced prediction methods by offering standardized APIs, GPU acceleration, and broad community support.
Microsoft Cognitive Toolkit (2016)	Enterprise-grade machine learning and deep learning framework released by Microsoft, designed to support large-scale model training and deployment. The toolkit expanded access to industrial-strength machine learning infrastructure for applied users.
PyTorch (2016)	Open-source machine learning framework released by Facebook (now Meta), emphasizing flexibility and ease of use. PyTorch enabled rapid experimentation with complex models and became widely adopted by researchers and practitioners due to its dynamic computation graph.
CatBoost (2017)	Release of the CatBoost library by Yandex, providing efficient gradient boosting with native handling of categorical features. CatBoost further reduced the technical barriers to applying advanced tree-based methods in settings with rich categorical data.

Return to Section 4.1

Appendix D: Variable Definitions

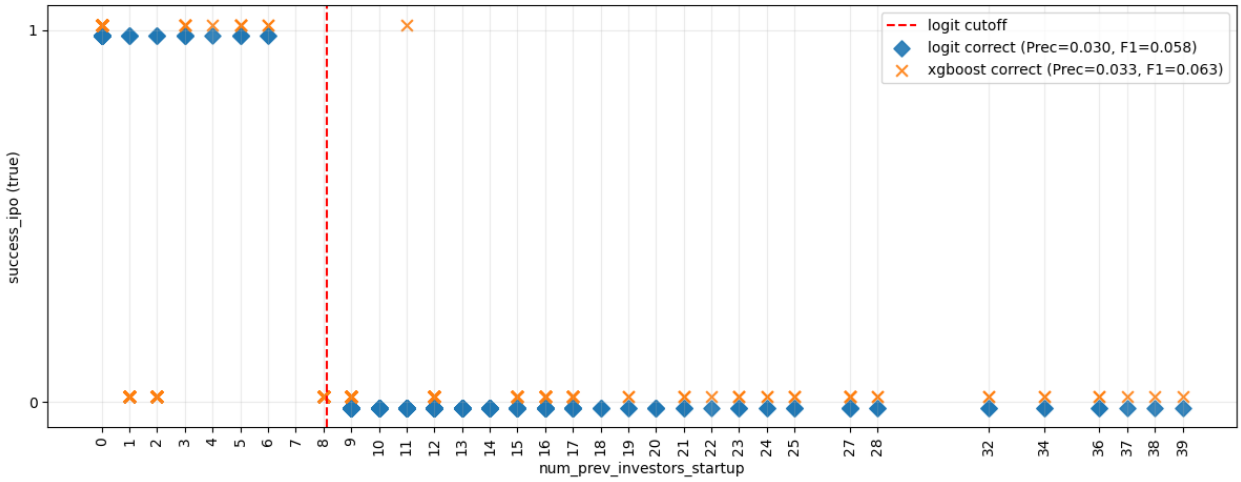
This exhibit summarizes all variables used in the empirical analysis. Panels A–D organize variables by conceptual group. **Panel A** defines investor–year outcome measures and benefit-related variables. **Panel B** defines distance measures capturing differences between realized investor portfolios and model-implied counterfactual portfolios. **Panel C** defines portfolio size variables and sample construction choices. **Panel D** describes the prediction models and out-of-sample estimation procedure. Panel labels in this table are distinct from stage-based panels used elsewhere in the paper, which group results by investment stage (all stages, early, mid, and late).

Panel A: Investor–Year Outcome Measures	
$HitActual_{i,t}$	The realized hit rate of investor i in year t, defined as the fraction of investments made by investor i in year t that result in a given success outcome. Success outcomes are defined separately for follow-on financing, acquisition, and IPO. Hit rates are computed within each investment stage when stage-specific analyses are conducted.
$HitModel_{i,t}$	The counterfactual hit rate implied by a prediction model for investor i in year t. For each investor-year, the model ranks all eligible investments by predicted success probability and selects the top k investments, where k equals the number of investments actually made by investor i in year t. The hit rate is computed as the fraction of these top-k investments that realize the corresponding success outcome.
$Benefit_{i,t}$	The difference between the model-implied hit rate and the realized investor hit rate for investor i in year t: $Benefit_{i,t} = HitModel_{i,t} - HitActual_{i,t}$. A positive value indicates that the model-selected portfolio outperforms the investor's realized portfolio in terms of hit rate.
$HasBenefit_{i,t}$	Indicator variable equal to one if $Benefit_{i,t} > 0$, and zero otherwise.
$Benefit_{i,t-1}$	Lagged benefit, defined as $Benefit_{i,t-1}$ for investor i in the previous calendar year.
$HasBenefit_{i,t-1}$	Lagged indicator for positive benefit, equal to one if $Benefit_{i,t-1} > 0$, and zero otherwise.
Panel B: Distance Measures Between Actual and Model Portfolios	
$DistIndex_{i,t}$	Composite distance index measuring the dissimilarity between investor i's realized portfolio in year t and the corresponding model-implied top-k portfolio. The index is constructed as a standardized average of distances along portfolio hotness, sector composition, and geographic exposure.
$DistHotness_{i,t}$	Difference in average market hotness between the model-implied top-k portfolio and the investor's realized portfolio for investor i in year t. Market hotness is defined as the intensity of recent investment activity in a startup's stage–sector–geography segment, measured by the number of financing rounds in that segment over a rolling 12-month window prior to year t.
$DistSector_{i,t}$	Cosine distance between the sector exposure vector of the investor's realized portfolio and that of the model-implied top-k portfolio in year t. Sector exposure vectors are constructed using startup category indicators.
$DistGeo_{i,t}$	Geographic distance between the investor's realized portfolio and the model-implied top-k portfolio in year t, computed using the distribution of investments across U.S. states.
Panel C: Portfolio Size and Sample Restrictions	
$NInv_{i,t}$	Number of investments made by investor i in year t. This value determines the size k of the model-implied counterfactual portfolio for that investor-year.
$k = \text{act\_p75}$	Counterfactual portfolio size used in benchmark analyses, defined as the number of investments made by an investor at the 75th percentile of the investor activity distribution in a given year.
Panel D: Prediction Models and Estimation Procedure	
Prediction Models	Three prediction models are used: Logit, XGBoost, and CatBoost. Models are trained separately for each success outcome (follow-on financing, acquisition, IPO) and investment stage where applicable.
Out-of-Sample Evaluation	All model-based quantities are computed using rolling 10-year training windows and are evaluated out of sample by predicting outcomes in the subsequent year.

Return to Section 5.1.1

Figure 1: Single-Feature Nonlinearity Diagnostic: Number of Previous Investors and IPO Success (Series A)

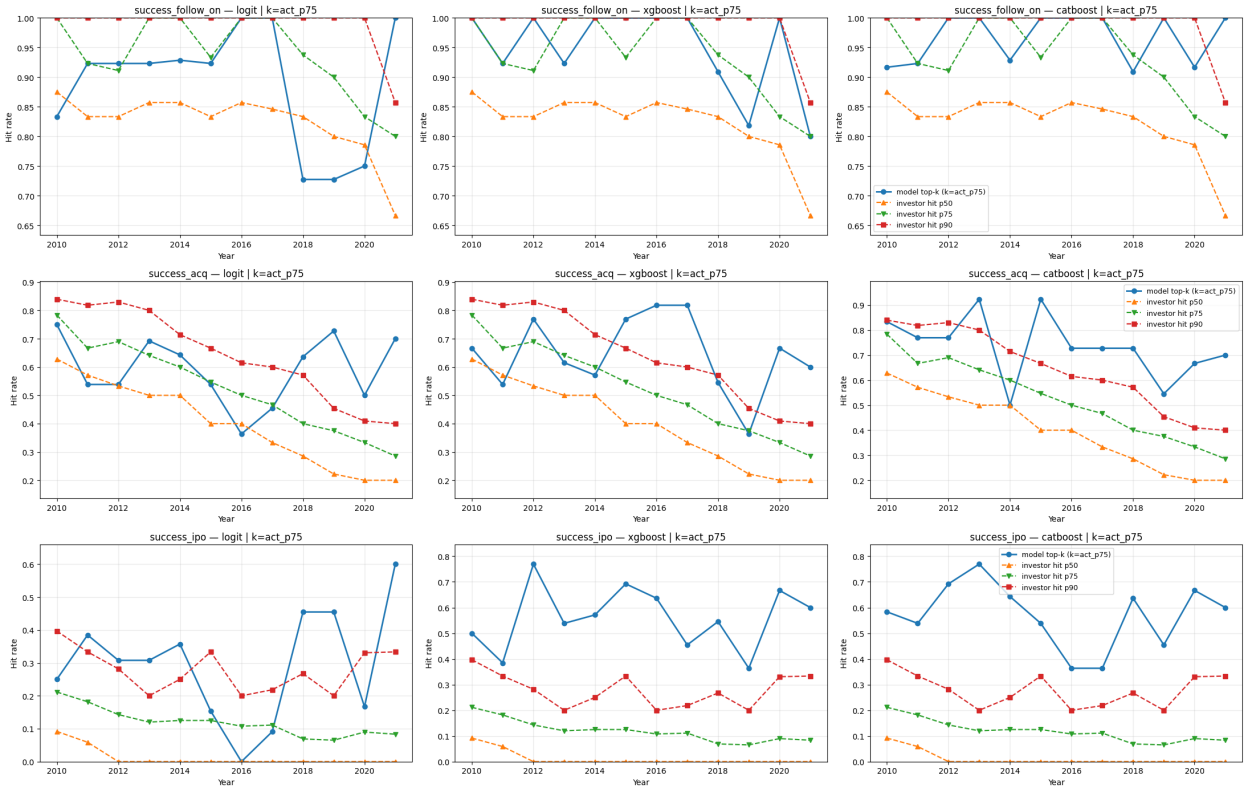
This figure illustrates a single-feature nonlinearity diagnostic for IPO outcomes using the number of previous investors in the startup at the time of a Series A financing round. The horizontal axis reports the number of prior investors, while the vertical axis indicates the realized IPO outcome. Each marker corresponds to an observation in the test sample. Blue diamonds denote observations correctly classified by a logit model, and orange crosses denote observations correctly classified by an XGBoost model. The vertical dashed line indicates the probability cutoff used by the logit model to generate binary predictions. Models are trained on Series A observations dated up to December 31st, 2018 and evaluated on a holdout sample dated from January 1st, 2019 onward. The legend reports model precision and F1 score, where precision measures the share of predicted IPOs that are realized IPOs, recall measures the share of realized IPOs that are correctly predicted, and the F1 score is the harmonic mean of precision and recall. For clarity, only correctly classified observations are shown.



Return to Section 4.2

Figure 2: Model Hit Rates at the 75th Percentile Investment Count Compared to Investor Benchmarks

This figure compares model-implied hit rates to observed investor hit rates when the number of investments is fixed at the 75th percentile of investor activity. Each subplot reports results for a given success outcome (follow-on financing, acquisition, or IPO) and prediction model (Logit, XGBoost, or CatBoost). All model-based quantities are computed using rolling 10-year training windows and are evaluated out-of-sample by predicting outcomes in the subsequent year. The solid blue line shows the hit rate of a counterfactual portfolio constructed by selecting the top-k investments according to the model's predicted success probabilities, where k equals the number of investments made by a 75th-percentile investor. Dashed lines report the realized hit rates of investors at the 50th, 75th, and 90th percentiles of the investor hit-rate distribution. The horizontal axis reports calendar year, and the vertical axis reports hit rates. All investments are pooled across stages.



Return to Section 5.1.2

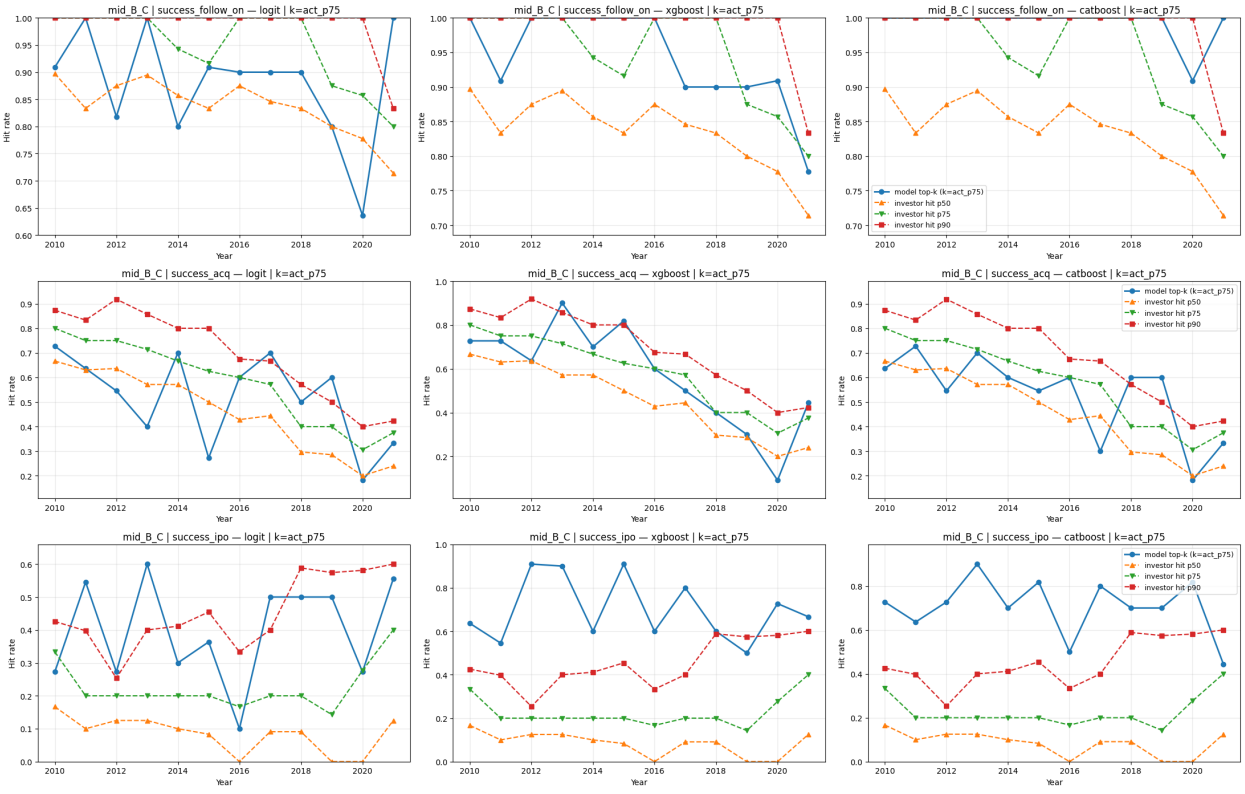
Figure 3: Model Hit Rates at the 75th Percentile Investment Count: Early-Stage

This figure compares model-implied hit rates to observed investor hit rates when the number of investments is fixed at the 75th percentile of investor activity, restricting the sample to early-stage (Seed and Series A) investments. Each subplot reports results for a given success outcome (follow-on financing, acquisition, or IPO) and prediction model (Logit, XGBoost, or CatBoost). All model-based quantities are computed using rolling 10-year training windows and are evaluated out-of-sample by predicting outcomes in the subsequent year. The solid blue line shows the hit rate of a counterfactual portfolio constructed by selecting the top-k investments according to the model's predicted success probabilities, where k equals the number of investments made by a 75th-percentile investor within the corresponding stage. Dashed lines report the realized hit rates of investors at the 50th, 75th, and 90th percentiles of the investor hit-rate distribution. The horizontal axis reports calendar year, and the vertical axis reports hit rates.

Return to Section 5.1.2

Figure 4: Model Hit Rates at the 75th Percentile Investment Count: Mid-Stage

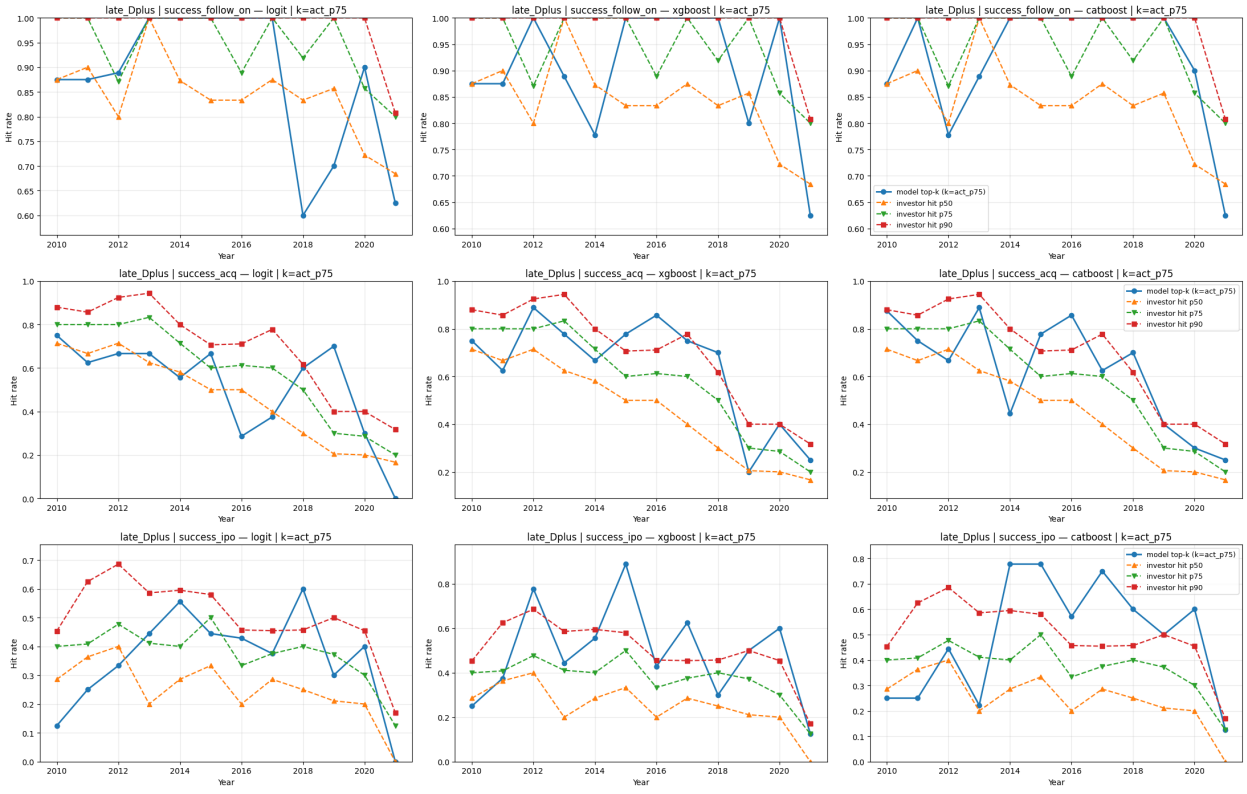
This figure compares model-implied hit rates to observed investor hit rates when the number of investments is fixed at the 75th percentile of investor activity, restricting the sample to mid-stage (Series B and Series C) investments. Each subplot reports results for a given success outcome (follow-on financing, acquisition, or IPO) and prediction model (Logit, XGBoost, or CatBoost). All model-based quantities are computed using rolling 10-year training windows and are evaluated out-of-sample by predicting outcomes in the subsequent year. The solid blue line shows the hit rate of a counterfactual portfolio constructed by selecting the top-k investments according to the model's predicted success probabilities, where k equals the number of investments made by a 75th-percentile investor within the corresponding stage. Dashed lines report the realized hit rates of investors at the 50th, 75th, and 90th percentiles of the investor hit-rate distribution. The horizontal axis reports calendar year, and the vertical axis reports hit rates.



Return to Section 5.1.2

Figure 5: Model Hit Rates at the 75th Percentile Investment Count: Late-Stage

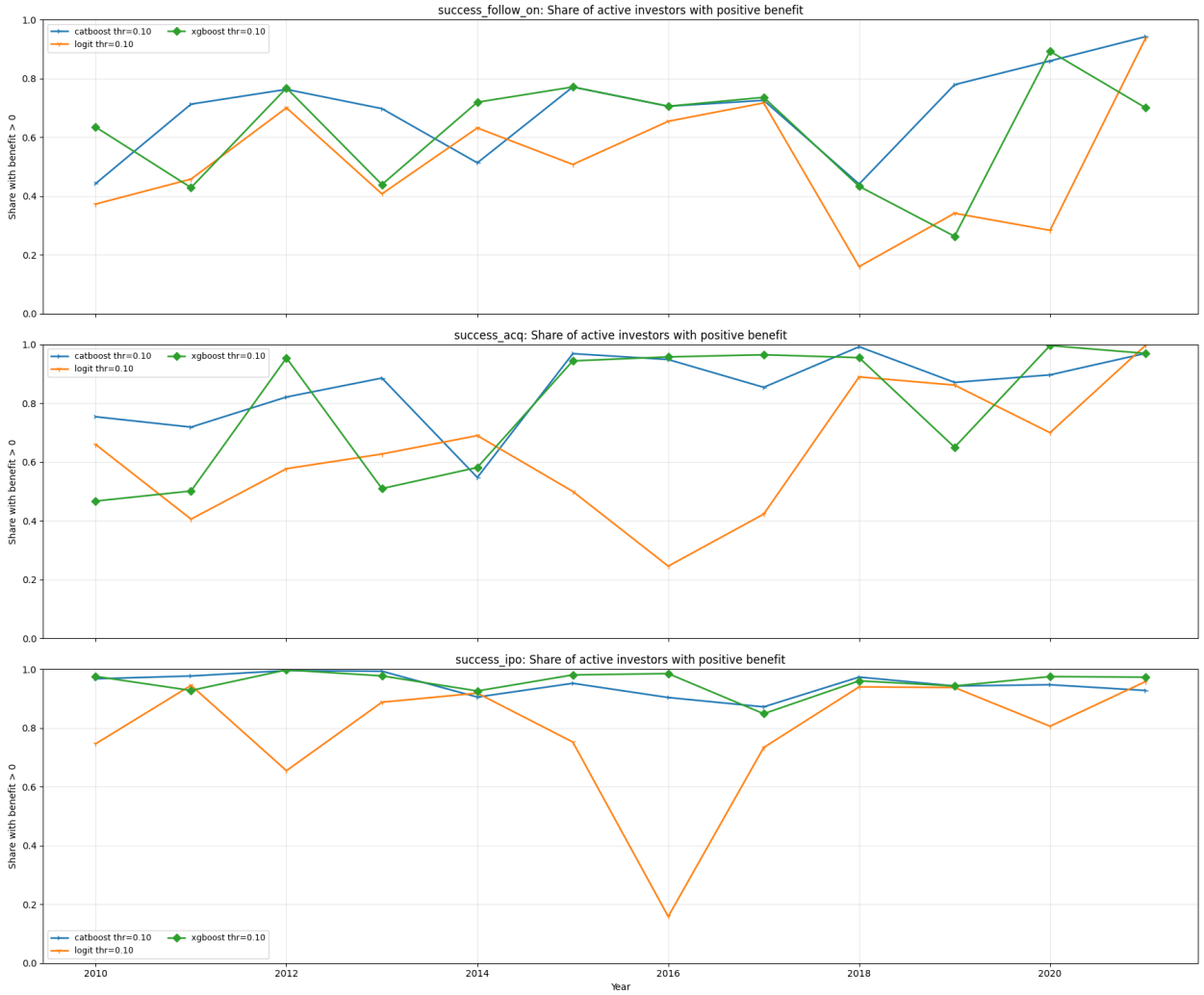
This figure compares model-implied hit rates to observed investor hit rates when the number of investments is fixed at the 75th percentile of investor activity, restricting the sample to late-stage (Series D onward) investments. Each subplot reports results for a given success outcome (follow-on financing, acquisition, or IPO) and prediction model (Logit, XGBoost, or CatBoost). All model-based quantities are computed using rolling 10-year training windows and are evaluated out-of-sample by predicting outcomes in the subsequent year. The solid blue line shows the hit rate of a counterfactual portfolio constructed by selecting the top-k investments according to the model's predicted success probabilities, where k equals the number of investments made by a 75th-percentile investor within the corresponding stage. Dashed lines report the realized hit rates of investors at the 50th, 75th, and 90th percentiles of the investor hit-rate distribution. The horizontal axis reports calendar year, and the vertical axis reports hit rates.



Return to Section 5.1.2

Figure 6: Share of Active Investors with Positive Benefit Over Time

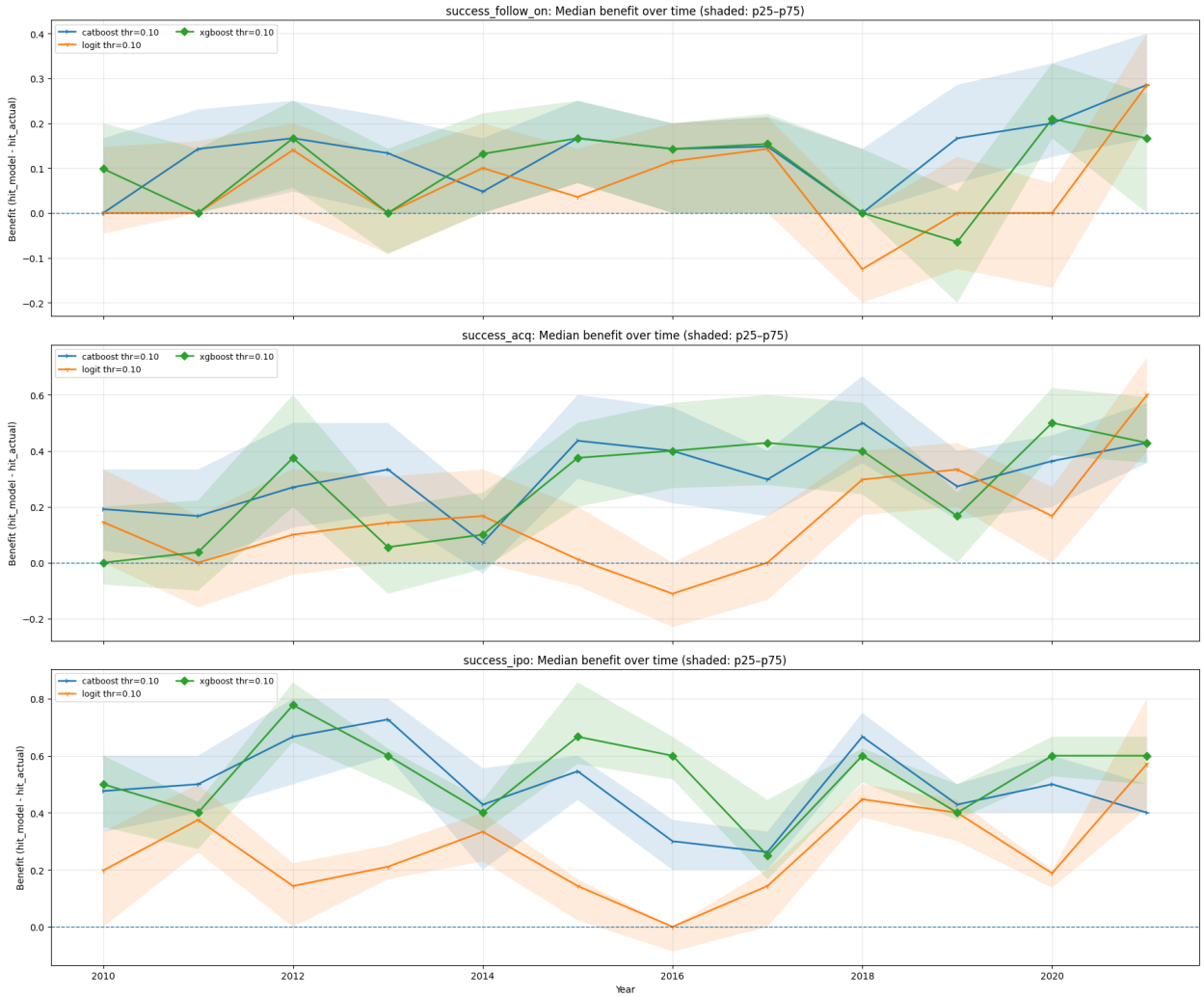
This figure reports, for each calendar year, the share of active investors whose estimated benefit is positive. Separate panels correspond to different success outcomes (follow-on financing, acquisition, and IPO). Within each panel, lines report results for different prediction models (Logit, XGBoost, and CatBoost). All model-based quantities are computed using rolling 10-year training windows and are evaluated out-of-sample by predicting outcomes in the subsequent year. Benefit is defined as the difference between the model-implied hit rate and the investor's realized hit rate, computed using a probability threshold of 0.10. The horizontal axis reports calendar year, and the vertical axis reports the fraction of active investors with positive benefit in a given year.



Return to Section 5.2

Figure 7: Median Benefit Over Time with Interquartile Range

This figure reports the evolution over time of the median estimated benefit, defined as the difference between the model-implied hit rate and the investor's realized hit rate. All model-based quantities are computed using rolling 10-year training windows and are evaluated out-of-sample by predicting outcomes in the subsequent year. Shaded areas represent the interquartile range (25th to 75th percentiles) of the benefit distribution. Separate panels correspond to different success outcomes (follow-on financing, acquisition, and IPO), and lines correspond to different prediction models (Logit, XGBoost, and CatBoost). All results are computed using a probability threshold of 0.10. The horizontal axis reports calendar year, and the vertical axis reports the benefit measure.



Return to Section 5.2

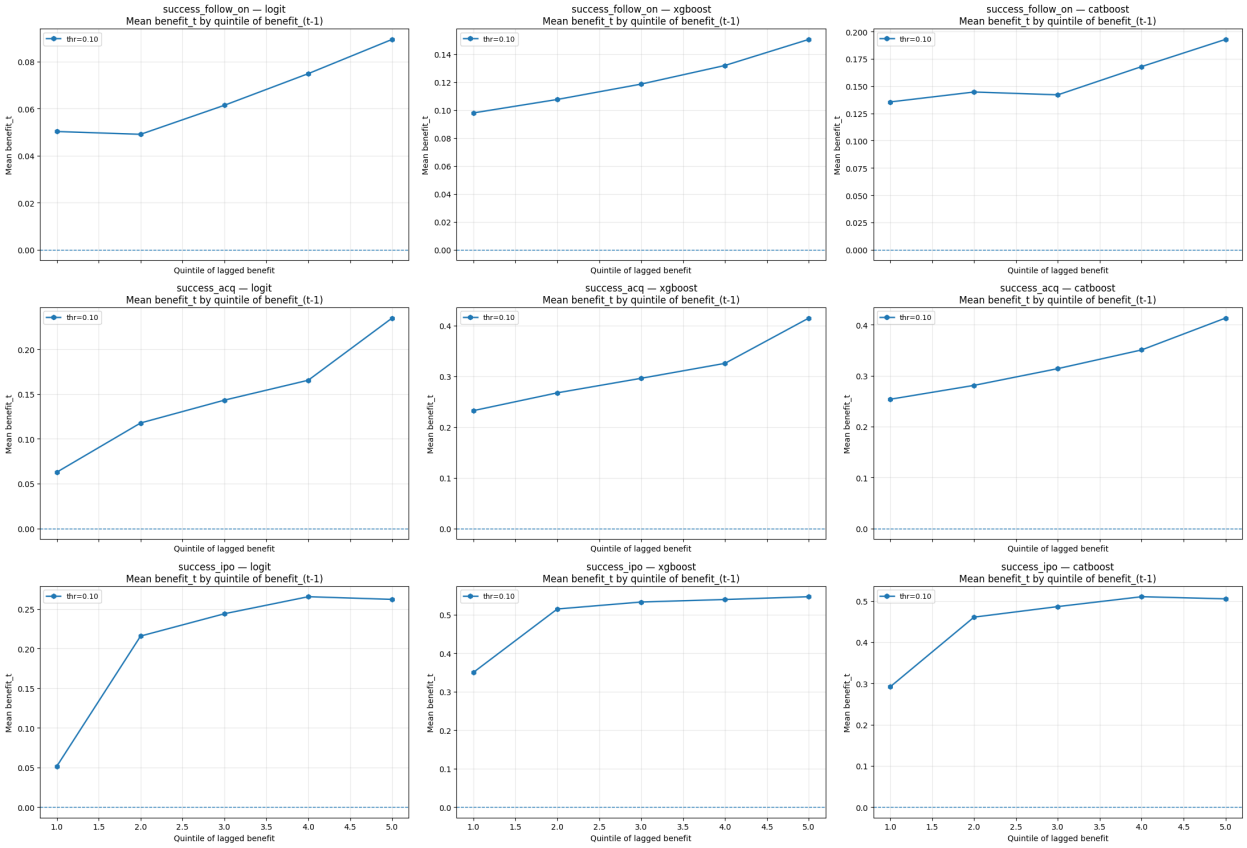
Figure 8: Sign Persistence of Benefit Over Time

This figure illustrates the persistence over time of the sign of the estimated benefit. For each calendar year, the figure reports three quantities: the unconditional probability that benefit is positive; the probability that benefit is positive conditional on being positive in the previous year; and the probability that benefit is positive conditional on being non-positive in the previous year. Separate panels correspond to different success outcomes (follow-on financing, acquisition, and IPO). All model-based quantities are computed using rolling 10-year training windows and are evaluated out-of-sample by predicting outcomes in the subsequent year. Benefit is computed using a probability threshold of 0.10. The horizontal axis reports calendar year, and the vertical axis reports probabilities.

Return to Section 5.2.3

Figure 9: Mean Benefit by Quintile of Lagged Benefit

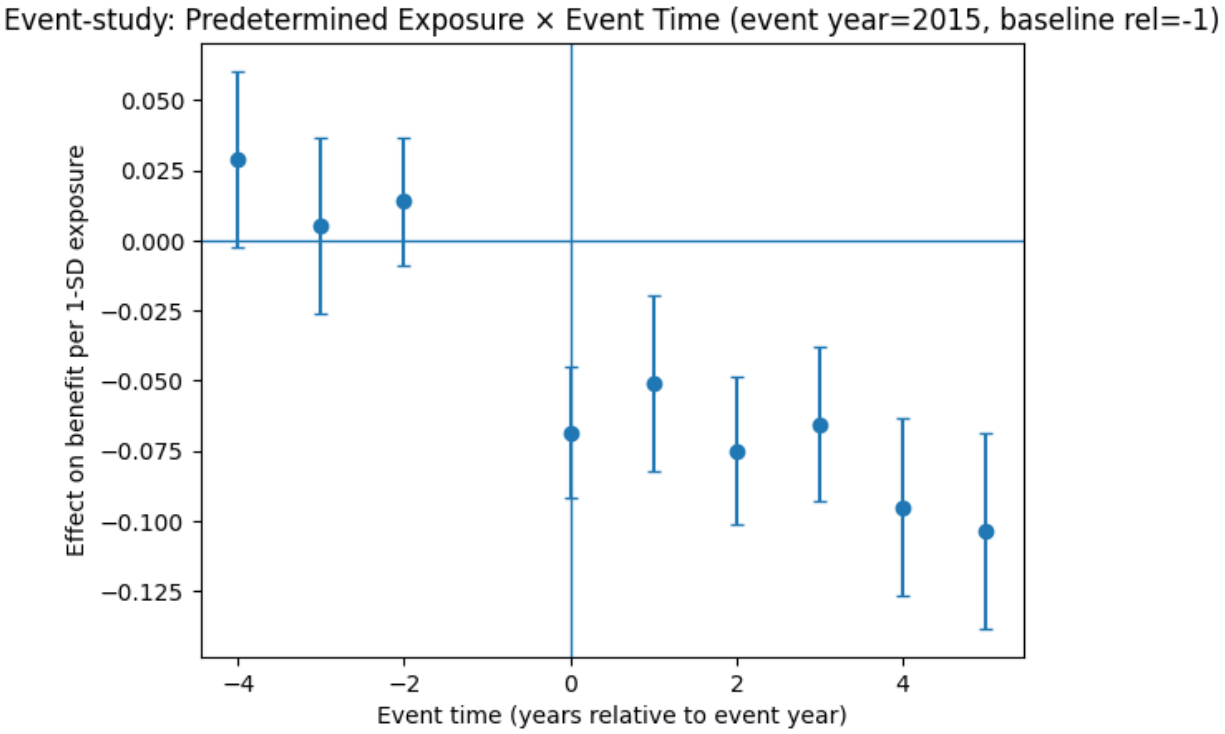
This figure reports the mean estimated benefit in period t by quintiles of lagged benefit in period $t-1$. Separate panels correspond to different success outcomes (follow-on financing, acquisition, and IPO) and different prediction models (Logit, XGBoost, and CatBoost). All model-based quantities are computed using rolling 10-year training windows and are evaluated out-of-sample by predicting outcomes in the subsequent year. The horizontal axis reports quintiles of the lagged benefit distribution, and the vertical axis reports the mean contemporaneous benefit. All results are computed using a probability threshold of 0.10.



Return to Section 5.2.3

Figure 10: Event-Study of Benefit Dynamics Around a Technology Shock

This figure reports event-study estimates of the effect of a predetermined exposure measure on the evolution of the estimated benefit around a common event year. The horizontal axis reports event time in years relative to the event year, normalized such that the year immediately preceding the event serves as the baseline. The vertical axis reports the estimated effect on benefit per one-standard-deviation increase in exposure. Points denote coefficient estimates and vertical bars denote confidence intervals based on clustered standard errors. Exposure is defined using pre-event information only. The underlying regression includes investor and year fixed effects.



Return to Section 5.3.2

Table 1: Summary Statistics

This table reports summary statistics for the main variables used in the paper, computed on the main estimation panel. Statistics are reported separately by investment stage, success outcome, and prediction model. See Variable Definitions (Appendix D) for detailed descriptions of all variables.

	Logit					XGBoost					CatBoost					N
	Mean	SD	p25	p50	p75	Mean	SD	p25	p50	p75	Mean	SD	p25	p50	p75	
Panel A: All Stages																
Follow-on																
Benefit_{it}	0.065	0.182	0.000	0.048	0.167	0.121	0.172	0.000	0.125	0.200	0.157	0.161	0.000	0.143	0.235	4,937
$\text{HitRate}_{it}^{\text{Actual}}$	0.810	0.151	0.727	0.833	0.909	0.810	0.151	0.727	0.833	0.909	0.810	0.151	0.727	0.833	0.909	4,937
$\text{HitRate}_{it}^{\text{Model}}$	0.875	0.115	0.800	0.900	1.000	0.931	0.098	0.875	1.000	1.000	0.966	0.054	0.938	1.000	1.000	4,937
$\text{Number of Investments}_{it}$	13.105	13.465	6.000	9.000	14.000	13.105	13.465	6.000	9.000	14.000	13.105	13.465	6.000	9.000	14.000	4,937
Distance_{it}	0.017	0.656	-0.442	-0.057	0.438	0.018	0.656	-0.441	-0.055	0.439	0.018	0.656	-0.442	-0.055	0.439	4,918
$\text{Distance}_{it}^{\text{Hotness}}$	-2.621	2.509	-3.736	-2.278	-1.122	-2.639	2.509	-3.757	-2.305	-1.150	-2.640	2.509	-3.757	-2.306	-1.151	4,918
$\text{Distance}_{it}^{\text{Sector}}$	0.354	0.160	0.234	0.325	0.458	0.354	0.161	0.234	0.325	0.458	0.354	0.161	0.234	0.325	0.458	4,918
$\text{Distance}_{it}^{\text{Geography}}$	0.515	0.171	0.399	0.470	0.567	0.515	0.171	0.398	0.470	0.566	0.515	0.171	0.398	0.470	0.566	4,918
Acquisition																
Benefit_{it}	0.145	0.258	0.000	0.143	0.324	0.307	0.254	0.143	0.308	0.500	0.322	0.230	0.167	0.333	0.500	4,937
$\text{HitRate}_{it}^{\text{Actual}}$	0.364	0.214	0.200	0.333	0.500	0.364	0.214	0.200	0.333	0.500	0.364	0.214	0.200	0.333	0.500	4,937
$\text{HitRate}_{it}^{\text{Model}}$	0.508	0.158	0.400	0.520	0.625	0.671	0.159	0.583	0.692	0.800	0.686	0.148	0.571	0.667	0.812	4,937
$\text{Number of Investments}_{it}$	13.105	13.465	6.000	9.000	14.000	13.105	13.465	6.000	9.000	14.000	13.105	13.465	6.000	9.000	14.000	4,937
Distance_{it}	0.016	0.657	-0.449	-0.057	0.434	0.018	0.656	-0.445	-0.057	0.438	0.018	0.657	-0.445	-0.056	0.437	4,918
$\text{Distance}_{it}^{\text{Hotness}}$	-2.603	2.509	-3.720	-2.265	-1.106	-2.578	2.508	-3.696	-2.249	-1.083	-2.593	2.509	-3.707	-2.267	-1.095	4,918
$\text{Distance}_{it}^{\text{Sector}}$	0.354	0.167	0.228	0.322	0.460	0.354	0.164	0.230	0.323	0.459	0.354	0.162	0.232	0.324	0.458	4,918
$\text{Distance}_{it}^{\text{Geography}}$	0.514	0.171	0.397	0.468	0.571	0.512	0.172	0.395	0.468	0.567	0.513	0.172	0.393	0.467	0.564	4,918
IPO																
Benefit_{it}	0.208	0.233	0.083	0.200	0.375	0.497	0.215	0.400	0.538	0.625	0.450	0.216	0.333	0.476	0.600	4,937
$\text{HitRate}_{it}^{\text{Actual}}$	0.084	0.152	0.000	0.000	0.111	0.084	0.152	0.000	0.000	0.111	0.084	0.152	0.000	0.000	0.111	4,937
$\text{HitRate}_{it}^{\text{Model}}$	0.292	0.175	0.167	0.276	0.429	0.581	0.154	0.500	0.600	0.667	0.534	0.157	0.400	0.550	0.647	4,937
$\text{Number of Investments}_{it}$	13.105	13.465	6.000	9.000	14.000	13.105	13.465	6.000	9.000	14.000	13.105	13.465	6.000	9.000	14.000	4,937
Distance_{it}	0.013	0.644	-0.427	-0.055	0.427	-0.001	0.642	-0.436	-0.046	0.408	0.010	0.648	-0.433	-0.040	0.423	4,918
$\text{Distance}_{it}^{\text{Hotness}}$	-2.119	2.509	-3.208	-1.777	-0.628	-1.769	2.514	-2.878	-1.471	-0.295	-2.028	2.505	-3.127	-1.699	-0.552	4,918
$\text{Distance}_{it}^{\text{Sector}}$	0.403	0.142	0.308	0.387	0.484	0.454	0.174	0.327	0.450	0.566	0.404	0.151	0.299	0.388	0.494	4,918
$\text{Distance}_{it}^{\text{Geography}}$	0.489	0.186	0.359	0.437	0.558	0.489	0.185	0.357	0.443	0.570	0.495	0.179	0.370	0.448	0.560	4,918
Panel B: Early Stage (Seed–A)																
Follow-on																
Benefit_{it}	0.086	0.176	0.000	0.071	0.192	0.156	0.163	0.000	0.143	0.247	0.131	0.167	0.000	0.125	0.211	2,835
$\text{HitRate}_{it}^{\text{Actual}}$	0.811	0.155	0.727	0.833	0.917	0.811	0.155	0.727	0.833	0.917	0.811	0.155	0.727	0.833	0.917	2,835
$\text{HitRate}_{it}^{\text{Model}}$	0.896	0.075	0.833	0.886	1.000	0.967	0.059	0.944	1.000	1.000	0.942	0.075	0.900	1.000	1.000	2,835
$\text{Number of Investments}_{it}$	11.975	12.657	6.000	8.000	13.000	11.975	12.657	6.000	8.000	13.000	11.975	12.657	6.000	8.000	13.000	2,835
Distance_{it}	-0.005	0.712	-0.499	-0.093	0.423	-0.005	0.712	-0.500	-0.097	0.426	-0.005	0.713	-0.501	-0.096	0.426	2,825
$\text{Distance}_{it}^{\text{Hotness}}$	-2.905	2.944	-4.279	-2.509	-1.039	-2.928	2.945	-4.316	-2.538	-1.054	-2.929	2.945	-4.320	-2.538	-1.056	2,825
$\text{Distance}_{it}^{\text{Sector}}$	0.338	0.155	0.227	0.309	0.427	0.338	0.155	0.226	0.308	0.426	0.338	0.155	0.226	0.308	0.426	2,825
$\text{Distance}_{it}^{\text{Geography}}$	0.539	0.176	0.415	0.485	0.597	0.538	0.176	0.415	0.486	0.597	0.538	0.176	0.415	0.486	0.597	2,825
Acquisition																
Benefit_{it}	0.246	0.275	0.000	0.231	0.429	0.267	0.251	0.111	0.259	0.429	0.356	0.274	0.167	0.357	0.571	2,835
$\text{HitRate}_{it}^{\text{Actual}}$	0.348	0.208	0.200	0.333	0.500	0.348	0.208	0.200	0.333	0.500	0.348	0.208	0.200	0.333	0.500	2,835
$\text{HitRate}_{it}^{\text{Model}}$	0.594	0.139	0.500	0.571	0.667	0.615	0.144	0.500	0.600	0.714	0.704	0.181	0.625	0.714	0.833	2,835
$\text{Number of Investments}_{it}$	11.975	12.657	6.000	8.000	13.000	11.975	12.657	6.000	8.000	13.000	11.975	12.657	6.000	8.000	13.000	2,835
Distance_{it}	-0.005	0.714	-0.501	-0.092	0.426	-0.006	0.714	-0.499	-0.096	0.423	-0.005	0.714	-0.499	-0.099	0.424	2,825
$\text{Distance}_{it}^{\text{Hotness}}$	-2.848	2.943	-4.234	-2.450	-0.976	-2.854	2.942	-4.233	-2.455	-0.961	-2.871	2.944	-4.243	-2.471	-0.984	2,825
$\text{Distance}_{it}^{\text{Sector}}$	0.337	0.158	0.222	0.305	0.428	0.337	0.157	0.223	0.306	0.427	0.337	0.156	0.224	0.307	0.427	2,825
$\text{Distance}_{it}^{\text{Geography}}$	0.536	0.177	0.410	0.484	0.596	0.534	0.178	0.407	0.481	0.596	0.535	0.178	0.409	0.484	0.595	2,825
IPO																
Benefit_{it}	0.139	0.168	0.000	0.125	0.222	0.403	0.194	0.286	0.400	0.516	0.383	0.147	0.294	0.375	0.500	2,835
$\text{HitRate}_{it}^{\text{Actual}}$	0.024	0.072	0.000	0.000	0.000	0.024	0.072	0.000	0.000	0.000	0.024	0.072	0.000	0.000	0.000	2,835
$\text{HitRate}_{it}^{\text{Model}}$	0.163	0.150	0.000	0.150	0.250	0.427	0.184	0.308	0.400	0.545	0.407	0.132	0.333	0.400	0.500	2,835
$\text{Number of Investments}_{it}$	11.975	12.657	6.000	8.000	13.000	11.975	12.657	6.000	8.000	13.000	11.975	12.657	6.000	8.000	13.000	2,835
Distance_{it}	0.005	0.696	-0.486	-0.068	0.457	0.051	0.660	-0.402	-0.012	0.472	0.025	0.686	-0.469	-0.037	0.467	2,825
$\text{Distance}_{it}^{\text{Hotness}}$	-2.431	2.940	-3.832	-2.038	-0.573	-2.370	2.930	-3.763	-2.048	-0.575	-2.493	2.931	-3.860	-2.158	-0.673	2,825
$\text{Distance}_{it}^{\text{Sector}}$	0.401	0.137	0.304	0.387	0.483	0.496	0.176	0.367	0.481	0.608	0.415	0.150	0.309	0.396	0.508	2,825
$\text{Distance}_{it}^{\text{Geography}}$	0.507	0.198	0.359	0.443	0.592	0.512	0.195	0.360	0.459	0.605	0.520	0.185	0.386	0.472	0.607	2,825

Return to Section 5.2.2

Summary Statistics (continued)

This table continues to report summary statistics for the main variables used in the paper, computed on the main estimation panel and reported separately by investment stage, success outcome, and prediction model. See Variable Definitions (Appendix D) for detailed descriptions of all variables.

	Logit					XGBoost					CatBoost					N
	Mean	SD	p25	p50	p75	Mean	SD	p25	p50	p75	Mean	SD	p25	p50	p75	
Panel C: Mid Stage (B–C)																
Follow-on																
Benefit_{it}	0.045	0.184	-0.058	0.000	0.167	0.121	0.149	0.000	0.113	0.200	0.168	0.146	0.047	0.154	0.250	1,074
$\text{HitRate}_{it}^{\text{Actual}}$	0.817	0.144	0.728	0.833	0.909	0.817	0.144	0.728	0.833	0.909	0.817	0.144	0.728	0.833	0.909	1,074
$\text{HitRate}_{it}^{\text{Model}}$	0.862	0.135	0.818	0.875	1.000	0.938	0.077	0.875	1.000	1.000	0.985	0.040	1.000	1.000	1.000	1,074
$\text{Number of Investments}_{it}$	9.935	5.887	6.000	8.000	12.000	9.935	5.887	6.000	8.000	12.000	9.935	5.887	6.000	8.000	12.000	1,074
Distance_{it}	0.000	0.474	-0.326	-0.044	0.295	-0.002	0.477	-0.329	-0.045	0.299	-0.002	0.477	-0.329	-0.043	0.299	1,067
$\text{Distance}_{it}^{\text{Hotness}}$	-1.770	1.778	-2.564	-1.467	-0.609	-1.792	1.770	-2.568	-1.519	-0.631	-1.792	1.770	-2.563	-1.519	-0.631	1,067
$\text{Distance}_{it}^{\text{Sector}}$	0.351	0.128	0.258	0.338	0.428	0.349	0.128	0.257	0.335	0.427	0.349	0.128	0.256	0.335	0.428	1,067
$\text{Distance}_{it}^{\text{Geography}}$	0.453	0.134	0.365	0.427	0.512	0.454	0.133	0.366	0.427	0.515	0.454	0.133	0.366	0.427	0.515	1,067
Acquisition																
Benefit_{it}	0.053	0.281	-0.125	0.000	0.222	0.084	0.226	-0.071	0.089	0.222	0.054	0.238	-0.125	0.000	0.200	1,074
$\text{HitRate}_{it}^{\text{Actual}}$	0.426	0.230	0.250	0.400	0.600	0.426	0.230	0.250	0.400	0.600	0.426	0.230	0.250	0.400	0.600	1,074
$\text{HitRate}_{it}^{\text{Model}}$	0.479	0.201	0.300	0.538	0.625	0.510	0.235	0.333	0.500	0.714	0.480	0.189	0.333	0.536	0.600	1,074
$\text{Number of Investments}_{it}$	9.935	5.887	6.000	8.000	12.000	9.935	5.887	6.000	8.000	12.000	9.935	5.887	6.000	8.000	12.000	1,074
Distance_{it}	-0.002	0.470	-0.333	-0.049	0.296	0.004	0.476	-0.326	-0.039	0.297	-0.001	0.475	-0.331	-0.050	0.295	1,067
$\text{Distance}_{it}^{\text{Hotness}}$	-1.777	1.760	-2.574	-1.497	-0.632	-1.740	1.749	-2.576	-1.472	-0.618	-1.778	1.763	-2.558	-1.513	-0.619	1,067
$\text{Distance}_{it}^{\text{Sector}}$	0.350	0.135	0.250	0.332	0.441	0.348	0.131	0.252	0.330	0.429	0.349	0.129	0.256	0.333	0.428	1,067
$\text{Distance}_{it}^{\text{Geography}}$	0.454	0.133	0.366	0.430	0.516	0.456	0.133	0.366	0.431	0.515	0.454	0.133	0.366	0.430	0.516	1,067
IPO																
Benefit_{it}	0.245	0.276	0.013	0.297	0.453	0.544	0.246	0.402	0.571	0.714	0.530	0.254	0.400	0.583	0.700	1,074
$\text{HitRate}_{it}^{\text{Actual}}$	0.144	0.198	0.000	0.077	0.200	0.144	0.198	0.000	0.077	0.200	0.144	0.198	0.000	0.077	0.200	1,074
$\text{HitRate}_{it}^{\text{Model}}$	0.389	0.194	0.300	0.429	0.538	0.688	0.158	0.571	0.667	0.818	0.674	0.157	0.584	0.684	0.800	1,074
$\text{Number of Investments}_{it}$	9.935	5.887	6.000	8.000	12.000	9.935	5.887	6.000	8.000	12.000	9.935	5.887	6.000	8.000	12.000	1,074
Distance_{it}	-0.013	0.563	-0.359	-0.014	0.355	-0.017	0.609	-0.409	0.023	0.410	-0.016	0.563	-0.379	-0.020	0.348	1,067
$\text{Distance}_{it}^{\text{Hotness}}$	-1.495	1.740	-2.295	-1.200	-0.361	-1.060	1.718	-1.938	-0.814	0.026	-1.318	1.724	-2.140	-1.036	-0.192	1,067
$\text{Distance}_{it}^{\text{Sector}}$	0.391	0.150	0.289	0.386	0.491	0.436	0.195	0.287	0.452	0.576	0.388	0.152	0.282	0.383	0.487	1,067
$\text{Distance}_{it}^{\text{Geography}}$	0.443	0.138	0.356	0.413	0.504	0.443	0.138	0.350	0.417	0.512	0.446	0.134	0.363	0.424	0.507	1,067
Panel D: Late Stage (D+)																
Follow-on																
Benefit_{it}	0.058	0.191	0.000	0.000	0.200	0.076	0.202	0.000	0.067	0.200	0.108	0.189	0.000	0.111	0.200	327
$\text{HitRate}_{it}^{\text{Actual}}$	0.820	0.155	0.714	0.833	1.000	0.820	0.155	0.714	0.833	1.000	0.820	0.155	0.714	0.833	1.000	327
$\text{HitRate}_{it}^{\text{Model}}$	0.878	0.136	0.800	0.900	1.000	0.896	0.129	0.821	0.938	1.000	0.927	0.137	0.900	1.000	1.000	327
$\text{Number of Investments}_{it}$	8.385	3.648	6.000	7.000	10.000	8.385	3.648	6.000	7.000	10.000	8.385	3.648	6.000	7.000	10.000	327
Distance_{it}	-0.224	0.404	-0.525	-0.238	-0.017	-0.225	0.406	-0.531	-0.254	-0.008	-0.225	0.406	-0.532	-0.250	-0.008	325
$\text{Distance}_{it}^{\text{Hotness}}$	-1.809	1.303	-2.489	-1.675	-0.989	-1.876	1.304	-2.541	-1.723	-1.071	-1.874	1.304	-2.532	-1.723	-1.071	325
$\text{Distance}_{it}^{\text{Sector}}$	0.316	0.115	0.232	0.293	0.381	0.314	0.117	0.230	0.293	0.381	0.314	0.117	0.230	0.293	0.380	325
$\text{Distance}_{it}^{\text{Geography}}$	0.379	0.109	0.290	0.375	0.441	0.384	0.109	0.297	0.379	0.449	0.384	0.109	0.298	0.379	0.449	325
Acquisition																
Benefit_{it}	0.087	0.259	-0.103	0.083	0.286	0.205	0.232	0.000	0.200	0.375	0.171	0.250	0.000	0.167	0.333	327
$\text{HitRate}_{it}^{\text{Actual}}$	0.416	0.245	0.205	0.400	0.600	0.416	0.245	0.205	0.400	0.600	0.416	0.245	0.205	0.400	0.600	327
$\text{HitRate}_{it}^{\text{Model}}$	0.503	0.253	0.300	0.571	0.700	0.621	0.221	0.400	0.667	0.800	0.587	0.251	0.375	0.600	0.800	327
$\text{Number of Investments}_{it}$	8.385	3.648	6.000	7.000	10.000	8.385	3.648	6.000	7.000	10.000	8.385	3.648	6.000	7.000	10.000	327
Distance_{it}	-0.204	0.410	-0.518	-0.238	0.046	-0.231	0.407	-0.532	-0.265	-0.024	-0.228	0.406	-0.532	-0.263	-0.010	325
$\text{Distance}_{it}^{\text{Hotness}}$	-2.015	1.344	-2.811	-1.816	-1.127	-1.886	1.298	-2.532	-1.723	-1.072	-1.874	1.303	-2.540	-1.717	-1.064	325
$\text{Distance}_{it}^{\text{Sector}}$	0.308	0.129	0.208	0.281	0.376	0.308	0.121	0.220	0.281	0.374	0.312	0.118	0.227	0.289	0.378	325
$\text{Distance}_{it}^{\text{Geography}}$	0.412	0.104	0.336	0.411	0.478	0.387	0.109	0.299	0.383	0.453	0.384	0.109	0.297	0.381	0.449	325
IPO																
Benefit_{it}	0.117	0.218	0.000	0.143	0.250	0.273	0.234	0.125	0.286	0.429	0.291	0.247	0.156	0.333	0.444	327
$\text{HitRate}_{it}^{\text{Actual}}$	0.248	0.188	0.118	0.222	0.394	0.248	0.188	0.118	0.222	0.394	0.248	0.188	0.118	0.222	0.394	327
$\text{HitRate}_{it}^{\text{Model}}$	0.364	0.144	0.333	0.385	0.429	0.521	0.192	0.400	0.500	0.625	0.539	0.196	0.429	0.583	0.667	327
$\text{Number of Investments}_{it}$	8.385	3.648	6.000	7.000	10.000	8.385	3.648	6.000	7.000	10.000	8.385	3.648	6.000	7.000	10.000	327
Distance_{it}	-0.205	0.421	-0.492	-0.233	0.045	-0.364	0.422	-0.673	-0.376	-0.075	-0.321	0.431	-0.628	-0.340	-0.046	325
$\text{Distance}_{it}^{\text{Hotness}}$	-1.627	1.317	-2.310	-1.482	-0.788	-1.536	1.304	-2.196	-1.363	-0.739	-1.638	1.305	-2.301	-1.489	-0.823	325
$\text{Distance}_{it}^{\text{Sector}}$	0.376	0.118	0.290	0.361	0.453	0.360	0.125	0.266	0.337	0.438	0.336	0.119	0.245	0.314	0.407	325
$\text{Distance}_{it}^{\text{Geography}}$	0.370	0.108	0.291	0.368	0.426	0.387	0.117	0.296	0.376	0.465	0.380	0.111	0.296	0.366	0.448	325

Return to Section 5.2.2

Table 2: Persistence of the Presence of Predictive Benefit

This table reports panel regression results analyzing the persistence of investors' predictive benefit over time. The dependent variable is $HasBenefit_{it}$, an indicator equal to one if investor i exhibits a positive predictive benefit in year t, and zero otherwise. The main explanatory variable is the lagged indicator $HasBenefit_{i,t-1}$. Results are reported separately by success outcome—follow-on financing, acquisition, and IPO—and by prediction model (Logit, XGBoost, and CatBoost). **Panel A** reports results pooling all investment stages, while **Panels B–D** report results separately for early-stage (Seed–A), mid-stage (B–C), and late-stage (D+) investments. All specifications include investor and year fixed effects, and standard errors are clustered at the investor level. The precise definition of all variables is provided in Appendix D. Clustered standard errors are reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

	Panel A: All Stages			**Panel B: Early Stage (Seed–A)**		
	Logit	XGBoost	CatBoost	Logit	XGBoost	CatBoost
Follow-on						
$\text{HasBenefit}_{i,t-1}$	-0.135***	-0.112***	-0.151***	-0.112***	-0.136***	-0.137***
	(0.015)	(0.015)	(0.016)	(0.020)	(0.022)	(0.023)
	[0.041]	[0.033]	[0.047]	[0.036]	[0.043]	[0.042]
Acquisition						
$\text{HasBenefit}_{i,t-1}$	-0.036**	-0.033**	-0.083***	0.029	-0.098***	-0.141***
	(0.017)	(0.015)	(0.017)	(0.021)	(0.020)	(0.020)
	[0.006]	[0.005]	[0.014]	[0.002]	[0.018]	[0.026]
IPO						
$\text{HasBenefit}_{i,t-1}$	-0.102***	-0.144***	-0.085**	0.063***	-0.207***	-0.158**
	(0.015)	(0.030)	(0.036)	(0.019)	(0.080)	(0.074)
	[0.021]	[0.032]	[0.010]	[0.007]	[0.019]	[0.011]

	Panel C: Mid Stage (B–C)			**Panel D: Late Stage (D+)**		
	Logit	XGBoost	CatBoost	Logit	XGBoost	CatBoost
Follow-on						
$\text{HasBenefit}_{i,t-1}$	-0.173***	-0.124***	-0.110***	-0.043	-0.174***	-0.160***
	(0.029)	(0.032)	(0.033)	(0.057)	(0.057)	(0.048)
	[0.051]	[0.029]	[0.023]	[0.004]	[0.039]	[0.041]
Acquisition						
$\text{HasBenefit}_{i,t-1}$	-0.036	-0.093***	-0.142***	-0.125**	-0.096	0.025
	(0.030)	(0.029)	(0.030)	(0.059)	(0.060)	(0.061)
	[0.003]	[0.016]	[0.033]	[0.010]	[0.005]	[0.000]
IPO						
$\text{HasBenefit}_{i,t-1}$	-0.094***	-0.082	-0.122**	-0.091	-0.041	0.019
	(0.033)	(0.057)	(0.057)	(0.068)	(0.046)	(0.059)
	[0.014]	[0.006]	[0.010]	[0.006]	[0.002]	[0.001]

Standard errors clustered by investor in parentheses

Within R^2 in brackets (two-way demeaning: investor and year)

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

Return to Section 5.2.3

Table 3: Persistence of Predictive Benefit

This table reports panel regression results analyzing the persistence of investors' predictive benefit over time. The dependent variable is $Benefit_{it}$, defined as the difference between the hit rate of a counterfactual portfolio constructed using model-based predictions and the investor's realized hit rate in year t. The main explanatory variable is the lagged benefit, $Benefit_{i,t-1}$. Results are reported separately by success outcome—follow-on financing, acquisition, and IPO—and by prediction model (Logit, XGBoost, and CatBoost). **Panel A** reports results pooling all investment stages, while **Panels B–D** report results separately for early-stage (Seed–A), mid-stage (B–C), and late-stage (D+) investments. All specifications include investor and year fixed effects, and standard errors are clustered at the investor level. The precise definition of all variables is provided in Appendix D. Clustered standard errors are reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

	Panel A: All Stages			**Panel B: Early Stage (Seed–A)**		
	Logit	XGBoost	CatBoost	Logit	XGBoost	CatBoost
Follow-on						
$\text{Benefit}_{i,t-1}$	-0.178***	-0.161***	-0.137***	-0.146***	-0.147***	-0.129***
	(0.021)	(0.022)	(0.021)	(0.027)	(0.029)	(0.029)
	[0.024]	[0.021]	[0.018]	[0.016]	[0.017]	[0.014]
Acquisition						
$\text{Benefit}_{i,t-1}$	-0.051**	-0.074***	-0.106***	-0.087***	-0.118***	-0.117***
	(0.021)	(0.020)	(0.021)	(0.028)	(0.027)	(0.026)
	[0.004]	[0.006]	[0.011]	[0.009]	[0.015]	[0.015]
IPO						
$\text{Benefit}_{i,t-1}$	-0.075***	-0.124***	-0.032	-0.140***	-0.123***	-0.110***
	(0.020)	(0.019)	(0.021)	(0.037)	(0.027)	(0.034)
	[0.006]	[0.017]	[0.001]	[0.012]	[0.014]	[0.010]

	Panel C: Mid Stage (B–C)			**Panel D: Late Stage (D+)**		
	Logit	XGBoost	CatBoost	Logit	XGBoost	CatBoost
Follow-on						
$\text{Benefit}_{i,t-1}$	-0.151***	-0.198***	-0.159***	-0.038	-0.236***	-0.110*
	(0.036)	(0.033)	(0.036)	(0.075)	(0.066)	(0.064)
	[0.017]	[0.028]	[0.019]	[0.001]	[0.035]	[0.009]
Acquisition						
$\text{Benefit}_{i,t-1}$	-0.125***	-0.064	-0.140***	-0.041	-0.062	-0.050
	(0.041)	(0.036)	(0.034)	(0.077)	(0.061)	(0.068)
	[0.009]	[0.002]	[0.014]	[0.001]	[0.003]	[0.002]
IPO						
$\text{Benefit}_{i,t-1}$	-0.116***	-0.049	-0.035	-0.086	-0.074	-0.139*
	(0.035)	(0.042)	(0.038)	(0.096)	(0.072)	(0.083)
	[0.010]	[0.001]	[0.001]	[0.002]	[0.003]	[0.007]

Standard errors clustered by investor in parentheses

Within R^2 in brackets (two-way demeaning: investor and year)

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

Return to Section 5.2.3

Table 4: Predictive Benefit and Distance to the Counterfactual Portfolio

This table reports panel regression results relating investors' predictive benefit to the distance between their realized investment portfolios and model-implied counterfactual portfolios. The dependent variable is $Benefit_{it}$. The main explanatory variable is $Dist_{it}$, a composite distance index capturing differences in portfolio composition between realized and counterfactual investments. Results are reported separately by success outcome—follow-on financing, acquisition, and IPO—and by prediction model (Logit, XGBoost, and CatBoost). **Panel A** reports results pooling all investment stages, while **Panels B–D** report results separately for early-stage (Seed–A), mid-stage (B–C), and late-stage (D+) investments. All specifications include investor and year fixed effects, and standard errors are clustered at the investor level. The precise definition of all variables is provided in Appendix D. Clustered standard errors are reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

	Panel A: All Stages			**Panel B: Early Stage (Seed–A)**		
	Logit	XGBoost	CatBoost	Logit	XGBoost	CatBoost
Acquisition						
Dist_{it}	-0.010	0.036***	0.024**	0.049***	0.032**	0.051***
	(0.008)	(0.009)	(0.009)	(0.013)	(0.014)	(0.014)
	[0.002]	[0.009]	[0.006]	[0.015]	[0.007]	[0.016]
Follow-on						
Dist_{it}	-0.006	-0.018**	0.004	0.008	0.006	0.003
	(0.007)	(0.007)	(0.007)	(0.009)	(0.010)	(0.010)
	[0.003]	[0.006]	[0.001]	[0.002]	[0.001]	[0.001]
IPO						
Dist_{it}	0.006	0.015*	-0.021***	-0.008	-0.002	-0.021***
	(0.006)	(0.008)	(0.007)	(0.007)	(0.008)	(0.008)
	[0.004]	[0.006]	[0.009]	[0.002]	[0.000]	[0.009]

	Panel C: Mid Stage (B–C)			**Panel D: Late Stage (D+)**		
	Logit	XGBoost	CatBoost	Logit	XGBoost	CatBoost
Acquisition						
Dist_{it}	0.062**	0.054**	0.041	-0.018	0.052	-0.020
	(0.025)	(0.023)	(0.024)	(0.041)	(0.033)	(0.033)
	[0.012]	[0.009]	[0.006]	[0.003]	[0.006]	[0.002]
Follow-on						
Dist_{it}	0.004	0.010	0.020	0.023	0.009	0.006
	(0.018)	(0.017)	(0.017)	(0.029)	(0.030)	(0.029)
	[0.001]	[0.002]	[0.003]	[0.004]	[0.001]	[0.001]
IPO						
Dist_{it}	0.006	0.054***	0.031*	-0.021	0.091**	0.028
	(0.015)	(0.016)	(0.016)	(0.035)	(0.035)	(0.034)
	[0.008]	[0.020]	[0.010]	[0.003]	[0.018]	[0.004]

Standard errors clustered by investor in parentheses

Within R^2 in brackets (two-way demeaning: investor and year)

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

Return to Section 5.2.3

Table 5: Predictive Benefit, Distance, and Persistence

This table reports panel regression results jointly relating investors' predictive benefit to both portfolio distance and its own persistence. The dependent variable is $Benefit_{it}$. The main explanatory variables are the composite distance index $Dist_{it}$ and the lagged benefit $Benefit_{i,t-1}$. Results are reported separately by success outcome—follow-on financing, acquisition, and IPO—and by prediction model (Logit, XGBoost, and CatBoost). **Panel A** reports results pooling all investment stages, while **Panels B–D** report results separately for early-stage (Seed–A), mid-stage (B–C), and late-stage (D+) investments. All specifications include investor and year fixed effects, and standard errors are clustered at the investor level. The precise definition of all variables is provided in Appendix D. Clustered standard errors are reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

	Logit	XGBoost	CatBoost	N		Logit	XGBoost	CatBoost	N
Panel A: All Stages					**Panel B: Early Stage (Seed–A)**				
Follow-on									
$\text{Benefit}_{i,t-1}$	-0.178***	-0.161***	-0.137***	4,918	$\text{Benefit}_{i,t-1}$	-0.145***	-0.147***	-0.129***	2,825
	(0.021)	(0.022)	(0.021)			(0.027)	(0.029)	(0.029)	
Dist_{it}	-0.002	-0.014*	0.006		Dist_{it}	0.006	0.005	0.001	
	(0.008)	(0.008)	(0.008)			(0.010)	(0.010)	(0.010)	
Within R^2	[0.026]	[0.037]	[0.012]		Within R^2	[0.022]	[0.022]	[0.024]	
Acquisition									
$\text{Benefit}_{i,t-1}$	-0.050**	-0.074***	-0.106***	4,918	$\text{Benefit}_{i,t-1}$	-0.091***	-0.120***	-0.120***	2,825
	(0.021)	(0.020)	(0.021)			(0.028)	(0.027)	(0.025)	
Dist_{it}	-0.012	0.028***	0.019*		Dist_{it}	0.040***	0.019	0.042***	
	(0.010)	(0.010)	(0.010)			(0.013)	(0.014)	(0.014)	
Within R^2	[-0.009]	[0.002]	[0.026]		Within R^2	[-0.023]	[0.009]	[0.027]	
IPO									
$\text{Benefit}_{i,t-1}$	-0.075***	-0.124***	-0.032	4,918	$\text{Benefit}_{i,t-1}$	-0.140***	-0.123***	-0.110***	2,825
	(0.020)	(0.019)	(0.021)			(0.037)	(0.027)	(0.034)	
Dist_{it}	-0.009	-0.016	0.009		Dist_{it}	0.022	0.010	0.009	
	(0.009)	(0.009)	(0.009)			(0.015)	(0.015)	(0.016)	
Within R^2	[0.006]	[0.017]	[0.001]		Within R^2	[0.012]	[0.014]	[0.010]	
Panel C: Mid Stage (B–C)					**Panel D: Late Stage (D+)**				
Follow-on									
$\text{Benefit}_{i,t-1}$	-0.147***	-0.198***	-0.153***	1,067	$\text{Benefit}_{i,t-1}$	-0.032	-0.234***	-0.109*	325
	(0.036)	(0.034)	(0.037)			(0.074)	(0.066)	(0.064)	
Dist_{it}	0.002	0.007	0.018		Dist_{it}	0.021	0.006	0.003	
	(0.018)	(0.017)	(0.017)			(0.030)	(0.030)	(0.030)	
Within R^2	[0.048]	[0.038]	[0.015]		Within R^2	[-0.003]	[0.087]	[-0.023]	

Clustered standard errors (by investor) in parentheses; within R^2 in brackets.
Investor and year fixed effects in all specifications.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Return to Section 5.2.3

Table 6: Decomposed Distance and Predictive Benefit

This table reports panel regression results relating investors' predictive benefit to decomposed measures of portfolio distance. The dependent variable is $Benefit_{it}$. The explanatory variables include the lagged benefit $Benefit_{i,t-1}$ and separate distance components capturing differences between realized and counterfactual portfolios along sectoral composition, geographic exposure, and investment hotness. Results are reported separately by success outcome—follow-on financing, acquisition, and IPO—and by prediction model (Logit, XGBoost, and CatBoost). **Panel A** reports results pooling all investment stages, while **Panels B–D** report results separately for early-stage (Seed–A), mid-stage (B–C), and late-stage (D+) investments. All specifications include investor and year fixed effects, and standard errors are clustered at the investor level. The precise definition of all variables is provided in Appendix D. Clustered standard errors are reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

	(1) Logit	(2) XGBoost	(3) CatBoost	(4) N	(1) Logit	(2) XGBoost	(3) CatBoost	(4) N
Panel A: All Stages					**Panel B: Early Stage (Seed–A)**			
$Benefit_{i,t-1}$ (Follow-on)	-0.177***	-0.160***	-0.137***	4,918	$Benefit_{i,t-1}$ (Follow-on)	-0.146***	-0.147***	-0.129***	2,823
	(0.021)	(0.022)	(0.021)			(0.027)	(0.029)	(0.029)	
$DistHotness_{it}$	-0.000	-0.000	-0.001		$DistHotness_{it}$	0.001	-0.002	-0.002	
	(0.002)	(0.002)	(0.002)			(0.002)	(0.003)	(0.003)	
$DistSector_{it}$	-0.055**	-0.082***	-0.016		$DistSector_{it}$	0.042	0.060**	0.031	
	(0.027)	(0.026)	(0.027)			(0.031)	(0.030)	(0.032)	
$DistGeo_{it}$	0.049	-0.001	0.059*		$DistGeo_{it}$	0.038	0.046	0.075**	
	(0.035)	(0.033)	(0.034)			(0.040)	(0.037)	(0.037)	
Within R^2	[0.027]	[0.037]	[0.011]		*Within* R^2	[0.024]	[0.037]	[0.017]	
$Benefit_{i,t-1}$ (Acquisition)	-0.051**	-0.074***	-0.106***	4,918	$Benefit_{i,t-1}$ (Acquisition)	-0.087***	-0.118***	-0.117***	2,823
	(0.020)	(0.020)	(0.021)			(0.028)	(0.027)	(0.026)	
$DistHotness_{it}$	0.001	0.002	0.002		$DistHotness_{it}$	-0.002	0.003	0.003	
	(0.002)	(0.002)	(0.002)			(0.003)	(0.003)	(0.003)	
$DistSector_{it}$	-0.018	0.023	0.060**		$DistSector_{it}$	-0.000	-0.029	0.027	
	(0.027)	(0.024)	(0.026)			(0.033)	(0.031)	(0.030)	
$DistGeo_{it}$	0.023	0.042	0.001		$DistGeo_{it}$	0.003	-0.010	0.021	
	(0.033)	(0.031)	(0.034)			(0.044)	(0.038)	(0.040)	
Within R^2	[0.005]	[0.011]	[0.017]		*Within* R^2	[0.012]	[0.021]	[0.017]	
$Benefit_{i,t-1}$ (IPO)	-0.075***	-0.124***	-0.032	4,918	$Benefit_{i,t-1}$ (IPO)	-0.139***	-0.123***	-0.110***	2,823
	(0.020)	(0.019)	(0.021)			(0.037)	(0.027)	(0.034)	
$DistHotness_{it}$	0.002	0.001	0.002		$DistHotness_{it}$	0.004	-0.002	-0.001	
	(0.002)	(0.002)	(0.002)			(0.004)	(0.003)	(0.004)	
$DistSector_{it}$	-0.049*	-0.002	-0.018		$DistSector_{it}$	0.010	-0.034	-0.043	
	(0.027)	(0.025)	(0.027)			(0.050)	(0.037)	(0.042)	
$DistGeo_{it}$	0.038	0.049	0.032		$DistGeo_{it}$	-0.042	-0.003	0.020	
	(0.040)	(0.036)	(0.039)			(0.051)	(0.043)	(0.053)	
Within R^2	[0.007]	[0.024]	[-0.004]		*Within* R^2	[0.008]	[0.023]	[0.016]	
Panel C: Mid Stage (B–C)					**Panel D: Late Stage (D+)**			
$Benefit_{i,t-1}$ (Follow-on)	-0.151***	-0.198***	-0.159***	1,071	$Benefit_{i,t-1}$ (Follow-on)	-0.038	-0.236***	-0.110*	327
	(0.036)	(0.033)	(0.036)			(0.075)	(0.066)	(0.064)	
$DistHotness_{it}$	-0.001	0.002	0.001		$DistHotness_{it}$	0.001	-0.001	0.002	
	(0.003)	(0.003)	(0.003)			(0.008)	(0.009)	(0.008)	
$DistSector_{it}$	-0.033	-0.043	-0.022		$DistSector_{it}$	0.023	0.046	-0.029	
	(0.034)	(0.032)	(0.035)			(0.091)	(0.087)	(0.080)	
$DistGeo_{it}$	0.013	0.007	0.006		$DistGeo_{it}$	0.017	0.109	0.107	
	(0.041)	(0.040)	(0.041)			(0.105)	(0.097)	(0.109)	
Within R^2	[0.022]	[0.046]	[0.026]		*Within* R^2	[-0.019]	[0.195]	[0.084]	
$Benefit_{i,t-1}$ (Acquisition)	-0.125***	-0.064*	-0.140***	1,071	$Benefit_{i,t-1}$ (Acquisition)	-0.041	-0.062	-0.050	327
	(0.041)	(0.036)	(0.034)			(0.077)	(0.061)	(0.068)	
$DistHotness_{it}$	0.003	-0.003	0.001		$DistHotness_{it}$	-0.004	-0.003	-0.004	
	(0.003)	(0.003)	(0.003)			(0.008)	(0.008)	(0.009)	
$DistSector_{it}$	0.066*	0.038	0.067**		$DistSector_{it}$	0.023	0.061	0.027	
	(0.034)	(0.033)	(0.033)			(0.087)	(0.077)	(0.085)	
$DistGeo_{it}$	0.009	0.027	0.009		$DistGeo_{it}$	0.158	0.111	0.164	
	(0.046)	(0.044)	(0.044)			(0.115)	(0.126)	(0.113)	
Within R^2	[0.029]	[0.009]	[0.038]		*Within* R^2	[0.016]	[0.028]	[0.010]	
$Benefit_{i,t-1}$ (IPO)	-0.116***	-0.049	-0.035	1,071	$Benefit_{i,t-1}$ (IPO)	-0.086	-0.074	-0.139*	327
	(0.035)	(0.042)	(0.038)			(0.096)	(0.072)	(0.083)	
$DistHotness_{it}$	-0.001	-0.002	-0.002		$DistHotness_{it}$	0.001	0.005	0.003	
	(0.003)	(0.004)	(0.003)			(0.010)	(0.009)	(0.009)	
$DistSector_{it}$	0.054*	0.037	0.059*		$DistSector_{it}$	0.026	0.072	0.071	
	(0.032)	(0.038)	(0.035)			(0.104)	(0.106)	(0.109)	
$DistGeo_{it}$	0.070*	0.047	0.031		$DistGeo_{it}$	-0.027	0.067	-0.076	
	(0.042)	(0.055)	(0.049)			(0.152)	(0.107)	(0.113)	
Within R^2	[0.032]	[0.001]	[0.000]		*Within* R^2	[0.023]	[0.012]	[0.036]	

Clustered standard errors (by investor) in parentheses
Within R^2 in brackets
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Return to Section 5.2.3

Table 7: Difference-in-Differences Estimates of Predictive Benefit Around the ML Shocks

This table reports difference-in-differences estimates examining how investors' predictive benefit responds to the diffusion of modern ML technologies. The dependent variable is $Benefit_{it}$. Exposure is defined as investors' predetermined predictive benefit, measured prior to the shock period. The shock period corresponds to the widespread release and adoption of modern ML libraries between 2015 and 2017; additional institutional details on the technological shock are provided in **Panel B** of Appendix C. The main coefficients of interest are interactions between investor exposure and post-shock indicators. Results are reported separately by success outcome—follow-on financing, acquisition, and IPO—and by prediction model. **Panel A** reports results pooling all investment stages, while **Panels B–D** report results separately for early-stage (Seed–A), mid-stage (B–C), and late-stage (D+) investments. All specifications include investor and year fixed effects, and standard errors are clustered at the investor level. The precise definition of all variables is provided in Appendix D. Clustered standard errors are reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

	(1) Logit	(2) XGBoost	(3) CatBoost	(4) N	(1) Logit	(2) XGBoost	(3) CatBoost	(4) N
Panel A: All Stages					**Panel B: Early Stage (Seed–A)**			
$\text{Exposure}_i \times \text{Post}_t$ (Follow-on)	-0.043***	-0.039***	-0.041***	2,421	-0.061***	-0.073***	-0.069***	1,616
	(0.009)	(0.008)	(0.008)		(0.013)	(0.012)	(0.012)	
	[0.018]	[0.015]	[0.016]		[0.012]	[0.021]	[0.020]	
$\text{Exposure}_i \times \text{Post}_t$ (Acquisition)	-0.057***	-0.050***	-0.049***	2,421	-0.052***	-0.077***	-0.069***	1,616
	(0.010)	(0.008)	(0.008)		(0.013)	(0.012)	(0.012)	
	[0.026]	[0.016]	[0.018]		[-0.005]	[0.013]	[0.011]	
$\text{Exposure}_i \times \text{Post}_t$ (IPO)	-0.022***	-0.038***	-0.041***	2,421	-0.009*	-0.021**	-0.015**	1,616
	(0.006)	(0.007)	(0.007)		(0.005)	(0.010)	(0.007)	
	[0.004]	[0.018]	[0.019]		[-0.001]	[0.009]	[0.010]	
Panel C: Mid Stage (B–C)					**Panel D: Late Stage (D+)**			
$\text{Exposure}_i \times \text{Post}_t$ (Follow-on)	-0.031***	-0.028***	-0.027***	1,067	-0.025	-0.041	-0.038	267
	(0.007)	(0.007)	(0.008)		(0.032)	(0.041)	(0.039)	
	[0.010]	[0.012]	[0.013]		[-0.012]	[0.004]	[0.006]	
$\text{Exposure}_i \times \text{Post}_t$ (Acquisition)	-0.027***	-0.031***	-0.027***	1,067	-0.064*	-0.060	-0.051	267
	(0.007)	(0.007)	(0.008)		(0.038)	(0.061)	(0.068)	
	[0.007]	[0.010]	[0.013]		[-0.039]	[0.003]	[0.012]	
$\text{Exposure}_i \times \text{Post}_t$ (IPO)	-0.094***	-0.082	-0.122**	1,067	-0.089*	-0.042	-0.015	267
	(0.033)	(0.057)	(0.057)		(0.048)	(0.046)	(0.034)	
	[0.004]	[0.008]	[0.010]		[0.115]	[-0.024]	[-0.083]	

Standard errors clustered by investor in parentheses

Within R^2 in square brackets

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

Return to Section 5.3.2