Predictability and Persistence in Deal Selection: Evidence from Venture Capital

Luiz Bissoto*
NEOMA Business School

November 20, 2026

Abstract

I study whether predictive technologies reveal economically meaningful deal-selection gains in venture capital (VC), and why such gains can persist in a competitive market. Using U.S. financing rounds from 2000 to 2021, I train investment-stage-specific machine learning (ML) models to predict follow-on financing, acquisition, and initial public offering (IPO) outcomes strictly out of sample. I compare actual VC portfolios with random portfolios, largest-round portfolios, and ML-ranked portfolios. In terms of the share of successful investments, actual VC portfolios outperform random selection by 4.7 p.p. on average across outcomes. ML-ranked portfolios raise this advantage to 30 p.p. on average, with the largest gains for IPOs. Relative to the largest-round benchmark, ML adds little for follow-on financing, modestly improves IPO prediction, and substantially improves acquisition prediction, suggesting that predictive tools add the most value where simple scale signals perform poorly. The ML-versus-actual gap is noisy but persistent within investment stage. Geography, sector, and measures of market hotness explain little of it. Exploiting a plausibly exogenous shock to ML implementation costs, I show that investors highly exposed to this shock experience stronger gap compression, especially in early- and mid-stage investments. Overall, the evidence is consistent with predictable selection gains persisting because predictive capabilities are costly to implement, and eroding as those capabilities diffuse.

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

JEL Codes: G11, G17, G23, G24

*Reims Campus, 59 rue Pierre Taittinger, 51100 Reims, France. Email: luiz.bissoto@neoma-bs.fr.

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 expanded the set of tools available for predictive analysis in many forecasting tasks. These tools have increasingly been applied to VC investment screening and classification (Bonelli (2026)), 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)). This paper focuses on the economic implications of this hypothesized predictability for VC deal selection. First, I quantify the gains that can be generated using information available at the time of investment. Then, I study whether simple selection rules already capture these gains, where they persist, and how they evolve as prediction technology diffuses. I refer to potential success-rate gains from exploiting predictive information before competitors fully incorporate it as *prediction rents*.

I organize the empirical analysis in three steps. First, I construct strictly out-of-sample ML predictions and compare portfolios ranked by these predictions with actual VC portfolios and two transparent benchmarks: random selection and selection by financing-round size.[2] This establishes the magnitude and content of potential prediction rents. Second, I test whether the ML-versus-actual success-rate gap contains a persistent investor-specific component.[3] Third, I examine why the gap persists, contrasting observable investment constraints—including geography, sector specialization, and exposure to crowded sectors—with plausible implementation frictions.

I work with financing rounds of U.S.-based startups from 2000 to 2021. For each round, I

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

[2] Round size provides a particularly demanding non-ML benchmark because it reflects investors' willingness to commit capital and therefore embeds their assessment of the investment opportunity.

[3] The ML-versus-actual gap can be interpreted as an inverse measure of how much of the ML-implied selection advantage an investor captures. A smaller gap indicates that the investor's realized portfolio is closer to the ML benchmark and therefore captures a larger share of the potential prediction rents; conversely, the gap measures the portion of those gains left unrealized.

construct predictive variables using only information available at the financing date, including investor experience, startup sector, and prior financing history. I distinguish three post-investment outcomes: subsequent financing, acquisition, and initial public offering (IPO). I estimate logit and extreme gradient boosting (XGBoost) models separately by outcome and financing-round type, using rolling training windows so that all predictions are strictly out of sample.[4] I use these predictions to construct counterfactual portfolios that are comparable in size to realized VC portfolios and evaluate performance both across all investments and separately for early-, mid-, and late-stage deals.

I find two main results. First, ML-based portfolios generate economically large selection gains, but the incremental value of ML varies sharply across outcomes. Actual VC portfolios outperform random selection by 4 p.p. for follow-on financing, 6.4 p.p. for acquisitions, and 3.8 p.p. for IPOs.[5] A simple rule that selects the largest financing rounds outperforms the same random benchmark by 15.5 p.p. for follow-on financing and 39.8 p.p. for IPOs, but performs 7.5 p.p. worse than random for acquisitions. The strong performance of this simple rule for two of the three outcomes confirms that round size is a demanding benchmark rather than a mechanical comparison. ML-selected portfolios outperform random selection by 15.1, 32.2, and 42.7 p.p., respectively. Thus, ML adds little beyond round size for follow-on financing, modestly improves IPO selection, and adds substantial value for acquisitions, where the scale signal points in the wrong direction. The gains are concentrated in early- and mid-stage investments, indicating that economically useful predictability emerges well before late-stage rounds, when success may be easier to infer from startup maturity and scale. This stage pattern weakens a pure deal-access interpretation: the largest predictive gains arise before deal flow access becomes especially restricted rather than being confined to highly competitive late-stage rounds.

Second, the ML-versus-actual gap contains a persistent investor-specific component and compresses as prediction technology becomes cheaper to implement. Positive gaps are widespread across investors and are particularly large for IPOs under the more flexible ML

[4] I focus on tree-based methods rather than neural networks because the data are predominantly structured and tabular, with a moderate number of predictors and substantial heterogeneity across observations.

[5] The random benchmark is constructed by randomly reassigning success labels among startups in the same year and investment stage, separately for each outcome. It preserves the size and success rate of the market opportunity set but removes the link between startup characteristics and success.

models. Because investor portfolios are often small and startup outcomes are discrete, annual investor-level gaps are measured with substantial noise. Even so, within investment stage, investors' average gaps over 2010–2015 are positively correlated with their average gaps over 2016–2021, with raw Pearson correlations of approximately 0.16 at early stages and 0.23 at mid stages. Investors with larger model-implied selection gains therefore tend to retain larger gaps over time, even relative to other investors operating at the same stage. Geography, sector specialization, and measures of sector hotness explain little of the gap. To examine implementation frictions, I use a separate pre-shock exposure measure based on average benefit over 2013–2015 and exploit a plausibly exogenous reduction in ML implementation costs associated with the release of major ML libraries and methods around 2016–2017. Higher-exposure investors experience stronger compression in 2016–2017, particularly in early- and mid-stage investments. This pattern is consistent with prediction rents persisting because predictive capabilities are costly to implement in practice and eroding as those capabilities become cheaper and more widely available.

I start by reporting the portfolio comparisons both pooling investment stages and separately for early-, mid-, and late-stage investments. The stage-specific results show that the ML-versus-actual advantage is not confined to mature startups: averaged across outcomes, the gap is approximately 24 p.p. at early stages and 20 p.p. at mid stages, compared with 14 p.p. at late stages. Therefore, economically meaningful predictability is already present in early stage rounds, rather than emerging only once startups reach later financing stages.

I then translate the portfolio comparison into an investor-level measure. For each investor-year, I construct an ML-implied portfolio by selecting the investments with the highest predicted success probabilities while holding fixed the number of investments the investor actually made. In the stage-specific analysis, I repeat the comparison within early-, mid-, and late-stage investments, preserving the investor's number of investments in each stage. I define *benefit* as the difference between the success rate of the ML-implied portfolio and that of the investor's actual portfolio. Because investor portfolios are often small and startup outcomes are discrete, annual benefit contains substantial sampling noise. I therefore assess persistence using multi-year averages. The main inference period, 2010–2021, naturally divides into two equal six-year blocks, 2010–2015 and 2016–2021, providing sufficiently long windows

to reduce annual measurement noise while preserving two non-overlapping observations for each investor. The cutoff also precedes the 2016–2017 reduction in ML implementation costs studied later in the paper, making persistence across the two periods a relatively demanding test: investor rankings must remain informative across a period in which prediction rents may themselves begin to compress. For each investor, I compute average benefit in each six-year block, retaining investors observed in at least three distinct years in both periods. I then correlate these averages across investors separately by investment stage, success outcome, and prediction model. The test therefore asks whether investors that were farther from the ML benchmark in 2010–2015 also tended to remain farther from it in 2016–2021.

Next, I examine whether the gap reflects observable differences between the firms selected by ML and the firms in which investors typically operate. For each investor-year, I measure the distance between the ML-implied and actual portfolios along geography, sector composition, and measures of sector hotness. I then regress investor-level benefit on either a composite distance index or its separate components, including investor and year fixed effects. Some of these measures, including the composite index and individual geography, sector, and hotness components, are statistically related to benefit in particular specifications. Their overall explanatory power, however, is very small: the composite specifications have within R^2 values between 0.001 and 0.005, while decomposing distance into its components raises them only to between 0.003 and 0.016. Observable differences in geography, sector, and market hotness therefore account for some of the gap in particular settings, but explain little of its overall variation.

In the final part of the analysis, I exploit a plausibly exogenous reduction in the implementation cost of prediction technology associated with the staggered release of several major ML libraries and methods around 2016–2017, including TensorFlow, PyTorch, Microsoft Cognitive Toolkit, and CatBoost. I interpret these releases as a broad reduction in the fixed cost and complexity of implementing modern prediction methods, rather than as adoption events for particular investors. For this diffusion design, I define *revealed exposure* as average benefit over 2013–2015, the three years immediately preceding the post-shock window. Investors with higher revealed exposure had greater scope to improve their selections using predictive information before implementation costs fell—without that requiring that

they necessarily subsequently adopted these technologies themselves. I then ask whether benefit compresses more strongly for these higher-exposure investors in 2016–2017. It does, particularly in early- and mid-stage investments, where the predictive gains are largest. As a chronological placebo, I instead measure exposure over 2010–2012 and compression over 2013–2014. Some compression is also present in the placebo period, so I interpret the shock-period estimates as evidence of *accelerated* compression around the decline in implementation costs rather than as causal evidence of individual investor adoption. The design does not distinguish whether this compression operates through investors' own adoption, competitors' adoption, or broader equilibrium effects as predictive capabilities diffuse, because investor-level adoption is unobserved and each of these channels can affect the ML-versus-actual gap. The relevant implication is that the gap narrows more rapidly when prediction technology becomes cheaper to deploy. Together with the persistence results, this evidence is consistent with prediction rents surviving while implementation capabilities remain costly and uneven, and becoming harder to sustain as those costs fall.

My contribution is twofold. First, I attempt to move the growing literature on ML-based startup prediction beyond the question of whether startup outcomes are predictable (Krishna et al. (2016); Retterath (2020); Ross et al. (2021); Kiss (2022); Razaghzadeh Bidgoli et al. (2024); Kalbande and Karmore (2024)) toward *why* they are. Using information available at the time of investment and standard, widely accessible prediction methods, I find economically meaningful out-of-sample predictability and show that some of it is already captured by a simple observable signal—financing round size. Second, I provide an implementation-and-diffusion account of this puzzle. Building on work on systematic distortions in VC choices and data-technology adoption (Davenport (2022); Lyonnet and Stern (2022); Bonelli (2026)), I exploit a plausibly exogenous technological shock to show that prediction rents compress most where investors had the greatest ex ante scope to benefit from prediction, supporting the view that diffusion contributes to their erosion.

I contribute primarily to two strands of literature. The first studies prediction and selection when decisions precede outcomes and market prices provide limited guidance. Work on startup and private-market prediction shows that observable characteristics contain information about subsequent financing and exit outcomes (Krishna et al. (2016); Retterath

(2020); Ross et al. (2021); Kiss (2022); Razaghzadeh Bidgoli et al. (2024); Kalbande and Karmore (2024)). A related literature studies VC screening and selection. Deal selection is a central source of value creation, investors place substantial weight on observable characteristics of entrepreneurs, businesses, and financing rounds, and screening effectiveness depends on expertise and available information (Gompers et al. (2020); Casamatta and Haritchabalet (2007); Block et al. (2019); Jang and Kaplan (2025)). More broadly, research in economics and finance uses algorithmic counterfactuals to evaluate human decisions against statistical predictions (Kleinberg et al. (2018); Erel et al. (2021); Lyonnet and Stern (2022); Davenport (2022)). Related evidence shows that flexible statistical methods can extract economically meaningful nonlinear predictive structure from conventional observables (Amini et al. (2021); Kelly et al. (2024); Chen et al. (2025)). Together, these literatures establish both the scope for prediction and the importance of selection, while leaving open how long economically valuable predictive differences can persist.

The second strand studies how information technologies, information-acquisition costs, and investor capabilities shape financial decisions and capital allocation. In lending, ML changes screening and the distribution of credit (Fuster et al. (2022)); in household finance, automated advice affects portfolio choice and can attenuate behavioral biases (D'Acunto et al. (2019)); and in capital markets, alternative data changes the cost and horizon of information acquisition (Dessaint et al. (2024)). Entrepreneurial-finance theories similarly emphasize costly information production, screening expertise, and limits to investor capacity (Yang and Zeng (2019); Casamatta and Haritchabalet (2007); Kanniainen and Keuschnigg (2004)), while lower experimentation costs can affect VC investment strategies and industry structure (Ewens et al. (2018)). Recent evidence brings these forces directly into VC: Bonelli (2026) shows that investors adopting data technologies screen historically familiar startups more effectively while financing fewer rare breakthrough firms. Taken together, this literature implies that the effects of predictive technologies depend not only on what they can predict, but also on the costs and capabilities required to incorporate predictions into investment decisions. My results add a market-level dimension by showing how the value of predictive capability changes as those constraints loosen.

2 Hypotheses Development

I develop three sequential hypotheses to study whether and why predictability can exist and persist in a competitive VC market. Hypothesis **H1** first asks whether information available at the time of investment contains predictive structure that ML methods can translate into economically meaningful selection gains. Conditional on such gains existing, Hypothesis **H2** asks whether they are persistent at the investor level. Conditional on both existence and persistence, Hypothesis **H3** examines whether these rents remain durable as prediction technology becomes cheaper to implement. The three hypotheses therefore move sequentially from the existence of prediction rents, to their persistence, and finally to their durability as implementation costs fall (Figure 1). I detail the economic reasoning and empirical implications of each hypothesis next.

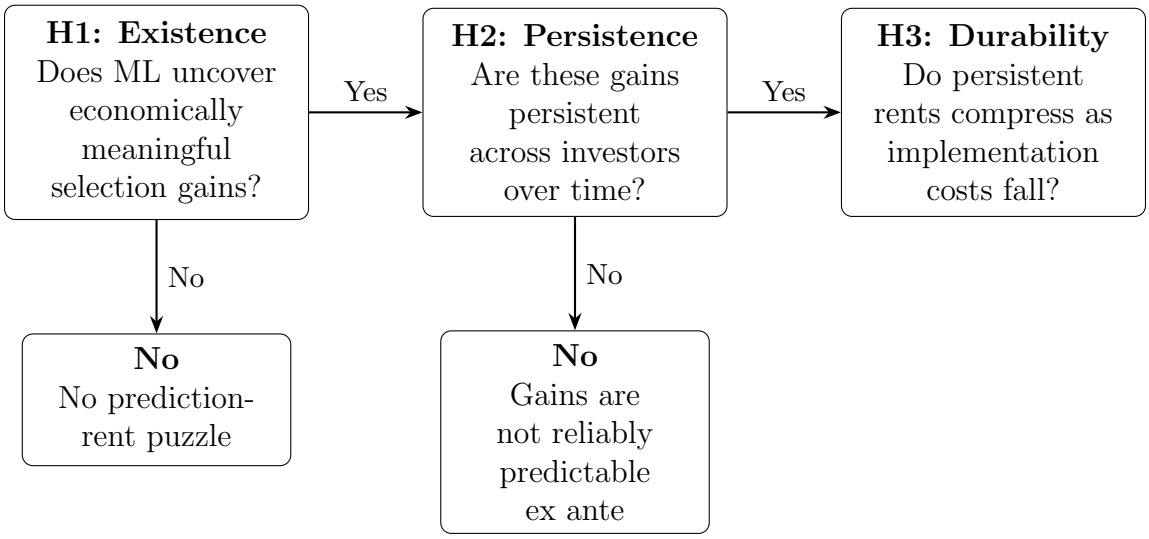


Figure 1: Sequential structure of the hypotheses.

H1. Existence of Prediction Rents. The first step is to establish whether information available at the time of investment contains predictive structure that can improve deal selection. Economically meaningful prediction rents exist when portfolios formed from out-of-sample ML rankings outperform comparable realized VC portfolios. I use random and round-size portfolios, which help characterize the source and magnitude of these gains but do not impose additional directional predictions.

Hypothesis 1 *Portfolios formed from out-of-sample ML rankings achieve higher success rates than comparable actual VC portfolios.*

A positive selection gain is not sufficient for investors to have an incentive to develop predictive capabilities. The gain must also be sufficiently persistent for investors to anticipate that improvements in prediction can benefit their own future selections. This motivates the second hypothesis.

H2. Persistence of Prediction Rents. Existing work documents persistence in realized VC fund performance (Harris et al., 2023). The object studied here is different: persistent differences in model-implied benefit imply that some investors remain systematically farther from the predictive benchmark than others. Because investors tend to maintain stable stage specializations and model-implied benefits differ systematically across stages, persistence should be evaluated among investors operating at comparable investment stages.

Hypothesis 2 *Model-implied selection benefits contain a persistent investor-specific component within investment stage.*

Persistence can arise for different reasons. In particular, stable differences in investors' feasible investment domains may generate persistent gaps between actual and model-implied portfolios even without differences in predictive capability. Therefore, I examine whether observable differences in geography, sector specialization, and market conditions account for the measured benefit. These tests evaluate an alternative opportunity-set explanation and do not impose a directional hypothesis on its explanatory power.

Persistence also does not imply that prediction rents are permanent. When the capabilities required to exploit predictive information become cheaper to implement, their scarcity should decline and the associated rents should become harder to sustain. This mechanism can explain why predictability may exist and persist in a competitive market without generating permanent rents, motivating the final hypothesis.

H3. Durability of Prediction Rents. A decline in implementation costs should compress the gap between model-implied and actual selection. The implication is strongest for

investors with greater ex ante exposure to prediction rents, because these investors have more scope for the diffusion of predictive capabilities to erode the measured advantage.

Hypothesis 3 *Reductions in the cost of implementing prediction technology compress model-implied selection benefits, with stronger compression among investors with greater ex ante exposure to prediction rents.*

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 (2026)). Second, it is more widely used by VC investors than close commercial alternatives such as PitchBook and Dealroom, supporting its relevance 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. Post-2021 information is used only to classify subsequent success outcomes, not to construct predictive features or investment opportunities. I restrict the sample to startups headquartered in the U.S. and to rounds denominated in U.S. dollars. 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 VC 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 in this section the outcome variables and the set of predictive features used in my 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. These events capture distinct economically important milestones in the startup lifecycle: continued access to external capital, an exit through acquisition, and an exit through the public market.

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 equals one if the relevant outcome occurs at *any* point after the focal financing round.[6]

These indicators do not measure realized deal-level returns and therefore do not distinguish, for example, between IPOs generating different return magnitudes. The lack of comprehensive investment-level return data for large VC investment samples is a known problem in the literature. Yimfor and Garfinkel (2023) explicitly examine how commonly observed entrepreneurial-finance success proxies map into VC fund returns and discuss follow-on funding as an alternative proxy for young startups. As such, I adopt the use of subsequent

[6]In the sample, follow-on financing is common, IPOs are rare, and acquisitions are substantially more frequent than IPOs.

financing and exit events as outcome-based measures of investment success, which is the literature standard (Cumming and MacIntosh (2003); Cochrane (2005); Abuzov (2025); Nanda et al. (2020); Hellmann et al. (2021)). Moreover, I estimate the prediction models separately by financing-round type. The relevant comparison is therefore, for example, between Series A investments that subsequently reach an IPO and other Series A investments, rather than between investments made at widely different points in the startup lifecycle. This stage-specific design reduces heterogeneity in the economic interpretation of a given outcome.

Each financing round is therefore associated with up to three distinct success labels, corresponding to the different definitions of success. I analyze the outcomes separately because they differ in their economic meaning, frequency, and relation to startup scale.

Predictive features. The predictive feature set is constructed to reflect the information available at the time of the financing round. All features are measured at the financing-round level and are based exclusively on information observed at or before 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. Name-based variables are included because firm names can convey economically relevant information to investors and markets; for example, Cooper et al. (2001) show that firms adding ".com" to their names during the internet boom experienced large valuation effects. I also include binary indicators for startup sector classifications based on company category (or sector) tags, using the most frequent categories in the data.[7]

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—

[7]The main benchmark results are qualitatively and quantitatively similar when I reestimate the prediction models after excluding all Crunchbase category indicators and investor-category overlap measures.

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[8] 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 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 earlier and later stages in the financing sequence, measured over windows *prior* to the focal round—to capture congestion and momentum across the

[8]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.

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 provides a complete list of all features used in the analysis, grouped by category and reported with their exact construction, while Table 1 provides their summary statistics.[9]

Portfolio benchmarks. The main portfolio analyses compare three ways of evaluating a portfolio with a fixed number k of investments. For each year, investment stage, and success outcome, eligible startups are ranked according to a given selection rule, and the k highest-ranked startups form the corresponding portfolio. The first comparison is a *permuted-label random benchmark*: I randomly reassign realized success outcomes across funded startups within each year, investment stage, and success outcome before evaluating the selected portfolio. This preserves the opportunity set, outcome frequency, stage composition, and portfolio size, but breaks the relationship between startup characteristics and realized success. The second is a *round-size benchmark*, which ranks startups by the amount raised in the focal financing round and selects the k largest rounds. This benchmark captures how much of the available selection advantage is visible from a simple observable rule requiring no ML model. The third is the *ML benchmark*, which ranks startups by their out-of-sample predicted success probabilities and selects the k highest-ranked investments. I report benchmark performance as the percentage-point improvement in the share of successful investments relative to the permuted-label random benchmark.

4 Motivating Evidence: ML Models Applied to VC

This section describes the prediction models and provides a simple illustration of why flexible methods may be useful in VC data. The exercise is not intended to optimize predictive

[9] 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 prediction rents exist, whether they are economically meaningful, how they are distributed across investors, and how they evolve as prediction technology diffuses. 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.

performance or to estimate the main portfolio gains. Its role is to show, in a transparent setting, how nonlinear models can extract structure from sparse startup data.

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 methods are widely used in applied ML for tabular data and are known to perform particularly well in settings with heterogeneous predictors, nonlinear effects, and interactions (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 patterns are learned directly from the data. Panel A of Appendix C summarizes these models and their properties, while Panel B reports the release dates of the open-source ML libraries used later in the technology-diffusion analysis of Section 5.3.

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. 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. 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 regularized prediction procedures rather than unrestricted fitting exercises.

All models are trained using ten-year rolling windows and evaluated strictly out of sample.

For each evaluation period, models are estimated using only information available before the financing rounds being predicted 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.

4.2 Illustrative single-feature exercise

To illustrate the difference between linear and nonlinear prediction in this setting, I use one predictor: the number of distinct investors that previously invested in a startup before its Series A round. The models are trained on Series A observations dated up to December 31, 2018, and used to predict whether startups raising Series A rounds from January 1, 2019 onward eventually complete an IPO.

This exercise does not use the full feature set and is not intended to estimate the main selection gains in the paper. The models produce predicted probabilities in the unit interval. For the illustration, I convert these probabilities into binary predictions using a fixed probability cutoff of 0.40. The cutoff is chosen only to make the classification boundary visible and plays no role in the main portfolio analysis, where portfolios are formed from within-year rankings of predicted probabilities rather than from fixed probability thresholds.

Figure 2 reports the results. The horizontal axis gives the number of prior investors, and the vertical axis indicates whether the startup eventually completes an IPO. The figure reports precision and the F1 score. Precision is the share of predicted IPOs that are realized IPOs, and the F1 score combines precision with recall, the share of realized IPOs identified by the model. Because IPOs are rare, these statistics are naturally low; their role here is to make the classification problem concrete rather than to provide the main evidence.

The figure highlights a useful distinction between linear and tree-based prediction. Logit maps the predictor into a monotonic probability function and therefore imposes a single directional relationship between prior investor attention and IPO likelihood. A tree-based model can instead assign different predictions to different regions of the same variable, allowing the relationship to be non-monotonic. In this exercise, XGBoost identifies correct predictions at values of prior investor attention that are not captured by the logit cutoff,

showing that even simple VC observables may contain nonlinear predictive structure.

5 Main Results

5.1 Investment performance with ML signals

I start by evaluating whether out-of-sample ML predictions translate into economically meaningful portfolio-level differences in success rates. I construct transparent and replicable counterfactual portfolios that use only information available at the financing date and are comparable in scale to actual investor portfolios, while abstracting from deal-access frictions. As mentioned in Section 3.2, I evaluate performance separately for follow-on financing, acquisition, and initial public offering outcomes, and compare ML-based portfolios with actual investor portfolios and simple non-ML benchmarks.

5.1.1 Methodology

The portfolio analysis is conducted at two evaluation levels. In the all-stage analysis, I form and evaluate actual and model-implied portfolios using all eligible financing rounds in a calendar year. In the stage-specific analysis, I repeat the same portfolio construction separately for early-stage rounds (Seed and Series A), mid-stage rounds (Series B and Series C), and late-stage rounds (Series D and later). In both cases, performance is measured using the same hit-rate metric and portfolio sizes are normalized to comparable investor activity levels. The prediction step is separate from this reporting aggregation: models are estimated and scored separately by exact financing-round type, and the resulting out-of-sample probabilities are pooled only when forming the broader all-stage or stage-group portfolio comparisons. I illustrate this approach next.

Performance metric. The main performance metric is the *hit rate*, defined as the fraction of investments in a portfolio that achieve the relevant success outcome. For a portfolio formed

in year t with k_t investments, the hit rate is

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

For ease of notation, I suppress the success-outcome and investment-stage subscripts. The same definition is applied to actual investor portfolios and model-based counterfactual portfolios.[10] For example, if an investor makes three Series A investments and one later succeeds, the Series A hit rate is 1/3. If the same investor makes five Series C investments and one succeeds, the Series C hit rate is 1/5. The pooled analysis combines the investments, producing a hit rate of 2/8, whereas the stage-specific analysis evaluates the two groups separately.

Investor benchmarks. For each calendar year and analysis sample, I compute the cross-sectional distribution of hit rates among *active* investors, defined as investors making at least five investments in that year. I report benchmark hit rates at the 50th, 75th, and 90th percentiles of this distribution. These percentiles summarize realized investor performance and provide the benchmarks against which model-based portfolios are evaluated. Throughout, investor percentiles refer to the hit-rate distribution rather than the distribution of investment activity.

Fixing portfolio size. To compare portfolios at a common investment scale, I set k_t equal to the number of investments made by an investor at the 75th percentile of the investment-count distribution in year t. In the pooled analysis, investment counts include all stages; in the stage-specific analysis, they are computed within the relevant stage group. The counterfactual portfolio therefore has a size representative of a relatively active investor in the corresponding market segment. Investment-count percentiles determine only portfolio size and do not enter the investor-performance benchmarks.

[10] Strebulaev and Dang (2024) emphasize that VC fund performance rankings can be highly sensitive to the realization of a single extreme success. Portfolio-level success rates are therefore a natural performance measure in a setting where returns are driven by rare successful outcomes and deal-level returns are generally unobserved.

Model-based portfolios. Within each year and analysis sample, I rank all eligible funded startups by their predicted probability of success and select the k_t highest-ranked startups. I then evaluate the selected portfolio using the hit-rate metric defined above. The full set of funded startups in the relevant year—and, for stage-specific results, the relevant stage group—constitutes the model-implied opportunity set. The resulting portfolio should therefore be interpreted as a benchmark for selection quality rather than as a claim that every investor could access every selected deal. Because portfolio formation is based on rankings rather than a fixed probability threshold, the selection rule adapts to the distribution of model signals in each year.

Out-of-sample evaluation. I generate predictions using rolling ten-year training windows. Models are estimated using financing rounds in a historical window and then used to score rounds in the subsequent calendar year, beginning with 2000–2009 as the training period for predictions in 2010 and rolling forward annually through 2021. Models are estimated separately for each success outcome and exact financing-round type. The resulting predictions are then combined either across all stages for the pooled analysis or within the early-, mid-, and late-stage groups for the stage-specific analysis. This procedure produces annual out-of-sample predictions under a constant lookback rule while preserving differences across financing-round types. Definitions of the main outcome and benchmark measures used throughout the empirical analysis are collected in Appendix D.1.

5.1.2 Performance Comparison

Table 2 reports the underlying hit rates, beginning with early-stage investments, followed by mid-stage, late-stage, and all-stage results. Within each stage block, the columns first report the hit rates of actual investors and the random benchmark, which establish the relevant outcome base rates, and then those of the round-size and ML portfolios. The final columns report the corresponding differences directly.

The early-stage block illustrates the main pattern. Actual VC portfolios have hit rates of 81.2% for follow-on financing, 35.5% for acquisitions, and 2.5% for IPOs, compared with random hit rates of approximately 76%, 28%, and 2%, respectively. Investors therefore select

better than chance, although the difference is small, particularly for IPO outcome. Ranking startups by round size raises the follow-on hit rate to 91.4% and the IPO hit rate to 35.8%, with an acquisition hit rate of 31.2%, slightly above the random hit rate. ML portfolios generate much larger acquisition hit rates, ranging from 60.0% under logit to 71.7% under CatBoost. The tree-based models also produce early-stage IPO hit rates above 40%, showing that rare right-tail outcomes contain substantial predictable structure well before startups reach mature financing rounds.

Moving to mid-stage investments, the base rates of acquisitions and IPOs rise, but the same outcome-specific distinction remains. The round-size portfolio performs well for follow-on financing and IPOs but slightly worse than random selection for acquisitions. ML improves on the size benchmark for all three outcomes. The most pronounced difference occurs for IPOs: the round-size portfolio has a hit rate of 36%, compared with 68.7% under XGBoost and 69.6% under CatBoost.

At late stages, round size becomes still more informative about follow-on financing and IPOs. The largest-round portfolio reaches hit rates of 91.4% and 49.1% for these outcomes, respectively. It remains a poor acquisition rule, however, producing a hit rate of 23.0%, compared with a random rate of approximately 43%. ML portfolios instead achieve acquisition hit rates between 53.9% and 64.0%. Thus, the acquisition result is not confined to a particular part of the startup lifecycle: at every stage, flexible predictions substantially outperform selection based on scale alone.

The final block pools investments across stages. Actual VC portfolios have hit rates of 81.2% for follow-on financing, 37.3% for acquisitions, and 8.4% for IPOs, compared with random rates of approximately 77%, 31%, and 4.6%, respectively. The pooled results therefore confirm that actual investors outperform random selection, while the much larger differences across the size and ML portfolios depend strongly on the success outcome.

Table 3 summarizes these comparisons by averaging across prediction models and expressing each result in p.p. relative to random selection. The rows within each stage block follow a common order: actual investors relative to random, round size relative to random, ML relative to random, and ML relative to round size. In early-stage investments, ML improves on random selection by 17.2 p.p. for follow-on financing, 36.4 p.p. for acquisitions, and 31.0

p.p. for IPOs. Relative to round size, the corresponding differences are 2.0, 33.2, and −2.8 p.p., showing that the principal incremental gain is for acquisitions. In the mid-stage block, ML exceeds round size for all three outcomes, with incremental gains of 8.9, 11.7, and 23.1 p.p. At late stages, the ML–size differences are −1.6 p.p. for follow-on financing, 36.3 p.p. for acquisitions, and −2.7 p.p. for IPOs. The incremental acquisition advantage therefore remains large even as the value added by ML for the other outcomes attenuates.

The all-stage block at the bottom provides the main pooled summary. Actual investors outperform random selection by 4 p.p. for follow-on financing, 6.4 p.p. for acquisitions, and 3.8 p.p. for IPOs. The round-size advantages are 15.5, −7.5, and 39.8 p.p., while the ML advantages are 15.1, 32.2, and 42.7 p.p. The final row reports the ML–size comparison directly: −0.4 p.p. for follow-on financing, 39.7 p.p. for acquisitions, and 2.9 p.p. for IPOs. Thus, in the model average, round size captures nearly all of the available advantage for follow-on financing and IPOs, whereas ML adds substantial predictive content for acquisitions.

This ordering has a natural economic interpretation. Round size aggregates investors' willingness to commit capital and therefore reflects market assessments of startup scale, maturity, and perceived quality. Larger rounds provide more runway and are closely associated with subsequent financing; they are also concentrated among companies sufficiently large and mature to pursue an IPO. Acquisitions follow a different process. Many acquisition targets do not first become large standalone companies, so ranking startups exclusively by round size can move the selected portfolio away from likely acquisition candidates. ML adds the most where this simple scale signal is misleading rather than merely incomplete.

The individual model results further show that this conclusion is driven primarily by the flexible tree-based methods. XGBoost and CatBoost generally outperform logit for acquisitions and IPOs, although no method dominates the size benchmark in every stage–outcome cell. This heterogeneity reinforces the central point: the value of flexible prediction depends on the economic structure of the outcome being predicted.

The permutation exercise confirms that the ML results are not a mechanical consequence of selecting the highest-ranked observations from a fixed opportunity set. Across 500 replications, ML performance exceeds the 95th percentile of the permuted-label benchmark in every reported stage, outcome, and model combination. The gains therefore depend on the

relationship between observable deal information, model rankings, and subsequently realized outcomes. The annual comparisons reveal a clear lifecycle pattern: ML-based selection performs most strongly before startups reach late-stage maturity, with the advantage emerging in early rounds, remaining substantial at mid stages, and declining only in later rounds.

Figure 3 establishes the pooled benchmark, holding portfolio size fixed at the investment count of a 75th-percentile investor. ML portfolios generally outperform the median investor and frequently approach or exceed the 75th-percentile investor benchmark. The clearest separation appears for acquisitions and, under the tree-based models, for IPOs, whereas follow-on financing remains closer to observed investor performance.

The stage-specific evidence shows that these gains emerge well before startups become mature. Figure 4 documents recurrent advantages in Seed and Series A investments, particularly for acquisitions and IPOs under the tree-based models. This early predictability is economically important because the models identify successful startups before they accumulate the scale and financing history that make later outcomes easier to anticipate. It also weakens a purely access-based interpretation under which ML merely identifies attractive late-stage firms that most investors could not realistically enter, since the gains are already substantial before access becomes especially restricted.

The advantage persists as startups move into Series B and Series C. Figure 5 shows the strongest separation for IPOs under XGBoost and CatBoost, consistent with the especially large mid-stage IPO advantage in the benchmark results. Acquisition and follow-on performance also remain favorable to ML, although the magnitude of the difference varies more across models and years. Together with the early-stage evidence, this pattern indicates that the main predictive gains arise while screening remains economically consequential and outcomes are not yet readily inferred from startup maturity or financing scale.

The pattern changes in later rounds. Figure 6 shows less systematic separation between ML portfolios and high-performing actual investors. By these stages, financing scale and startup maturity already reveal substantial information about follow-on financing and IPO prospects, leaving less scope for predictive models to improve upon observable market signals. Deal access, syndication, and investor reputation may also become more important relative to screening as startups approach exit-relevant milestones. This attenuation is consistent with

evidence that later-stage VC performance may operate partly through access to attractive rounds and syndicates (Nanda et al. (2020)) rather than through selection alone.

Taken together, the evidence provides partial but economically informative support for Hypothesis **H1**. ML rankings outperform actual investor portfolios and random selection, but they do not uniformly outperform the round-size benchmark. For follow-on financing and IPOs, a substantial share of predictability is already visible through financing-round scale, although flexible models add meaningful information in some stages. For acquisitions, the size rule performs poorly while ML rankings perform strongly throughout the startup lifecycle. Predictive modeling in this case therefore adds the most precisely where a simple market-based signal of scale points in the wrong direction.

5.2 Who benefits from ML signals

Having established that ML-based portfolios can generate meaningful selection gains in aggregate, I examine next how these gains map into investor-level outcomes. Specifically, I study whether the benefits from ML-implied selection are widespread across investors and whether they contain an investor-specific component within investment stage, as emphasized in Hypothesis **H2**. Because annual hit rates are computed from relatively small investor-year portfolios, I treat persistence estimates as measurement exercises rather than as direct evidence of investor skill.

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

$$Benefit_{i,t} = HitModel_{i,t} - HitActual_{i,t}. \tag{2}$$

Because the counterfactual ML-based portfolio is scaled to match the investor's actual level of investment activity in year t, 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 actual choices outperformed the ML-based benchmark. Thus, benefit measures the investor-year prediction rent left uncaptured relative to the ML benchmark: investors with lower benefit are those whose actual portfolios already capture more of the selection gains identified by ML. 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 uses the full set of startups funded in the market in a given year as the counterfactual opportunity set. It therefore abstracts from deal-access frictions that may prevent a particular investor from participating in every startup identified by ML signals. This assumption is useful for measuring selection quality, but it means that benefit should be interpreted as a frictionless benchmark rather than as a literal prescription for immediate investment decisions in every identified deal.

5.2.1 Summary Statistics

I report summary statistics for the investor-year panel in Table 4. Across outcomes and models, positive benefit is widespread and meaningful in magnitude. Average benefits are largest for IPOs, especially under tree-based models, reflecting the large gap between actual investor portfolios and model-implied portfolios for rare right-tail outcomes. These patterns differ across investment stages, which motivates reporting persistence and forecastability within stage rather than relying only on pooled estimates.

The investor-level figures show both prevalence and instability. In Figure 7, positive benefit is common across outcomes and years, meaning that many active investors would have had a higher success rate under the ML-implied portfolio than under their actual portfolio. The pattern is especially pronounced for IPO outcomes under tree-based models. I report the median benefit series over time in Figure 8. The series is positive in most years but varies substantially, consistent with the fact that investor-year success rates are noisy and depend on discrete startup outcomes.

5.2.2 Persistence and Measurement Reliability

Next, I study whether investor-level benefit is persistent across two non-overlapping six-year periods. The estimates are in Table 5. For each investment-stage, success-outcome, and prediction-model combination, I compute each investor's average benefit over 2010–2015 and over 2016–2021. I retain investors observed in at least three distinct years in each period and compute the Pearson and Spearman correlations between the two averages across this matched set of investors. The table reports averages of these correlations across the nine outcome-model combinations within each stage. The strict activity requirement leaves too few matched late-stage investors for the same reliability summary, so I report early-stage, mid-stage, and pooled results. The within-stage Pearson correlations are 0.155 at early stages and 0.231 at mid stages. The pooled correlation is higher, 0.419, because it also reflects persistent stage specialization: investors that repeatedly invest in stages where ML-implied gaps are larger will appear more persistent even if within-stage selection quality is only moderately persistent.

The next columns assess measurement reliability within the 2010–2015 block. I split that period into alternating calendar years—2010, 2012, and 2014 in one half, and 2011, 2013, and 2015 in the other—and correlate investor average benefit across the two halves. The split-half correlations are 0.271 at early stages, 0.225 at mid stages, and 0.345 in the pooled sample. Applying the Spearman–Brown correction, $2r/(1+r)$, gives estimated reliabilities for the full 2010–2015 six-year benefit measure of 0.417, 0.325, and 0.503, respectively. These values indicate that substantial measurement noise remains even after benefit is averaged over multiple years. The final columns report variance decomposition results: investor and year fixed effects jointly explain about 0.61 of the pooled variance, with a residual share of about 0.39.

This evidence supports Hypothesis **H2**, but in a qualified sense. Persistence across the two six-year periods is positive but moderate within investment stage and considerably stronger in the pooled sample, where persistent stage specialization also contributes to the correlation. At the same time, the split-half results show that even multi-year measures of benefit contain substantial noise. The ML-versus-actual gap is therefore neither perfectly

stable nor purely transitory: it contains an investor-specific component that remains visible across widely separated periods once stage specialization and measurement reliability are taken into account.

The measurement-noise result is also useful for interpreting apparent compression in benefit. Panel A of Table 6 uses the same 2010–2015 and 2016–2021 averages as the persistence exercise. Because the change is defined as later-period average benefit minus earlier-period average benefit, the slope from regressing this change on earlier-period benefit is mechanically equal to the slope from regressing later-period benefit on earlier-period benefit minus one. A negative change-on-level coefficient therefore does not by itself imply economic mean reversion. Panel B provides a separate annual diagnostic using Arellano–Bond dynamic-panel specifications. Taken together, these exercises reinforce the distinction between persistent cross-investor differences in benefit and substantial transitory variation in its measured annual realizations.

5.2.3 Portfolio Distance and Investor-Level Constraints

In this section, I study whether the gap arises because ML-implied portfolios select firms outside investors' usual investment domains. *Distance* captures observable differences between ML-implied and actual portfolios along dimensions such as sector, geography, and market hotness, serving as proxy for static differences in investors' investment domains and opportunity sets. Appendix D.2 provides the exact construction of the composite index and its individual components. I estimate two sets of investor-year regressions. The first uses a composite distance index:

$$Benefit_{i,t}^{s,o,m} = \beta Distance_{i,t}^{s,o,m} + \alpha_i + \delta_t + \varepsilon_{i,t}^{s,o,m}, \tag{3}$$

where $Benefit_{i,t}^{s,o,m}$ is the difference between the ML-implied hit rate and the actual hit rate for investor i in year t, investment stage s, success outcome o, and prediction model m. $Distance_{i,t}^{s,o,m}$ is a standardized index of observable distance between the investor's actual portfolio and the corresponding ML-implied portfolio. Investor fixed effects, α_i, absorb time-invariant differences across investors, while year fixed effects, δ_t, absorb aggregate changes

in the startup market.

The coefficient β asks whether investors have larger benefit when the ML-implied portfolio is farther from their actual portfolio in observable space. A positive coefficient would mean that the model adds more value precisely when it selects firms that are farther from the investor's usual investment domain. A low within R^2, by contrast, means that these observable distance measures explain little of the within-investor variation in benefit.

The second specification decomposes distance into its components:

$$\begin{aligned} Benefit_{i,t}^{s,o,m} = \beta_1 HotnessDistance_{i,t}^{s,o,m} + \beta_2 SectorDistance_{i,t}^{s,o,m} \\ + \beta_3 GeographicDistance_{i,t}^{s,o,m} + \alpha_i + \delta_t + \varepsilon_{i,t}^{s,o,m}. \end{aligned} \tag{4}$$

This specification asks whether any specific observable dimension accounts for the gap. $HotnessDistance$, $SectorDistance$, and $GeographicDistance$ measure differences in market activity, sector composition, and location, respectively. Standard errors are clustered at the investor level in all specifications.

Table 7 reports the estimates. The composite specification has very low explanatory power: the within R^2 is 0.001 in the pooled sample, 0.003 at early stages, and 0.005 at mid stages. Decomposing distance increases explanatory power only modestly. The decomposed within R^2 is 0.003 in the pooled sample, 0.006 at early stages, and 0.016 at mid stages. Some distance components are statistically related to benefit in particular outcome-model combinations. Sector distance is significant in four of nine pooled specifications and four of nine early-stage specifications, while market-hotness measures are significant in five of nine mid-stage specifications. Thus, these observable differences are related to benefit in some specifications, but collectively explain very little of the gap between ML-implied and actual performance. In Table 8, I report the underlying coefficients for each stage, outcome, and model. These associations show that model-implied portfolios sometimes differ from actual portfolios along observable investment-domain dimensions. However, the low within R^2 values again indicate that these observable dimensions explain little of the benefit gap.

Overall, observable portfolio distance and its individual components explain little of the gap between actual and model-implied selection. These findings weaken an observable

opportunity-set explanation for persistent benefit: static differences in investors' usual investment domains account for only a small share of the gap.

5.3 Prediction Technology Diffusion and Rent Erosion

I conclude by studying whether prediction rents erode as prediction technology becomes cheaper and more widely available, as formalized in Hypothesis **H3**. This analysis is motivated by the preceding results showing that predictive benefit is real, partly persistent, and not explained by static observable constraints such as sectoral focus, geography, or market positioning. 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 the diffusion of prediction 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 around 2016–2017. TensorFlow was released in November 2015, so I treat 2015 as effectively pre-shock in the main timing convention and use 2016–2017 as the post-shock implementation window. These releases include scalable gradient-boosting and neural-network frameworks that made advanced prediction methods easier to implement 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 prediction tools beyond specialized research teams.[11]

Panel structure and exposure definition. The unit of observation is an investor–year. To capture heterogeneity in exposure to prediction-rent erosion, I construct an investor-level measure of pre-shock exposure using average investor-level benefit over 2013–2015. This

[11]The relevant libraries and their release dates are summarized in Panel B of Appendix C, introduced earlier in Section 4.1.

three-year window is distinct from the 2010–2015 block used in the persistence exercise and places the exposure measure immediately before the 2016–2017 post-shock implementation window. Measured using the same portfolio construction as in the main analysis, it produces a single, time-invariant measure of *revealed exposure* for each investor, which is then assigned to all investor–year observations in the estimation window. Exposure is computed exclusively using pre-shock data and is held fixed throughout the analysis. Appendix D.3 collects the definitions of revealed exposure and the additional portfolio-size, reliability, and prediction-procedure measures used in the investor-level and diffusion analyses.

The measure captures how much an investor's actual portfolio decisions differed from ML-implied counterfactual portfolios before advanced ML tools became cheaper and more widely accessible. As such, exposure does not require that an investor actively used ML methods in the pre-shock period. Instead, it measures the investor's ex ante exposure to prediction rents: investors with higher pre-shock benefit are those for whom ML-implied selection would have generated larger gains before prediction technology diffused.

Specification. I estimate the following two-period difference-in-differences regression:

$$Benefit_{i,t} = \beta \left(Exposure_i \times Post_t \right) + \alpha_i + \delta_t + \varepsilon_{i,t}, \tag{5}$$

where $Benefit_{i,t}$ is investor-year benefit, $Exposure_i$ is average pre-shock benefit measured over 2013–2015, $Post_t$ equals one in 2016–2017 and zero otherwise, α_i are investor fixed effects, and δ_t are calendar-year fixed effects.[12] Exposure is standardized within each stage, outcome, and model specification, so β measures the post-shock change in benefit associated with a one-standard-deviation increase in pre-shock revealed exposure. Standard errors are clustered at the investor level.

Placebo design. I also estimate a chronological placebo design. The placebo measures exposure over 2010–2012 and defines the post period as 2013–2014, before the main ML infrastructure shock. This test asks whether high-benefit investors were already experiencing

[12] The main effect of $Exposure_i$ is absorbed by investor fixed effects because exposure is time invariant, while the main effect of $Post_t$ is absorbed by calendar-year fixed effects because it varies only by calendar year.

comparable compression before the implementation-cost decline. The placebo is not expected to be exactly zero, because measured benefit contains sampling noise and some general reallocation or reversion may occur over time. The relevant comparison is therefore whether actual shock-period compression is larger and more systematic than placebo compression.

Interpretation. This specification is a continuous-treatment difference-in-differences design, where treatment intensity is given by investors' predetermined exposure to prediction rents. The coefficient measures the differential post-shock change in benefit associated with higher pre-shock exposure. Because benefit is measured as a difference in hit rates, coefficients are expressed in p.p. units after multiplying by 100. For example, a coefficient of -0.04 means that a one-standard-deviation increase in pre-shock exposure is associated with a 4 p.p. larger post-shock decline in benefit. Negative post-shock coefficients therefore indicate that investors with larger ex ante prediction rents experience larger subsequent declines in benefit following the reduction in ML implementation costs. The mechanism can operate through investors' own adoption of predictive tools, competitors' adoption, or broader equilibrium changes in deal selection as predictive information becomes less scarce.

5.3.2 The Evolution of Prediction Rents around Technology Diffusion

I compare actual shock-period compression with placebo compression in Table 9. The strongest evidence is concentrated at early stages. The actual shock coefficient averages -0.044 across outcome-model cells and is significant at the 5% level in eight of nine specifications, while the placebo coefficient averages -0.016 and is significant in three of nine specifications. Because exposure is standardized and benefit is measured as the hit-rate gap between ML-implied and actual portfolios, the actual coefficient implies that a one-standard-deviation increase in pre-shock exposure is associated with a 4.4 p.p. larger decline in that gap after the implementation-cost shock. Economically, investors with greater ex ante prediction rents therefore move substantially closer to the ML-implied selection frontier after implementation costs fall. The actual-minus-placebo difference averages -0.031 and is significant in four of nine specifications, implying about 3.1 p.p. greater compression around the actual shock than during the earlier placebo period for a one-standard-deviation increase

in exposure. The pooled and mid-stage estimates point in the same direction but provide weaker evidence of acceleration. In the pooled sample, actual compression averages 4.1 p.p., compared with 2.8 p.p. in the placebo period, leaving an average actual-minus-placebo difference of 1.5 p.p. that is significant in only one of nine specifications. At mid stages, actual compression averages 4.8 p.p., compared with 2.8 p.p. in the placebo period, for an average difference of 2.3 p.p. that is significant in two of nine specifications. Thus, higher-exposure investors generally experience greater subsequent compression, but the evidence that this compression accelerates specifically around the decline in ML implementation costs is strongest at early stages.

These averages can mask whether accelerated compression is broad-based or concentrated in particular outcomes, prediction models, or post-shock years. I therefore examine the underlying stage–outcome–model cells and estimate the post-shock interaction separately for 2016 and 2017. The estimates are reported in Table 10. At early stages, the clearest acceleration appears in follow-on financing and IPO prediction. For follow-on financing, the actual-minus-placebo estimates range from -0.053 to -0.075 across the three models, implying that a one-standard-deviation increase in exposure is associated with 5.3–7.5 p.p. more compression around the actual shock than in the placebo period; the differences are significant at the 5% level for XGBoost and CatBoost. For IPOs, the corresponding differences are -0.070 under XGBoost and -0.046 under CatBoost, implying 7.0 and 4.6 p.p. of additional compression, respectively, and both are statistically significant. By contrast, the early-stage acquisition specifications exhibit negative actual-shock coefficients but no statistically significant actual-minus-placebo differences, indicating compression without clear evidence that it accelerated around the implementation-cost decline. At mid stages, the evidence of acceleration is likewise concentrated rather than uniform: the statistically significant actual-minus-placebo differences occur for IPO prediction under Logit and CatBoost, at -0.072 and -0.083, respectively. The year-specific estimates further sharpen the timing of the early-stage result: the 2016 coefficient is more negative than the 2017 coefficient in eight of nine cells, suggesting that compression is concentrated near the beginning of the post-shock window rather than building gradually through 2017. Taken together, the cell-level results indicate accelerated compression in specific settings rather than a uniform post-2015

decline in benefit. Therefore, I interpret the evidence as accelerated compression around the decline in ML implementation costs, rather than as direct causal evidence of investor adoption.

Overall, the results show that prediction rents in VC are economically meaningful but not static. Investors with greater pre-shock exposure to ML-implied gains experience stronger compression in benefit when prediction technology becomes cheaper to implement, especially in the early-stage settings where predictive gains are largest. This evidence supports Hypothesis **H3**: persistent gaps between actual and model-implied selection evolve with the diffusion of predictive capabilities in competitive private markets. More broadly, the results show that predictable selection gains are not immediately arbitraged away, but become harder to preserve as sophisticated prediction tools become more accessible. The design does not separately identify whether this compression reflects adoption by exposed investors, adoption by competitors, or broader erosion of investor-specific opportunities, but each channel is consistent with the same economic mechanism: diffusion of predictive capability compresses rents from superior selection.

6 Conclusion

I study whether ML can identify economically meaningful improvements in VC deal selection and why such gains can persist in a competitive investment market. I find substantial out-of-sample predictability, but the incremental value of ML depends on what the market already reveals. Actual VC portfolios outperform random selection, and financing round size alone captures much of the predictability in follow-on financing and IPO outcomes. For acquisitions, however, this simple market signal performs poorly and ML adds substantial predictive value beyond it. The gains also emerge at early and mid stages, before startup maturity and financing scale make outcomes easier to infer. These results show that economically relevant information about future venture outcomes is present at the time of investment, but that the value of sophisticated prediction depends on how much of that information is already reflected in observable market signals.

The gap between actual and model-implied selection is not immediately eliminated by

competition. It contains a persistent investor-specific component, while observable differences in investors' usual investment domains explain little of it. At the same time, this gap is not permanent. When the cost of implementing advanced prediction methods falls, investors with greater ex ante scope for improvement experience stronger subsequent compression in the gap, with the clearest evidence at early stages. My results therefore suggest that predictive advantages can persist when exploiting them requires scarce organizational capability or costly implementation, but become harder to preserve as predictive technology becomes cheaper and more widely available.

More broadly, my results separate the existence of predictability from the ability to appropriate its economic value. A market can contain exploitable predictive structure without that structure being immediately incorporated into investment decisions, because useful predictions still have to be converted into an implementable selection process. Conversely, wider access to predictive tools can erode the gains from superior selection even when the underlying predictability remains. The durability of data-driven advantages therefore depends not only on whether better predictions are possible, but also on the costs and capabilities required to act on them before competitors do. In competitive markets, predictive advantages can coexist with competition when predictive capability is difficult to deploy, but become harder to sustain as that capability diffuses.

References

Rustam Abuzov. Busy venture capitalists and investment performance. *Journal of Financial and Quantitative Analysis*, 60, 2025. doi: 10.1017/S0022109024000759.

Shahram Amini, Ryan Elmore, Özde Öztekin, and Jack Strauss. Can machines learn capital structure dynamics? *Journal of Corporate Finance*, 70:102073, 2021. doi: 10.1016/j.jcorpfin.2021.102073.

Joern Block, Christian Fisch, Silvio Vismara, and René Andres. Private equity investment criteria: An experimental conjoint analysis of venture capital, business angels, and family offices. *Journal of Corporate Finance*, 58:329–352, 2019. doi: 10.1016/j.jcorpfin.2019.05.009.

Maxime Bonelli. Data-driven investors. *The Review of Financial Studies*, 39(7):1909–1969, 2026. doi: 10.1093/rfs/hhaf078.

Catherine Casamatta and Carole Haritchabalet. Experience, screening and syndication in venture capital investments. *Journal of Financial Intermediation*, 16(3):368–398, 2007. ISSN 1042-9573.

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.

Michael J. Cooper, Orlin Dimitrov, and P. Raghavendra Rau. A rose.com by any other name. *The Journal of Finance*, 56(6):2371–2388, 2001. doi: 10.1111/0022-1082.00408.

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.

Francesco D'Acunto, Nagpurnanand Prabhala, and Alberto G. Rossi. The promises and pitfalls of robo-advising. *The Review of Financial Studies*, 32(5):1983–2020, 2019. doi: 10.1093/rfs/hhz014.

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.

Olivier Dessaint, Thierry Foucault, and Laurent Frésard. Does alternative data improve financial forecasting? the horizon effect. *The Journal of Finance*, 79(3):2237–2287, 2024. doi: 10.1111/jofi.13323.

Isil Erel, Léa H. Stern, Chenhao Tan, and Michael S. Weisbach. Selecting directors using

machine learning. *The Review of Financial Studies*, 34(7):3226–3264, 2021. doi: 10.1093/rfs/hhab050.

Michael Ewens, Ramana Nanda, and Matthew Rhodes-Kropf. Cost of experimentation and the evolution of venture capital. *Journal of Financial Economics*, 128(3):422–442, 2018. ISSN 0304-405X.

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

Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. Predictably unequal? the effects of machine learning on credit markets. *The Journal of Finance*, 77(1):5–47, 2022.

Paul A. Gompers, Will Gornall, Steven N. Kaplan, and Ilya A. Strebulaev. How do venture capitalists make decisions? *Journal of Financial Economics*, 135(1):169–190, 2020. ISSN 0304-405X.

Robert S. Harris, Tim Jenkinson, Steven N. Kaplan, and Ruediger Stucke. Has persistence persisted in private equity? evidence from buyout and venture capital funds. *Journal of Corporate Finance*, 81:102361, 2023. doi: 10.1016/j.jcorpfin.2023.102361.

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.

Young Soo Jang and Steven N. Kaplan. Venture capital start-up selection. Working Paper 33483, National Bureau of Economic Research, 2025.

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.

Vesa Kanniainen and Christian Keuschnigg. Start-up investment with scarce venture capital support. *Journal of Banking & Finance*, 28(8):1935–1959, 2004. doi: 10.1016/j.jbankfin.2003.07.005.

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.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 2018. doi: 10.1093/qje/qjx032.

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. Conditionally accepted at The Review of Financial Studies; revised May 2026.

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.

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.

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.

Ming Yang and Yao Zeng. Financing entrepreneurial production: Security design with flexible information acquisition. *The Review of Financial Studies*, 32(3):819–863, 2019. doi: 10.1093/rfs/hhy084.

Emmanuel Yimfor and Jon A. Garfinkel. Predicting success in entrepreneurial finance research. *Journal of Corporate Finance*, 81:102359, 2023. doi: 10.1016/j.jcorpfin.2023.102359.

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. **Panel B** provides an illustrative financing-round observation 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: 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. **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

Feature	Construction and economic interpretation
Capital raised	Total amount of capital raised in the current financing round, measured in U.S. dollars. This captures the scale of the financing event and the amount of capital committed to the startup at the decision date.
Syndicate size	Number of distinct investors participating in the current financing round. This captures how broadly the deal is shared across investors and whether the round attracts a larger financing syndicate.
Lead investor indicator	Indicator equal to one if at least one investor is identified as a lead investor in the round, and zero otherwise. This captures whether the deal has an investor taking a more prominent role.
Number of lead investors	Count of investors identified as lead investors in the round. This distinguishes rounds with a single lead from rounds with multiple lead investors.
Startup name length	Number of characters in the startup name. This is a simple text-based characteristic observed before the financing decision.
Startup name word count	Number of words in the startup name. This captures whether the firm name is short and brand-like or longer and more descriptive.
Startup name unique characters	Number of distinct characters appearing in the startup name. This captures an additional dimension of name structure without using future information.
Startup category indicators	Binary indicators constructed from Crunchbase category tags. Each indicator equals one if the startup is assigned to the corresponding category and zero otherwise. The main category indicators include software, health care, information technology, software-as-a-service (SaaS), and enterprise software. These variables capture the startup's sector and business domain at the time of financing.

Panel B: Investor Experience and Historical Activity

Feature	Construction and economic interpretation
Cumulative prior investment count	Number of financing rounds completed by the participating investors before the current round. At the round level, the measure is aggregated across the syndicate using the mean, sum, and maximum. These variables capture how experienced the investors are before making the current investment.
Cumulative prior invested capital	Total capital previously invested by the participating investors before the current round. At the round level, the measure is aggregated across the syndicate using the mean, sum, and maximum. These variables capture the scale of investors' prior activity and available investment experience.
Rolling prior investment count	Number of prior investments made by the participating investors over rolling windows of 12, 24, 36, 48, 60, and 120 months before the current round. For each horizon, the measure is aggregated across the syndicate using the mean, sum, and maximum. These variables capture recent investor activity and whether investors have been active in the market immediately before the deal.
Rolling prior invested capital	Total capital previously invested by the participating investors over rolling windows of 12, 24, 36, 48, 60, and 120 months before the current round. For each horizon, the measure is aggregated across the syndicate using the mean, sum, and maximum. These variables capture recent capital deployment rather than only deal counts.
Investor category exposure	Historical category exposure of each participating investor, measured from prior investments in each startup category before the current round. These variables summarize the sectors in which the investors have accumulated experience.

Return to Section 3.2

Appendix B: Machine Learning Feature Set (Continued)

Panel C: Investor–Startup Matching Features

Feature	Construction and economic interpretation
Category overlap, mean	Mean weighted Jaccard similarity between the startup's category profile and each participating investor's historical category exposure. The measure is computed using only investments made before the current round. It captures how closely the syndicate's average prior experience matches the startup's sector profile.
Category overlap, maximum	Maximum weighted Jaccard similarity between the startup's category profile and any participating investor's historical category exposure. This captures whether at least one investor in the syndicate has strong prior experience in the startup's sector.
Prior investor network depth	Number of distinct investors that invested in the startup before the current financing round. This captures how much investor attention and financing history the startup has already accumulated.

Panel D: Market Conditions and Competitive Environment

Feature	Construction and economic interpretation
Market activity, same stage	Number of financing rounds in the same investment stage in the overall U.S. market over rolling windows of 12, 24, 36, 48, 60, and 120 months before the current round. These variables capture market-wide activity and congestion in the same part of the financing cycle.
State-level market activity, same stage	Number of financing rounds in the same investment stage and in the startup's U.S. state over rolling windows of 12, 24, 36, 48, 60, and 120 months before the current round. These variables capture local financing conditions and the density of nearby VC activity.
Adjacent-stage activity, previous stage	Number of financing rounds in the immediately preceding investment stage over rolling windows of 12, 24, 36, 48, 60, and 120 months before the current round. These variables capture recent upstream financing activity that may feed into the current stage.
Adjacent-stage activity, next stage	Number of financing rounds in the immediately subsequent investment stage over rolling windows of 12, 24, 36, 48, 60, and 120 months before the current round. These variables capture downstream financing activity and market conditions for startups approaching later stages.

Panel E: Time Controls and Out-of-Sample Design

Feature	Construction and economic interpretation
Calendar-month indicators	Month-of-year indicators for the financing round announcement month. These variables capture seasonality in VC activity and outcome realization.
Month index	Continuous monthly time index based on the financing round announcement date. This variable captures slow-moving trends in VC activity, market development, and data coverage over time.
Out-of-sample timing discipline	All predictive features are constructed using only information available at or before the financing round announcement date. Prediction models are trained using rolling 10-year windows and evaluated out of sample by predicting outcomes in the subsequent year. This timing rule prevents the models from using future information.

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. **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 (November 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.
CNTK (January 2016)	Enterprise-grade machine learning and deep learning framework, also known as the Microsoft Cognitive Toolkit, designed to support large-scale model training and deployment. The toolkit expanded access to industrial-strength machine learning infrastructure for applied users.
PyTorch (January 2017)	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 (July 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.1: Variable Definitions

This exhibit defines the main investor–year variables and benchmark concepts used in the empirical analysis. **Panel A** defines actual and model-implied success-rate measures, together with the random and round-size benchmark concepts used in the portfolio-comparison tables.

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 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.
$Benefit_{i,t}$	The difference between the model-implied hit rate and the realized investor hit rate: $Benefit_{i,t} = HitModel_{i,t} - HitActual_{i,t}$. A positive value indicates that the model-selected portfolio outperforms the investor's realized portfolio in terms of hit rate.
$HasBenefit_{i,t}$	Indicator 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.
Random benchmark hit rate	Hit rate from a top-k portfolio after realized success labels are randomly permuted within year, investment stage, and success outcome. This is reported as "Random hit" in Appendix Table 2. The benchmark preserves the opportunity set, outcome frequency, and portfolio-size distribution, but breaks the link between observable characteristics and realized success. The main permutation diagnostics use 500 replications with a fixed seed.
Round-size benchmark hit rate	Hit rate from a top-k portfolio that ranks startups only by the amount raised in the focal financing round. This is reported as "Size hit" in Appendix Table 2. The benchmark measures the selection advantage available from a simple non-ML observable rule.
Advantage over random benchmark	Improvement, in p.p., in the share of successful investments relative to the permuted-label random benchmark. These concepts appear in Table 3 as "Actual - random," "Round size - random," and "ML - random."

Return to Section 5.1.1

Appendix D.2: Variable Definitions (Continued)

This exhibit continues Appendix D. **Panel B** defines observable distance measures capturing differences between actual investor portfolios and model-implied counterfactual portfolios.

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. Market hotness is 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.
$DistSector_{i,t}$	Cosine distance between the sector exposure vector of the investor's realized portfolio and that of the model-implied top-k portfolio. 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, computed using the distribution of investments across U.S. states.

Return to Section 5.2.3

Appendix D.3: Variable Definitions (Continued)

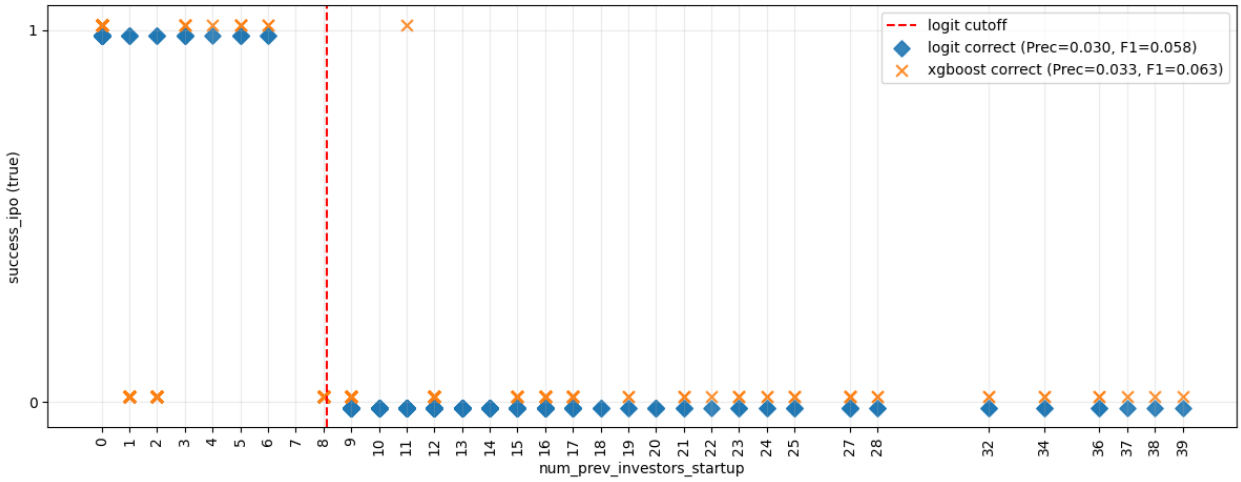
This exhibit continues Appendix D. **Panel C** defines portfolio-size, reliability, and exposure variables used in the investor-level and diffusion analyses. **Panel D** summarizes the prediction models and out-of-sample estimation procedure.

Panel C: Portfolio Size, Reliability, and Exposure	
$NInv_{i,t}$	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 figures, defined as the number of investments made by an investor at the 75th percentile of the investor activity distribution in a given year.
Split-period persistence	Correlation across investors between average benefit measured over 2010–2015 and average benefit measured over 2016–2021, computed separately by investment stage, success outcome, and prediction model. Investors are required to be observed in at least three distinct years in each period.
Split-half reliability	Correlation across investors between average benefit measured in 2010, 2012, and 2014 and average benefit measured in 2011, 2013, and 2015, computed separately by investment stage, success outcome, and prediction model.
Spearman-Brown reliability	Reliability correction computed as $2r/(1+r)$, where r is the split-half correlation. This statistic approximates the reliability of the full 2010–2015 six-year benefit measure implied by the two three-year measures.
Revealed exposure	Investor-level exposure to prediction-rent erosion, defined as average benefit over the 2013–2015 pre-shock window. Investors with higher revealed exposure are those whose realized portfolios differed more from ML-implied selections immediately before prediction technology became cheaper to implement.
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 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 5.3

Figure 2: Single-Feature Nonlinearity Diagnostic: Number of Previous Investors and IPO Success (Series A)

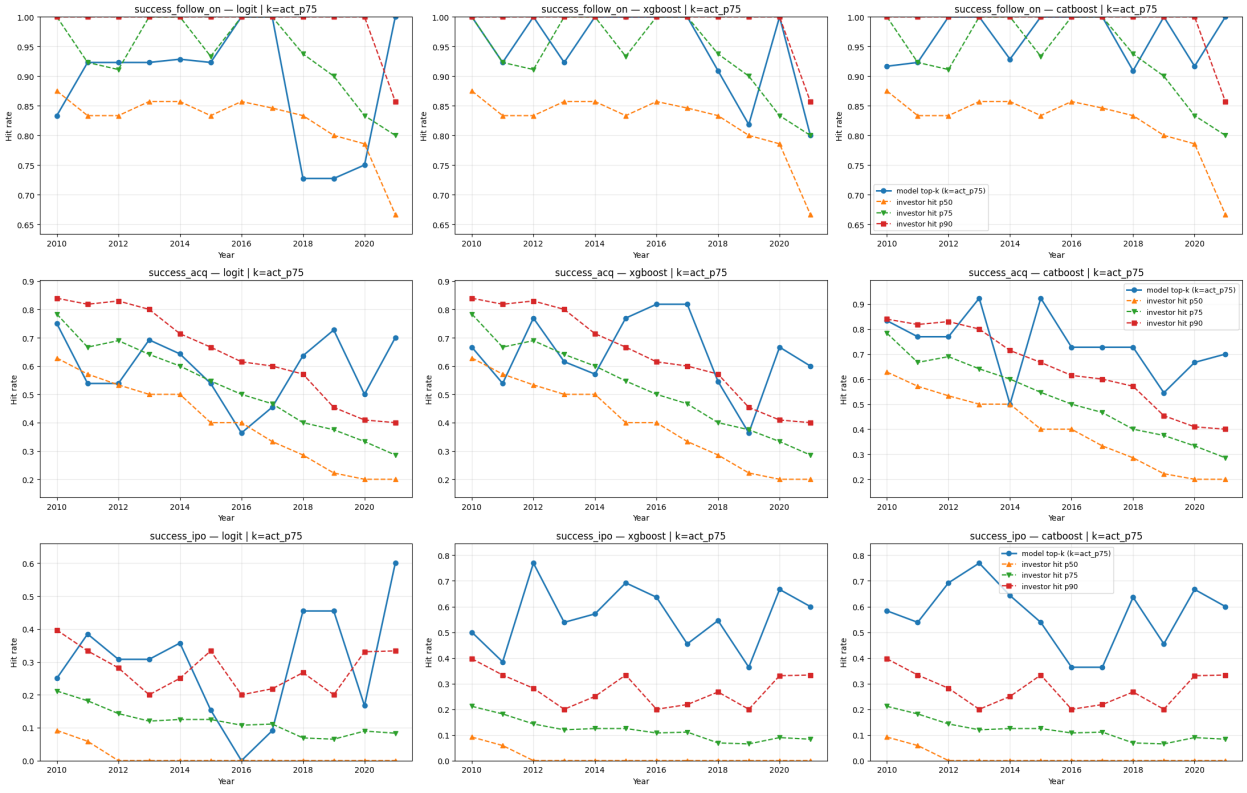
This figure illustrates a single-feature nonlinearity diagnostic for IPO outcomes using the number of previous investors in the startup at the time of a Series A financing round. The horizontal axis reports the number of prior investors, while the vertical axis indicates the realized IPO outcome. Each 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 31, 2018, and evaluated on a holdout sample dated from January 1, 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 3: Model Hit Rates at the 75th Percentile Investment Count Compared to Investor Benchmarks

This figure compares model-implied success rates to observed investor success 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). Predictions are generated within financing-round type using rolling 10-year training windows and evaluated out of sample in the subsequent year; this pooled figure then aggregates all scored rounds by calendar year. The solid blue line shows the success 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 show actual investor success rates at the 50th, 75th, and 90th percentiles. The horizontal axis reports calendar year, and the vertical axis reports success rates.



Return to Section 5.1.2

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

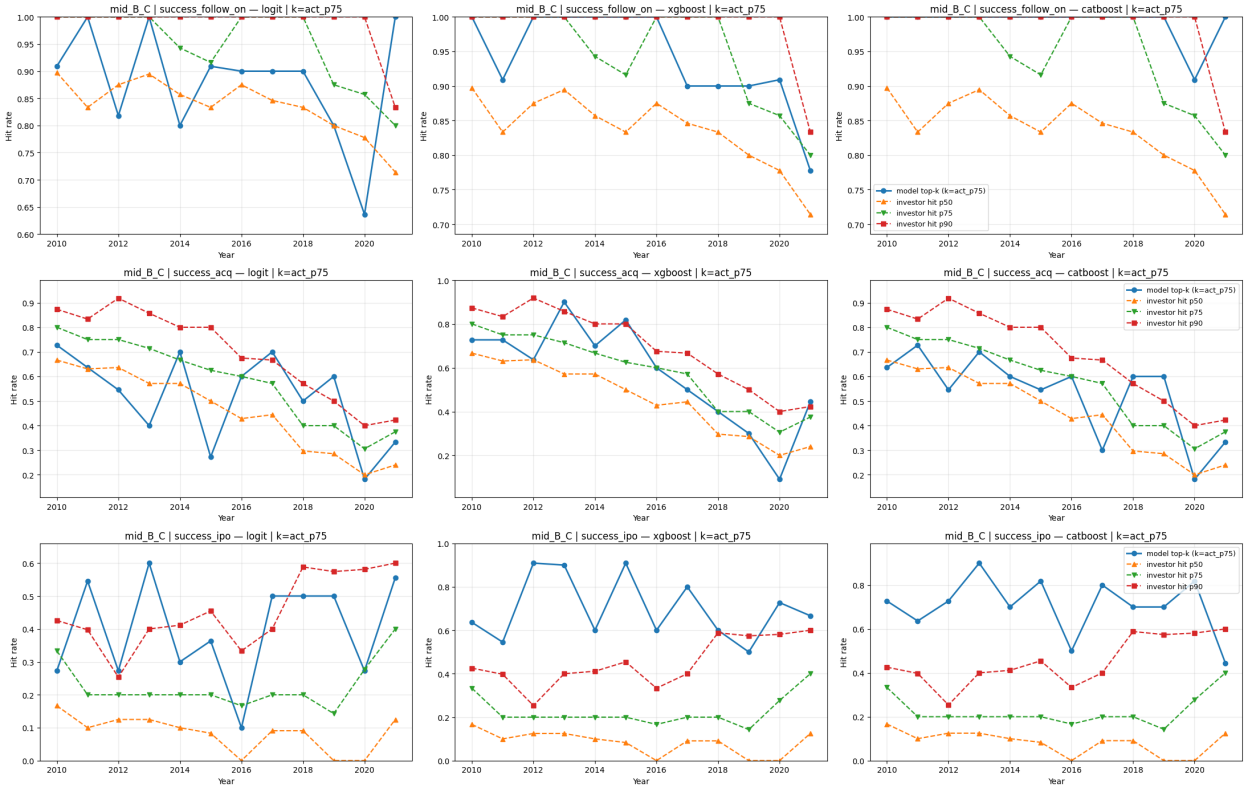
This figure compares model-implied success rates to observed investor success rates when the number of investments is fixed at the 75th percentile of investor activity, restricting the sample to early-stage (Seed and Series A) investments. Predictions are generated within financing-round type using rolling 10-year training windows and then aggregated into the early-stage group by calendar year. The solid blue line shows the success 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 show actual investor success rates at the 50th, 75th, and 90th percentiles.



Return to Section 5.1.2

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

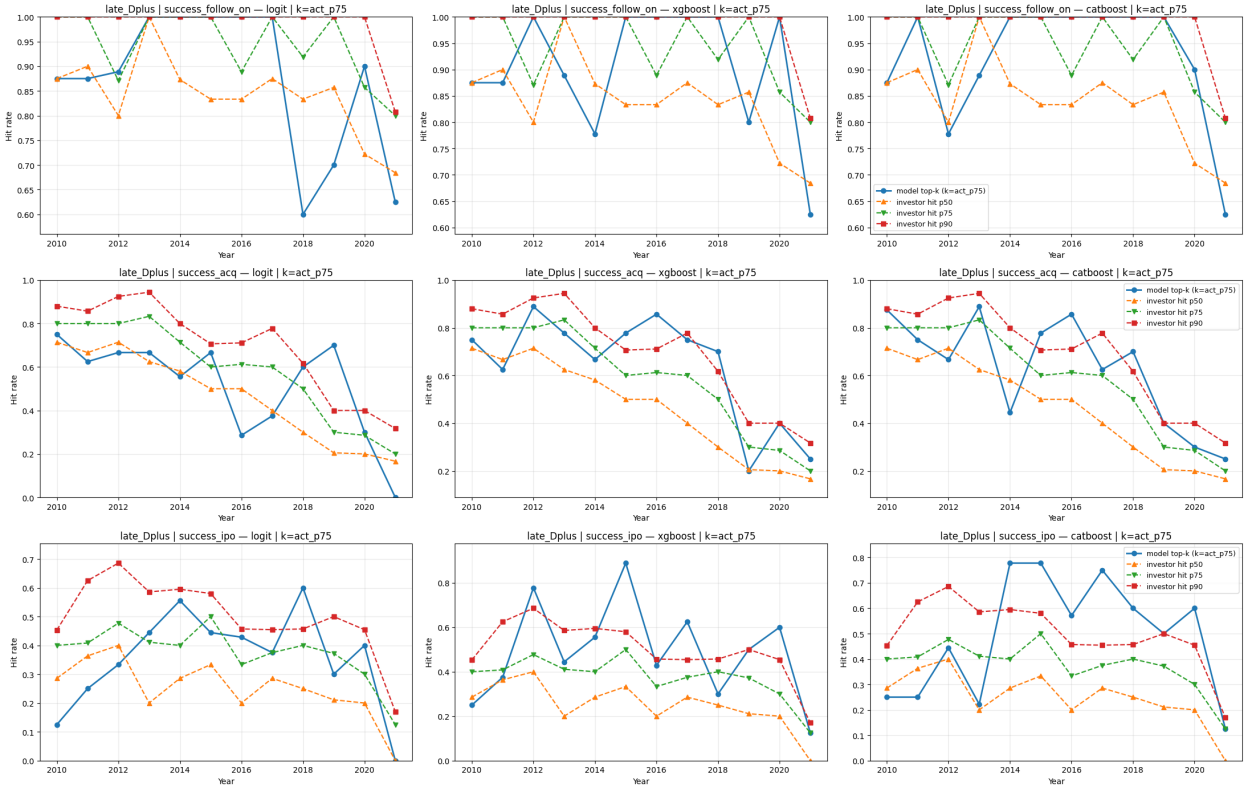
This figure compares model-implied success rates to observed investor success rates when the number of investments is fixed at the 75th percentile of investor activity, restricting the sample to mid-stage (Series B and Series C) investments. Predictions are generated within financing-round type using rolling 10-year training windows and then aggregated into the mid-stage group by calendar year. The solid blue line shows the success 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 show actual investor success rates at the 50th, 75th, and 90th percentiles.



Return to Section 5.1.2

Figure 6: Model Hit Rates at the 75th Percentile Investment Count: Late-Stage

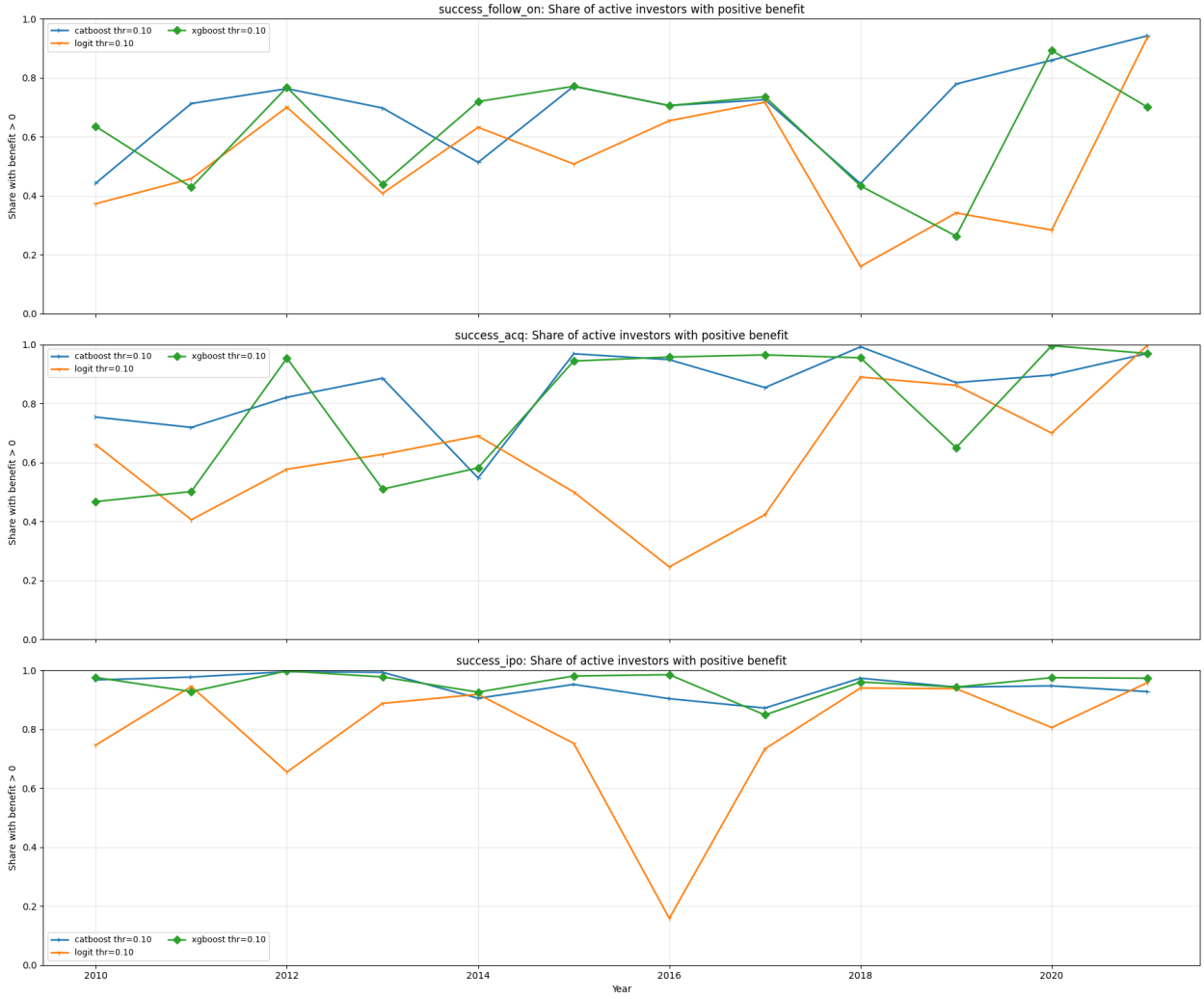
This figure compares model-implied success rates to observed investor success rates when the number of investments is fixed at the 75th percentile of investor activity, restricting the sample to late-stage (Series D onward) investments. Predictions are generated within financing-round type using rolling 10-year training windows and then aggregated into the late-stage group by calendar year. The solid blue line shows the success 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 show actual investor success rates at the 50th, 75th, and 90th percentiles.



Return to Section 5.1.2

Figure 7: Share of Active Investors with Positive Benefit Over Time

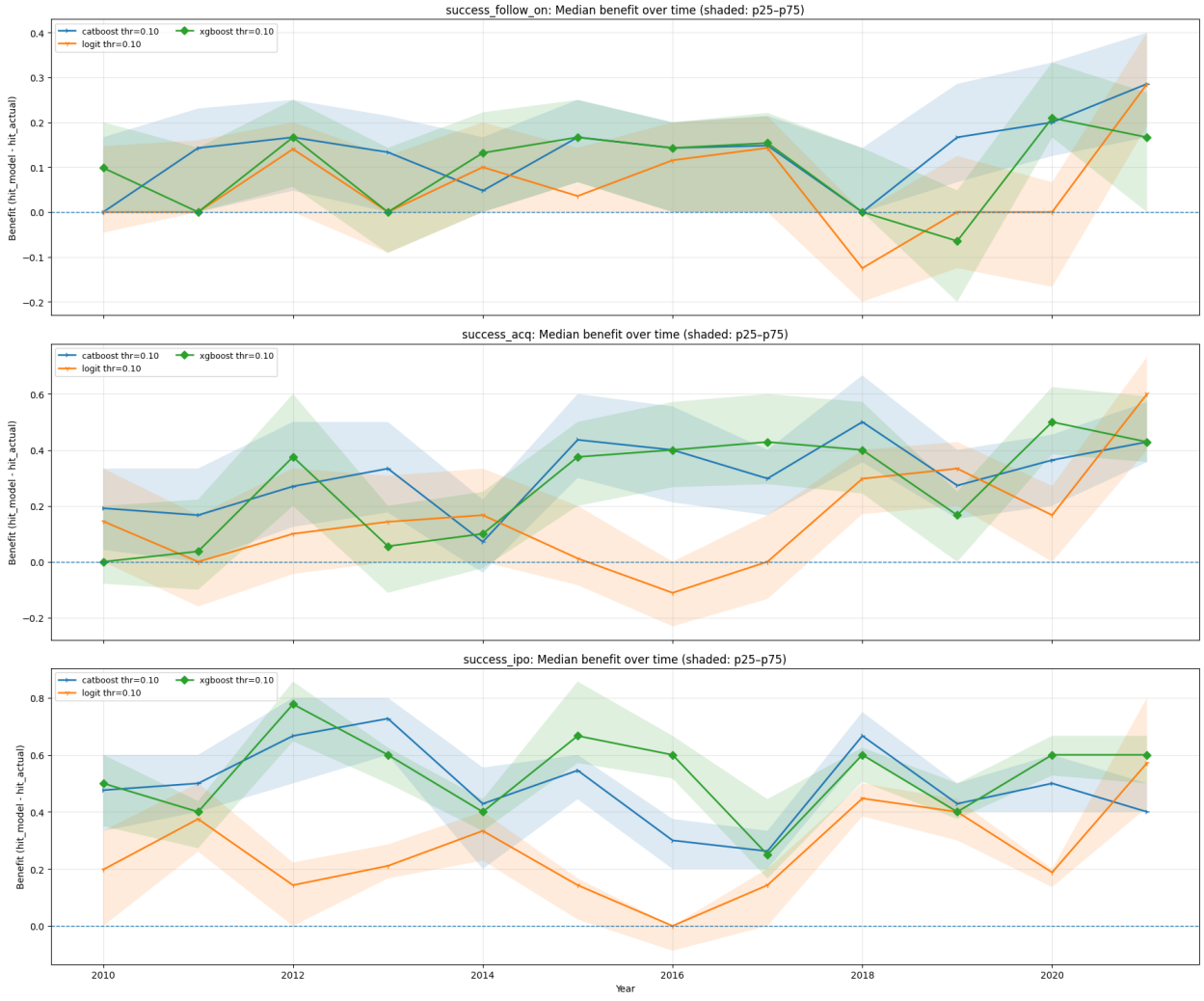
This figure plots, for each calendar year, the share of active investors whose estimated benefit is positive. Benefit is the success-rate difference between the model-implied top-k portfolio and the investor's actual portfolio, where k equals the investor's actual number of investments. Separate panels correspond to follow-on financing, acquisition, and IPO outcomes; within each panel, lines correspond to Logit, XGBoost, and CatBoost results. The figure is descriptive and shows the prevalence of positive benefit over time.



Return to Section 5.2.1

Figure 8: Median Benefit Over Time with Interquartile Range

This figure plots the evolution over time of median benefit, defined as the success-rate difference between the model-implied top-k portfolio and the investor's actual portfolio. Shaded areas represent the interquartile range of the investor-year benefit distribution. Separate panels correspond to follow-on financing, acquisition, and IPO outcomes, and lines correspond to Logit, XGBoost, and CatBoost. The figure is descriptive and shows the time-series variation in measured benefit.



Return to Section 5.2.1

Table 1: Summary Statistics for ML Features and Outcomes

This table 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 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. All variables are constructed using only information available at or before the financing round announcement date.

	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

Table 2: Benchmark Results by Stage, Outcome, and Model

This table reports the stage, outcome, and prediction-model cells underlying the benchmark analysis. The first four numeric columns report raw hit rates, in percent, for actual investor portfolios, the permuted-label random benchmark, the round-size benchmark, and the ML benchmark. The next three columns report p.p. improvements relative to the permuted-label random benchmark. The ML-size column reports the difference between ML and the round-size benchmark. Stage abbreviations are ES for early stage, MS for mid stage, LS for late stage, and AS for all stages. Model abbreviations are L for logit, XGB for XGBoost, and CB for CatBoost. The permuted-label benchmark is recomputed within each stage-outcome-model cell using the same top-k rule and model-specific ranked opportunity set; therefore, random hit rates can vary slightly across models even when actual hit rates are the same within a stage-outcome cell. ML benefit exceeds the 95th percentile of the 500-replication permuted-label benchmark in every reported row.

Stage	Outcome	Model	Actual hit	Random hit	Size hit	ML hit
ES	Follow-on	L	81.2	76	91.4	89.5
ES	Follow-on	XGB	–	76.3	–	96.4
ES	Follow-on	CB	–	76.2	–	94.2
ES	Acquisition	L	35.5	27.9	31.2	60
ES	Acquisition	XGB	–	28.1	–	61.5
ES	Acquisition	CB	–	28	–	71.7
ES	IPO	L	2.5	2.1	35.8	15.8
ES	IPO	XGB	–	2	–	42.3
ES	IPO	CB	–	2.1	–	40.9
MS	Follow-on	L	82.2	78.7	84	85.9
MS	Follow-on	XGB	–	78.4	–	94.1
MS	Follow-on	CB	–	78.5	–	98.8
MS	Acquisition	L	43	39.6	38	47.9
MS	Acquisition	XGB	–	39.1	–	51.6
MS	Acquisition	CB	–	39.2	–	49.8
MS	IPO	L	15.6	9.9	36	39
MS	IPO	XGB	–	9.9	–	68.7
MS	IPO	CB	–	10	–	69.6
LS	Follow-on	L	82.2	78.9	91.4	87.9
LS	Follow-on	XGB	–	78.8	–	89
LS	Follow-on	CB	–	79.5	–	92.4
LS	Acquisition	L	43.8	43.4	23	53.9
LS	Acquisition	XGB	–	43.8	–	64
LS	Acquisition	CB	–	43	–	59.9
LS	IPO	L	26.8	19.3	49.1	34.8
LS	IPO	XGB	–	19.5	–	52.2
LS	IPO	CB	–	19.2	–	52.2
AS	Follow-on	L	81.2	77.3	92.7	87.4
AS	Follow-on	XGB	–	77.3	–	92.9
AS	Follow-on	CB	–	77.1	–	96.6
AS	Acquisition	L	37.3	31	23.5	51.9
AS	Acquisition	XGB	–	31	–	67.6
AS	Acquisition	CB	–	30.9	–	70.1
AS	IPO	L	8.4	4.6	44.4	29.5
AS	IPO	XGB	–	4.7	–	58.6
AS	IPO	CB	–	4.6	–	53.9

Return to Section 5.1.2

Table 3: Selection Benchmark Ladder by Success Outcome

This table summarizes the detailed benchmark cells in Table 2. It compares actual investor portfolios, a round-size benchmark, and ML-ranked portfolios against a permuted-label random benchmark, separately by investment stage and success outcome. Entries are p.p. improvements in the share of successful investments. The rows labeled Actual - random, Round size - random, and ML - random report performance relative to the permuted-label benchmark; the row labeled ML - round size reports the incremental ML advantage over the round-size benchmark. Follow-on, acquisition, and IPO columns are averaged across prediction models (Logit, XGBoost, and CatBoost) within the corresponding stage and outcome. Avg., Min., and Max. summarize the three outcome-specific entries in the same row. The permuted-label benchmark randomly reassigns success labels within year, investment stage, and success outcome, preserving outcome frequency and portfolio size while breaking the link between observable characteristics and realized success. The round-size benchmark ranks startups by the amount raised in the focal financing round. The ML benchmark ranks startups by out-of-sample predicted success probabilities. In all stage-outcome-model cells summarized here, ML benefit exceeds the 95th percentile of the 500-replication random benchmark.

Stage	Comparison	Follow-on	Acquisition	IPO	Avg.	Min.	Max.
Early stage	Actual - random	5	7.5	0.5	4.3	0.5	7.5
Early stage	Round size - random	15.2	3.2	33.8	17.4	3.2	33.8
Early stage	ML - random	17.2	36.4	31	28.2	17.2	36.4
Early stage	ML - round size	2	33.2	-2.8	10.8	-2.8	33.2
Mid stage	Actual - random	3.7	3.7	5.6	4.3	3.7	5.6
Mid stage	Round size - random	5.5	-1.3	26.1	10.1	-1.3	26.1
Mid stage	ML - random	14.4	10.5	49.2	24.7	10.5	49.2
Mid stage	ML - round size	8.9	11.7	23.1	14.6	8.9	23.1
Late stage	Actual - random	3.1	0.4	7.5	3.7	0.4	7.5
Late stage	Round size - random	12.3	-20.4	29.7	7.2	-20.4	29.7
Late stage	ML - random	10.7	15.9	27.1	17.9	10.7	27.1
Late stage	ML - round size	-1.6	36.3	-2.7	10.7	-2.7	36.3
All stages	Actual - random	4	6.4	3.8	4.7	3.8	6.4
All stages	Round size - random	15.5	-7.5	39.8	15.9	-7.5	39.8
All stages	ML - random	15.1	32.2	42.7	30	15.1	42.7
All stages	ML - round size	-0.4	39.7	2.9	14.1	-0.4	39.7

Return to Section 5.1.2

Table 4: 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 Appendix D.1, Appendix D.2, and Appendix D.3 for detailed definitions of the 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}	0.07	0.18	0.00	0.05	0.17	0.12	0.17	0.00	0.12	0.20	0.16	0.16	0.00	0.14	0.23	4,937
$\text{HitRate}_{it}^{\text{Actual}}$	0.81	0.15	0.73	0.83	0.91	0.81	0.15	0.73	0.83	0.91	0.81	0.15	0.73	0.83	0.91	4,937
$\text{HitRate}_{it}^{\text{Model}}$	0.88	0.12	0.80	0.90	1.00	0.93	0.10	0.88	1.00	1.00	0.97	0.05	0.94	1.00	1.00	4,937
$\text{Number of Investments}_{it}$	13.11	13.46	6.00	9.00	14.00	13.11	13.46	6.00	9.00	14.00	13.11	13.46	6.00	9.00	14.00	4,937
Distance_{it}	0.02	0.66	-0.44	-0.06	0.44	0.02	0.66	-0.44	-0.06	0.44	0.02	0.66	-0.44	-0.06	0.44	4,918
$\text{Distance}_{it}^{\text{Hotness}}$	-2.62	2.51	-3.74	-2.28	-1.12	-2.64	2.51	-3.76	-2.31	-1.15	-2.64	2.51	-3.76	-2.31	-1.15	4,918
$\text{Distance}_{it}^{\text{Sector}}$	0.35	0.16	0.23	0.33	0.46	0.35	0.16	0.23	0.33	0.46	0.35	0.16	0.23	0.33	0.46	4,918
$\text{Distance}_{it}^{\text{Geography}}$	0.52	0.17	0.40	0.47	0.57	0.52	0.17	0.40	0.47	0.57	0.52	0.17	0.40	0.47	0.57	4,918
Acquisition																
Benefit_{it}	0.14	0.26	0.00	0.14	0.32	0.31	0.25	0.14	0.31	0.50	0.32	0.23	0.17	0.33	0.50	4,937
$\text{HitRate}_{it}^{\text{Actual}}$	0.36	0.21	0.20	0.33	0.50	0.36	0.21	0.20	0.33	0.50	0.36	0.21	0.20	0.33	0.50	4,937
$\text{HitRate}_{it}^{\text{Model}}$	0.51	0.16	0.40	0.52	0.62	0.67	0.16	0.58	0.69	0.80	0.69	0.15	0.57	0.67	0.81	4,937
$\text{Number of Investments}_{it}$	13.11	13.46	6.00	9.00	14.00	13.11	13.46	6.00	9.00	14.00	13.11	13.46	6.00	9.00	14.00	4,937
Distance_{it}	0.02	0.66	-0.45	-0.06	0.43	0.02	0.66	-0.45	-0.06	0.44	0.02	0.66	-0.45	-0.06	0.44	4,918
$\text{Distance}_{it}^{\text{Hotness}}$	-2.60	2.51	-3.72	-2.27	-1.11	-2.58	2.51	-3.70	-2.25	-1.08	-2.59	2.51	-3.71	-2.27	-1.09	4,918
$\text{Distance}_{it}^{\text{Sector}}$	0.35	0.17	0.23	0.32	0.46	0.35	0.16	0.23	0.32	0.46	0.35	0.16	0.23	0.32	0.46	4,918
$\text{Distance}_{it}^{\text{Geography}}$	0.51	0.17	0.40	0.47	0.57	0.51	0.17	0.40	0.47	0.57	0.51	0.17	0.39	0.47	0.56	4,918
IPO																
Benefit_{it}	0.21	0.23	0.08	0.20	0.38	0.50	0.21	0.40	0.54	0.62	0.45	0.22	0.33	0.48	0.60	4,937
$\text{HitRate}_{it}^{\text{Actual}}$	0.08	0.15	0.00	0.00	0.11	0.08	0.15	0.00	0.00	0.11	0.08	0.15	0.00	0.00	0.11	4,937
$\text{HitRate}_{it}^{\text{Model}}$	0.29	0.17	0.17	0.28	0.43	0.58	0.15	0.50	0.60	0.67	0.53	0.16	0.40	0.55	0.65	4,937
$\text{Number of Investments}_{it}$	13.11	13.46	6.00	9.00	14.00	13.11	13.46	6.00	9.00	14.00	13.11	13.46	6.00	9.00	14.00	4,937
Distance_{it}	0.01	0.64	-0.43	-0.06	0.43	-0.00	0.64	-0.44	-0.05	0.41	0.01	0.65	-0.43	-0.04	0.42	4,918
$\text{Distance}_{it}^{\text{Hotness}}$	-2.12	2.51	-3.21	-1.78	-0.63	-1.77	2.51	-2.88	-1.47	-0.29	-2.03	2.50	-3.13	-1.70	-0.55	4,918
$\text{Distance}_{it}^{\text{Sector}}$	0.40	0.14	0.31	0.39	0.48	0.45	0.17	0.33	0.45	0.57	0.40	0.15	0.30	0.39	0.49	4,918
$\text{Distance}_{it}^{\text{Geography}}$	0.49	0.19	0.36	0.44	0.56	0.49	0.18	0.36	0.44	0.57	0.49	0.18	0.37	0.45	0.56	4,918
Panel B: Early Stage (Seed–A)																
Follow-on																
Benefit_{it}	0.09	0.18	0.00	0.07	0.19	0.16	0.16	0.00	0.14	0.25	0.13	0.17	0.00	0.12	0.21	2,835
$\text{HitRate}_{it}^{\text{Actual}}$	0.81	0.15	0.73	0.83	0.92	0.81	0.15	0.73	0.83	0.92	0.81	0.15	0.73	0.83	0.92	2,835
$\text{HitRate}_{it}^{\text{Model}}$	0.90	0.07	0.83	0.89	1.00	0.97	0.06	0.94	1.00	1.00	0.94	0.07	0.90	1.00	1.00	2,835
$\text{Number of Investments}_{it}$	11.97	12.66	6.00	8.00	13.00	11.97	12.66	6.00	8.00	13.00	11.97	12.66	6.00	8.00	13.00	2,835
Distance_{it}	-0.01	0.71	-0.50	-0.09	0.42	-0.01	0.71	-0.50	-0.10	0.43	-0.01	0.71	-0.50	-0.10	0.43	2,825
$\text{Distance}_{it}^{\text{Hotness}}$	-2.90	2.94	-4.28	-2.51	-1.04	-2.93	2.94	-4.32	-2.54	-1.05	-2.93	2.94	-4.32	-2.54	-1.06	2,825
$\text{Distance}_{it}^{\text{Sector}}$	0.34	0.15	0.23	0.31	0.43	0.34	0.15	0.23	0.31	0.43	0.34	0.15	0.23	0.31	0.43	2,825
$\text{Distance}_{it}^{\text{Geography}}$	0.54	0.18	0.41	0.48	0.60	0.54	0.18	0.41	0.49	0.60	0.54	0.18	0.41	0.49	0.60	2,825
Acquisition																
Benefit_{it}	0.25	0.28	0.00	0.23	0.43	0.27	0.25	0.11	0.26	0.43	0.36	0.27	0.17	0.36	0.57	2,835
$\text{HitRate}_{it}^{\text{Actual}}$	0.35	0.21	0.20	0.33	0.50	0.35	0.21	0.20	0.33	0.50	0.35	0.21	0.20	0.33	0.50	2,835
$\text{HitRate}_{it}^{\text{Model}}$	0.59	0.14	0.50	0.57	0.67	0.61	0.14	0.50	0.60	0.71	0.70	0.18	0.62	0.71	0.83	2,835
$\text{Number of Investments}_{it}$	11.97	12.66	6.00	8.00	13.00	11.97	12.66	6.00	8.00	13.00	11.97	12.66	6.00	8.00	13.00	2,835
Distance_{it}	-0.01	0.71	-0.50	-0.09	0.43	-0.01	0.71	-0.50	-0.10	0.42	-0.01	0.71	-0.50	-0.10	0.42	2,825
$\text{Distance}_{it}^{\text{Hotness}}$	-2.85	2.94	-4.23	-2.45	-0.98	-2.85	2.94	-4.23	-2.46	-0.96	-2.87	2.94	-4.24	-2.47	-0.98	2,825
$\text{Distance}_{it}^{\text{Sector}}$	0.34	0.16	0.22	0.30	0.43	0.34	0.16	0.22	0.31	0.43	0.34	0.16	0.22	0.31	0.43	2,825
$\text{Distance}_{it}^{\text{Geography}}$	0.54	0.18	0.41	0.48	0.60	0.53	0.18	0.41	0.48	0.60	0.54	0.18	0.41	0.48	0.59	2,825
IPO																
Benefit_{it}	0.14	0.17	0.00	0.12	0.22	0.40	0.19	0.29	0.40	0.52	0.38	0.15	0.29	0.38	0.50	2,835
$\text{HitRate}_{it}^{\text{Actual}}$	0.02	0.07	0.00	0.00	0.00	0.02	0.07	0.00	0.00	0.00	0.02	0.07	0.00	0.00	0.00	2,835
$\text{HitRate}_{it}^{\text{Model}}$	0.16	0.15	0.00	0.15	0.25	0.43	0.18	0.31	0.40	0.55	0.41	0.13	0.33	0.40	0.50	2,835
$\text{Number of Investments}_{it}$	11.97	12.66	6.00	8.00	13.00	11.97	12.66	6.00	8.00	13.00	11.97	12.66	6.00	8.00	13.00	2,835
Distance_{it}	0.01	0.70	-0.49	-0.07	0.46	0.05	0.66	-0.40	-0.01	0.47	0.03	0.69	-0.47	-0.04	0.47	2,825
$\text{Distance}_{it}^{\text{Hotness}}$	-2.43	2.94	-3.83	-2.04	-0.57	-2.37	2.93	-3.76	-2.05	-0.57	-2.49	2.93	-3.86	-2.16	-0.67	2,825
$\text{Distance}_{it}^{\text{Sector}}$	0.40	0.14	0.30	0.39	0.48	0.50	0.18	0.37	0.48	0.61	0.41	0.15	0.31	0.40	0.51	2,825
$\text{Distance}_{it}^{\text{Geography}}$	0.51	0.20	0.36	0.44	0.59	0.51	0.20	0.36	0.46	0.60	0.52	0.18	0.39	0.47	0.61	2,825

Return to Section 5.2.1

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 Appendix D.1, Appendix D.2, and Appendix D.3 for detailed definitions of the 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}	0.04	0.18	-0.06	0.00	0.17	0.12	0.15	0.00	0.11	0.20	0.17	0.15	0.05	0.15	0.25	1,074
$\text{HitRate}_{it}^{\text{Actual}}$	0.82	0.14	0.73	0.83	0.91	0.82	0.14	0.73	0.83	0.91	0.82	0.14	0.73	0.83	0.91	1,074
$\text{HitRate}_{it}^{\text{Model}}$	0.86	0.14	0.82	0.88	1.00	0.94	0.08	0.88	1.00	1.00	0.98	0.04	1.00	1.00	1.00	1,074
$\text{Number of Investments}_{it}$	9.94	5.89	6.00	8.00	12.00	9.94	5.89	6.00	8.00	12.00	9.94	5.89	6.00	8.00	12.00	1,074
Distance_{it}	0.00	0.47	-0.33	-0.04	0.29	-0.00	0.48	-0.33	-0.04	0.30	-0.00	0.48	-0.33	-0.04	0.30	1,067
$\text{Distance}_{it}^{\text{Hotness}}$	-1.77	1.78	-2.56	-1.47	-0.61	-1.79	1.77	-2.57	-1.52	-0.63	-1.79	1.77	-2.56	-1.52	-0.63	1,067
$\text{Distance}_{it}^{\text{Sector}}$	0.35	0.13	0.26	0.34	0.43	0.35	0.13	0.26	0.34	0.43	0.35	0.13	0.26	0.34	0.43	1,067
$\text{Distance}_{it}^{\text{Geography}}$	0.45	0.13	0.36	0.43	0.51	0.45	0.13	0.37	0.43	0.52	0.45	0.13	0.37	0.43	0.52	1,067
Acquisition																
Benefit_{it}	0.05	0.28	-0.12	0.00	0.22	0.08	0.23	-0.07	0.09	0.22	0.05	0.24	-0.12	0.00	0.20	1,074
$\text{HitRate}_{it}^{\text{Actual}}$	0.43	0.23	0.25	0.40	0.60	0.43	0.23	0.25	0.40	0.60	0.43	0.23	0.25	0.40	0.60	1,074
$\text{HitRate}_{it}^{\text{Model}}$	0.48	0.20	0.30	0.54	0.62	0.51	0.23	0.33	0.50	0.71	0.48	0.19	0.33	0.54	0.60	1,074
$\text{Number of Investments}_{it}$	9.94	5.89	6.00	8.00	12.00	9.94	5.89	6.00	8.00	12.00	9.94	5.89	6.00	8.00	12.00	1,074
Distance_{it}	-0.00	0.47	-0.33	-0.05	0.30	0.00	0.48	-0.33	-0.04	0.30	-0.00	0.47	-0.33	-0.05	0.29	1,067
$\text{Distance}_{it}^{\text{Hotness}}$	-1.78	1.76	-2.57	-1.50	-0.63	-1.74	1.75	-2.58	-1.47	-0.62	-1.78	1.76	-2.56	-1.51	-0.62	1,067
$\text{Distance}_{it}^{\text{Sector}}$	0.35	0.14	0.25	0.33	0.44	0.35	0.13	0.25	0.33	0.43	0.35	0.13	0.26	0.33	0.43	1,067
$\text{Distance}_{it}^{\text{Geography}}$	0.45	0.13	0.37	0.43	0.52	0.46	0.13	0.37	0.43	0.52	0.45	0.13	0.37	0.43	0.52	1,067
IPO																
Benefit_{it}	0.24	0.28	0.01	0.30	0.45	0.54	0.25	0.40	0.57	0.71	0.53	0.25	0.40	0.58	0.70	1,074
$\text{HitRate}_{it}^{\text{Actual}}$	0.14	0.20	0.00	0.08	0.20	0.14	0.20	0.00	0.08	0.20	0.14	0.20	0.00	0.08	0.20	1,074
$\text{HitRate}_{it}^{\text{Model}}$	0.39	0.19	0.30	0.43	0.54	0.69	0.16	0.57	0.67	0.82	0.67	0.16	0.58	0.68	0.80	1,074
$\text{Number of Investments}_{it}$	9.94	5.89	6.00	8.00	12.00	9.94	5.89	6.00	8.00	12.00	9.94	5.89	6.00	8.00	12.00	1,074
Distance_{it}	-0.01	0.56	-0.36	-0.01	0.35	-0.02	0.61	-0.41	0.02	0.41	-0.02	0.56	-0.38	-0.02	0.35	1,067
$\text{Distance}_{it}^{\text{Hotness}}$	-1.50	1.74	-2.29	-1.20	-0.36	-1.06	1.72	-1.94	-0.81	0.03	-1.32	1.72	-2.14	-1.04	-0.19	1,067
$\text{Distance}_{it}^{\text{Sector}}$	0.39	0.15	0.29	0.39	0.49	0.44	0.20	0.29	0.45	0.58	0.39	0.15	0.28	0.38	0.49	1,067
$\text{Distance}_{it}^{\text{Geography}}$	0.44	0.14	0.36	0.41	0.50	0.44	0.14	0.35	0.42	0.51	0.45	0.13	0.36	0.42	0.51	1,067
Panel D: Late Stage (D+)																
Follow-on																
Benefit_{it}	0.06	0.19	0.00	0.00	0.20	0.08	0.20	0.00	0.07	0.20	0.11	0.19	0.00	0.11	0.20	327
$\text{HitRate}_{it}^{\text{Actual}}$	0.82	0.15	0.71	0.83	1.00	0.82	0.15	0.71	0.83	1.00	0.82	0.15	0.71	0.83	1.00	327
$\text{HitRate}_{it}^{\text{Model}}$	0.88	0.14	0.80	0.90	1.00	0.90	0.13	0.82	0.94	1.00	0.93	0.14	0.90	1.00	1.00	327
$\text{Number of Investments}_{it}$	8.38	3.65	6.00	7.00	10.00	8.38	3.65	6.00	7.00	10.00	8.38	3.65	6.00	7.00	10.00	327
Distance_{it}	-0.22	0.40	-0.53	-0.24	-0.02	-0.23	0.41	-0.53	-0.25	-0.01	-0.23	0.41	-0.53	-0.25	-0.01	325
$\text{Distance}_{it}^{\text{Hotness}}$	-1.81	1.30	-2.49	-1.68	-0.99	-1.88	1.30	-2.54	-1.72	-1.07	-1.87	1.30	-2.53	-1.72	-1.07	325
$\text{Distance}_{it}^{\text{Sector}}$	0.32	0.12	0.23	0.29	0.38	0.31	0.12	0.23	0.29	0.38	0.31	0.12	0.23	0.29	0.38	325
$\text{Distance}_{it}^{\text{Geography}}$	0.38	0.11	0.29	0.38	0.44	0.38	0.11	0.30	0.38	0.45	0.38	0.11	0.30	0.38	0.45	325
Acquisition																
Benefit_{it}	0.09	0.26	-0.10	0.08	0.29	0.20	0.23	0.00	0.20	0.38	0.17	0.25	0.00	0.17	0.33	327
$\text{HitRate}_{it}^{\text{Actual}}$	0.42	0.24	0.20	0.40	0.60	0.42	0.24	0.20	0.40	0.60	0.42	0.24	0.20	0.40	0.60	327
$\text{HitRate}_{it}^{\text{Model}}$	0.50	0.25	0.30	0.57	0.70	0.62	0.22	0.40	0.67	0.80	0.59	0.25	0.38	0.60	0.80	327
$\text{Number of Investments}_{it}$	8.38	3.65	6.00	7.00	10.00	8.38	3.65	6.00	7.00	10.00	8.38	3.65	6.00	7.00	10.00	327
Distance_{it}	-0.20	0.41	-0.52	-0.24	0.05	-0.23	0.41	-0.53	-0.27	-0.02	-0.23	0.41	-0.53	-0.26	-0.01	325
$\text{Distance}_{it}^{\text{Hotness}}$	-2.02	1.34	-2.81	-1.82	-1.13	-1.89	1.30	-2.53	-1.72	-1.07	-1.87	1.30	-2.54	-1.72	-1.06	325
$\text{Distance}_{it}^{\text{Sector}}$	0.31	0.13	0.21	0.28	0.38	0.31	0.12	0.22	0.28	0.37	0.31	0.12	0.23	0.29	0.38	325
$\text{Distance}_{it}^{\text{Geography}}$	0.41	0.10	0.34	0.41	0.48	0.39	0.11	0.30	0.38	0.45	0.38	0.11	0.30	0.38	0.45	325
IPO																
Benefit_{it}	0.12	0.22	0.00	0.14	0.25	0.27	0.23	0.12	0.29	0.43	0.29	0.25	0.16	0.33	0.44	327
$\text{HitRate}_{it}^{\text{Actual}}$	0.25	0.19	0.12	0.22	0.39	0.25	0.19	0.12	0.22	0.39	0.25	0.19	0.12	0.22	0.39	327
$\text{HitRate}_{it}^{\text{Model}}$	0.36	0.14	0.33	0.39	0.43	0.52	0.19	0.40	0.50	0.62	0.54	0.20	0.43	0.58	0.67	327
$\text{Number of Investments}_{it}$	8.38	3.65	6.00	7.00	10.00	8.38	3.65	6.00	7.00	10.00	8.38	3.65	6.00	7.00	10.00	327
Distance_{it}	-0.20	0.42	-0.49	-0.23	0.04	-0.36	0.42	-0.67	-0.38	-0.07	-0.32	0.43	-0.63	-0.34	-0.05	325
$\text{Distance}_{it}^{\text{Hotness}}$	-1.63	1.32	-2.31	-1.48	-0.79	-1.54	1.30	-2.20	-1.36	-0.74	-1.64	1.30	-2.30	-1.49	-0.82	325
$\text{Distance}_{it}^{\text{Sector}}$	0.38	0.12	0.29	0.36	0.45	0.36	0.12	0.27	0.34	0.44	0.34	0.12	0.24	0.31	0.41	325
$\text{Distance}_{it}^{\text{Geography}}$	0.37	0.11	0.29	0.37	0.43	0.39	0.12	0.30	0.38	0.47	0.38	0.11	0.30	0.37	0.45	325

Return to Section 5.2.1

Table 5: Investor-Specific Persistence and Measurement Reliability

This table reports persistence and reliability diagnostics for investor-year benefit. Benefit is the success-rate difference between the ML-implied top-k portfolio and the investor's actual portfolio. For each stage, outcome, and prediction-model combination, the Pearson and Spearman columns correlate investors' average benefit in 2010–2015 with their average benefit in 2016–2021. Investors are included only if benefit is observed in at least three distinct years in each period; reported entries are averages across the nine outcome-model combinations within each stage. Late-stage results are omitted because the strict matched-investor sample is too small for the same reliability summary. The split-half column is computed within 2010–2015 by correlating average benefit in 2010, 2012, and 2014 with average benefit in 2011, 2013, and 2015. Spearman–Brown reliability applies the correction $2r/(1+r)$ and approximates the reliability of the full 2010–2015 six-year benefit measure. The final two columns report the share of investor-year benefit variance explained jointly by investor and year fixed effects and the remaining residual share.

Stage	Pearson persist.	Spearman persist.	Split-half	SB reliab.	Inv.+year R^2	Residual
Early stage	0.155	0.085	0.271	0.417	0.604	0.396
Mid stage	0.231	0.162	0.225	0.325	0.574	0.426
All stages	0.419	0.334	0.345	0.503	0.615	0.385

Return to Section 5.2.2

Table 6: Diagnostic Checks for Persistence Estimates

Panel A uses the same persistence measures as Table 5: earlier-period benefit is average benefit over 2010–2015 and later-period benefit is average benefit over 2016–2021. It reports the arithmetic identity that the slope from regressing later-period minus earlier-period benefit on earlier-period benefit equals the slope from regressing later-period benefit on earlier-period benefit minus one. **Panel B** reports a separate annual diagnostic using Arellano–Bond dynamic-panel estimates of the autoregressive coefficient in investor-year benefit, with explicit year controls and robust standard errors clustered by investor.

Panel A: Change-Slope Identity

Stage	Mean level slope	Mean change slope	Max identity error
All stages	0.364	-0.636	5.96e-08
Early stage	0.143	-0.857	0
Mid stage	0.258	-0.742	2.98e-08

Panel B: Dynamic-Panel Persistence

Stage	Mean ρ	Median ρ	Mean 95% CI	Specs
All stages	0.066	0.064	[0.006, 0.127]	9
Early stage	0.044	0.041	[-0.030, 0.117]	9
Mid stage	-0.013	0.001	[-0.118, 0.092]	9

Return to Section 5.2.2

Table 7: Observable Portfolio Distance and Predictive Benefit

This table summarizes regressions relating investor-year benefit to observable distance between actual and ML-implied portfolios. The composite specification uses a standardized distance index combining measures of market hotness, sector composition, and geography. The decomposed specification includes the three components separately. In the column headings, Comp. denotes the composite specification, Decomp. denotes the decomposed specification, and sig. denotes the number of outcome-model combinations, out of nine, in which the corresponding distance measure is statistically significant at conventional levels. Entries report within R^2 values and significance counts. All regressions are estimated at the investor-year level with investor and year fixed effects, and standard errors are clustered at the investor level. Table 8 reports the underlying coefficients.

Stage	Comp. R^2	Comp. sig.	Decomp. R^2	Hotness sig.	Sector sig.	Geo. sig.
All stages	0.001	4/9	0.003	1/9	4/9	3/9
Early stage	0.003	3/9	0.006	1/9	4/9	2/9
Mid stage	0.005	4/9	0.016	5/9	2/9	2/9

Return to Section 5.2.3

Table 8: Distance Regressions by Stage, Outcome, and Model

This table reports the stage, outcome, and prediction-model cells underlying Table 7. The composite column reports the coefficient on the standardized portfolio-distance index. The decomposed columns report coefficients on measures of market hotness, sector distance, and geographic distance. Regressions are estimated at the investor-year level with investor and year fixed effects; standard errors are clustered at the investor level. Stars denote significance at the 10%, 5%, and 1% levels.

Stage	Outcome	Model	Comp.	Comp. R^2	Hotness	Sector	Geo.	Dec. R^2
All stages	Follow-on	Logit	-0.002	0.000	0.001	-0.008**	0.008	0.002
All stages	Follow-on	XGBoost	-0.014**	0.001	0.001	-0.013***	0.001	0.004
All stages	Follow-on	CatBoost	0.005	0.000	0.000	-0.003	0.011**	0.002
All stages	Acquisition	Logit	-0.012	0.001	-0.001	-0.009*	-0.001	0.001
All stages	Acquisition	XGBoost	0.027***	0.003	0.001	0.010**	0.018***	0.005
All stages	Acquisition	CatBoost	0.017**	0.001	-0.001	0.008*	0.010*	0.002
All stages	IPO	Logit	0.004	0.000	0.005*	-0.000	-0.001	0.001
All stages	IPO	XGBoost	0.008	0.001	0.008**	0.004	-0.006	0.003
All stages	IPO	CatBoost	-0.018***	0.003	0.005	-0.010***	-0.014***	0.008
Early stage	Follow-on	Logit	0.005	0.000	0.009**	-0.005	0.002	0.003
Early stage	Follow-on	XGBoost	0.003	0.000	0.007	-0.004	0.001	0.002
Early stage	Follow-on	CatBoost	-0.001	0.000	0.006	-0.005	-0.003	0.002
Early stage	Acquisition	Logit	0.038***	0.007	-0.005	0.027***	0.017**	0.016
Early stage	Acquisition	XGBoost	0.017	0.001	-0.002	0.014**	0.003	0.003
Early stage	Acquisition	CatBoost	0.039***	0.007	-0.002	0.027***	0.011	0.014
Early stage	IPO	Logit	-0.005	0.001	-0.005	-0.003	0.006	0.003
Early stage	IPO	XGBoost	-0.001	0.000	0.003	-0.000	-0.005	0.001
Early stage	IPO	CatBoost	-0.018***	0.006	-0.001	-0.008***	-0.010**	0.009
Mid stage	Follow-on	Logit	0.004	0.000	0.010	0.008	-0.011	0.004
Mid stage	Follow-on	XGBoost	0.011	0.001	0.008	0.010	-0.006	0.003
Mid stage	Follow-on	CatBoost	0.021	0.003	0.001	0.019***	-0.000	0.008
Mid stage	Acquisition	Logit	0.044**	0.006	0.024*	0.012	0.011	0.007
Mid stage	Acquisition	XGBoost	0.045**	0.008	0.024**	0.004	0.019	0.010
Mid stage	Acquisition	CatBoost	0.029	0.003	0.032***	-0.005	0.008	0.010
Mid stage	IPO	Logit	0.002	0.000	0.034***	0.001	-0.031***	0.028
Mid stage	IPO	XGBoost	0.049***	0.021	0.039***	0.030***	-0.019**	0.057
Mid stage	IPO	CatBoost	0.029**	0.007	0.034***	0.011	-0.012	0.021

Return to Section 5.2

Table 9: Accelerated Compression Around the ML Implementation-Cost Shock

This table compares actual shock-period compression with chronological placebo compression. The dependent variable is investor-year benefit. Exposure is predetermined pre-shock benefit. The actual design measures exposure over 2013–2015 and post-shock compression over 2016–2017. The placebo design measures exposure over 2010–2012 and placebo compression over 2013–2014, before the main ML infrastructure shock. Entries report average coefficients across the nine outcome-model cells, the number of cells significant at conventional levels, the average actual-minus-placebo difference, the number of cells in which that difference is statistically significant at conventional levels, and the number of cells in which the actual coefficient is more negative than the placebo coefficient. Regressions include investor and year fixed effects, and standard errors are clustered at the investor level. Table 10 reports the underlying actual, year-specific, placebo, and actual-minus-placebo estimates.

Stage	Actual	Actual sig.	Placebo	Placebo sig.	Actual-placebo	Diff. sig.	More neg.
All stages	-0.041	9/9	-0.028	3/9	-0.015	1/9	8/9
Early stage	-0.044	8/9	-0.016	3/9	-0.031	4/9	6/9
Mid stage	-0.048	7/9	-0.028	1/9	-0.023	2/9	8/9

Return to Section 5.3.2

Table 10: Accelerated Compression Estimates by Stage, Outcome, and Model

This table reports the stage, outcome, and prediction-model cells underlying Table 9. Actual is the coefficient on pre-shock exposure, measured as average benefit over 2013–2015, interacted with the 2016–2017 post-shock indicator. The 2016 and 2017 columns report year-specific post-shock estimates. Placebo uses exposure measured over 2010–2012 and a placebo post period of 2013–2014. Actual-placebo reports the difference between the actual and placebo coefficients. Regressions include investor and year fixed effects; standard errors are clustered at the investor level. Stars denote significance at the 10%, 5%, and 1% levels.

Stage	Outcome	Model	Actual	2016	2017	Placebo	Actual-placebo
All stages	Follow-on	Logit	-0.043***	-0.050***	-0.033***	-0.007	-0.038*
All stages	Follow-on	XGBoost	-0.034***	-0.039**	-0.026**	-0.013	-0.020
All stages	Follow-on	CatBoost	-0.039***	-0.046***	-0.028**	-0.026	-0.014
All stages	Acquisition	Logit	-0.040***	-0.055***	-0.018	-0.026*	-0.018
All stages	Acquisition	XGBoost	-0.031***	-0.027*	-0.036**	-0.047***	0.019
All stages	Acquisition	CatBoost	-0.066***	-0.049***	-0.089***	-0.045***	-0.032**
All stages	IPO	Logit	-0.036***	-0.054***	-0.014	-0.031*	-0.006
All stages	IPO	XGBoost	-0.051***	-0.072***	-0.025	-0.024*	-0.026
All stages	IPO	CatBoost	-0.031***	-0.048***	-0.011	-0.031**	-0.001
Early stage	Follow-on	Logit	-0.045***	-0.058**	-0.024	0.007	-0.053*
Early stage	Follow-on	XGBoost	-0.050***	-0.064***	-0.026	0.021	-0.070**
Early stage	Follow-on	CatBoost	-0.060***	-0.071***	-0.042**	0.012	-0.075***
Early stage	Acquisition	Logit	-0.037**	-0.046**	-0.025	-0.058***	0.016
Early stage	Acquisition	XGBoost	-0.052***	-0.064***	-0.034	-0.049***	-0.021
Early stage	Acquisition	CatBoost	-0.037**	-0.062***	-0.003	-0.058***	0.019
Early stage	IPO	Logit	-0.012	-0.007	-0.018*	-0.031*	0.021
Early stage	IPO	XGBoost	-0.081***	-0.102***	-0.057***	-0.011	-0.070***
Early stage	IPO	CatBoost	-0.018***	-0.033***	-0.002	0.018	-0.046**
Mid stage	Follow-on	Logit	-0.052***	-0.038	-0.070*	-0.050*	-0.004
Mid stage	Follow-on	XGBoost	-0.061***	-0.049**	-0.075**	-0.047**	-0.011
Mid stage	Follow-on	CatBoost	-0.059***	-0.056**	-0.063**	-0.036	-0.021
Mid stage	Acquisition	Logit	-0.074***	-0.118**	-0.022	-0.056	-0.025
Mid stage	Acquisition	XGBoost	-0.035	-0.068	0.002	-0.034	0.004
Mid stage	Acquisition	CatBoost	-0.034	-0.066	0.002	-0.058	0.029
Mid stage	IPO	Logit	-0.043***	-0.094***	0.017	0.010	-0.072**
Mid stage	IPO	XGBoost	-0.038***	-0.065***	-0.009	-0.015	-0.018
Mid stage	IPO	CatBoost	-0.037***	-0.131***	0.072***	0.032	-0.083***

Return to Section 5.3.2