PitchBook
a Morningstar company

Institutional Research Group

pbinstitutionalresearch@pitchbook.com

Originally published March 2023
Updated on June 28, 2026

Contents

PitchBook VC Exit Predictor

Summary

The PitchBook VC Exit Predictor leverages machine learning and our vast database of VC companies, financing rounds, and investors to objectively assess whether a startup is likely to exit successfully and when that exit is likely to occur. The VC Exit Predictor framework comprises two primary components: a classification model that predicts the probability a startup will ultimately be acquired, go public, or not exit because it becomes self-sustaining or fails, and a discrete-time hazard model that predicts the probability an exit will occur within the next one, three, and five years. The exit type predictions are then used to calculate a naïve expected return for an investment in a company's next financing round, based on historical returns derived from cap table data. Finally, these expected returns are normalized across the VC universe by percentile rank to create a single metric called the Opportunity Score—a value between 0 and 100, where 100 represents the most attractive investment opportunity and 0 the least attractive.

We evaluated the exit-type classification model using a two-level hierarchical structure: how well it differentiates between companies that experience a successful exit and those that do not, and, among successful exits, how well it distinguishes between M&A and public listings. On out-of-sample data, the model achieved 71% accuracy in predicting success versus no exit and 94% in predicting between M&A and public listings, conditioned on a successful exit. There are nuances in evaluating the modeling setup and interpreting the overall accuracy numbers, which are discussed in more detail in the Evaluation section.

We then evaluated the exit timing model separately, with predictions conditioned on a successful exit. This decision was primarily driven by the fact that we often learn that a startup has failed with a delay and do not know exactly when the failure occurred. We also modeled the timing of M&A and public listing exits separately, given significant differences in their exit timing distributions and relationships with the model inputs. Across one-, three-, and five-year time horizons, the public listing model achieved average accuracies of 90.5%, 69.6%, and 61.3%, respectively. Across the same time horizons, the M&A model achieved average accuracies of 89.6%, 65.2%, and 56.1%, respectively.

Data

Individual data observations used to train and evaluate the models are associated with VC financing rounds, while inclusion is determined at the company level. We included companies that had raised at least two VC financing rounds (including angel and seed) and are no longer VC-backed, meaning they have undergone a merger or public listing, filed for bankruptcy, ceased operations, or become self-sustaining. Because many startup failures are not disclosed, a company that has not received a VC financing round in more than six years was deemed to have failed or to have become self-sustaining, based on an analysis of the empirical distribution of time between VC rounds. The inclusion criteria resulted in more than 109,000 observations from 52,000 distinct companies in the full dataset.[1] The accompanying table provides additional detail on the outcome distribution.

Outcome distribution

	Count	Proportion
No exit	28,014	54.0%
Merger	22,014	42.4%
Public listing	1,848	3.6%
Total	51,876	100.0%

Source: PitchBook • Geography: Global • As of June 28, 2026

Theoretically, model inputs could be generated daily for a company between its second VC financing date and its exit date. However, this is computationally infeasible and would produce highly correlated observations, since many features would not change from one day to the next. Significant feature updates typically occur after a financing event. Therefore, we generated a single observation per VC financing round for each qualifying company.[2] The prediction date for each observation was determined by randomly sampling from a uniform distribution between the close date of the current round and the close date of the next event (next VC round or exit). The prediction date determines which information is included in the input—only data known at the time of prediction is allowed to avoid look-ahead bias. Randomly sampling the prediction date, rather than using the close date of the current round, allows the model to learn how time affects outcomes. For example, a company that raised its last round a year ago has a better chance of successfully exiting than one that has not raised in four years, all else equal. In addition, this matches the structure of the data the model will be used on for inference (current VC-backed companies), where the time from the last close date will differ across companies.

Model inputs

The inputs, or features, for the machine learning classification model were compiled from the extensive information on each startup's PitchBook profile. In total, there are 36 features for each observation, which can be categorized into three main groups: company, deals, and investors. Almost all inputs are shared between the exit type and timing models, but the timing model also includes three additional macro inputs that capture the current public equity market return and IPO environment.

Company-level inputs can be further broken down into static and point-in-time information. Static features are basic descriptive data points about a company that rarely change, including industry/vertical, geographic location, and number of founders. Point-in-time features, on the other hand, are attributes whose values depend on when a prediction is made. This broad category includes data ranging from the business stage (e.g., product development, revenue generation, profitability) to patents. Additional point-in-time inputs include the number of employees, company age, acquisitions, and related news articles. Inputs in the deals category comprise data from announced, current, and past financing rounds, with a focus on VC deals. Key variables include the number of VC rounds, stock type, close date, and deal size.[3]

Engineering features from investor data, particularly those related to individual investor entities, present a challenge due to their high dimensionality. There are nearly 10,000 distinct investors in this analysis, and only a small fraction invests in each company. Rather than treating investors as a sparse, high-dimensional categorical feature, we developed a method to rank them by importance and experience within the VC universe. This ranking method relies on the well-known hypothesis that influential VC investors frequently collaborate by investing in the same companies, often likened to an exclusive social club. To capture this dynamic, we modeled VC investors as a social network where two entities are connected if they have invested in the same company. The connections, or edges, are then weighted by the number of distinct co-invested companies between pairs of investors. To quantify the idea that investors should be highly ranked if 1) they have VC investing experience and 2) they are connected with other experienced VC investors, we calculated the eigenvector centrality of the investor network.[4] In addition to the investor ranking, other inputs in this category include average capital invested per distinct investor, investor counts, counts by types other than VC (e.g., corporate VC), follow-on counts, lead-investor frequency, and geographic location.

VC Exit Predictor

Modeling

The first step in the modeling process is to split the data into training and test sets so the model is trained and evaluated on mutually exclusive samples. Extra care is needed at this step because multiple observations with the same outcome for the same company can lead to information leakage[5] between the training and test data. To avoid this pitfall, we partitioned the data at the company level so all of a company's

observations were either in the training set or the test set. Therefore, when the model made predictions on the test set, it had no prior information about any of the companies. We performed stratified random sampling to assign each company to the training or test set with an 80/20 split, which results in around 87,000 and 22,000 observations in the training and test sets, respectively.

Another aspect of the modeling process that deserved attention was the presence of an unbalanced outcome distribution. Unbalanced class distributions in supervised machine learning classification can cause the model to overemphasize the majority class during training, potentially leading to biased predictions in its favor. In this case, the class distribution is imbalanced toward mergers and no exits, with public listings accounting for less than 10% of the data. Startups that go public are often the most lucrative for investors and, therefore, are important for the model to perform well on. To mitigate the impact of class imbalance, we implemented SMOTE,[6] an oversampling method that creates synthetic observations for the minority class(es). Synthetic observations are created by randomly sampling along the hyperplanes (in the case of more than two dimensions) connecting all the k minority-class nearest neighbors, where k is a hyperparameter. The synthetic observations are therefore logical perturbations of the original data. These observations are used strictly during model training and are not considered during evaluation. The oversampling process effectively gives each class equal weight in the loss function during training.

The specific algorithm we employed is XGBoost,[7] a gradient-boosted classification tree model. Since its introduction, this algorithm has achieved state-of-the-art performance on many traditional machine learning tasks with two-dimensional input features. For this task, it outperformed a multinomial linear model, a multilayer perceptron (MLP) neural network, and a recurrent neural network with long short-term memory (LSTM) layers, in which financing rounds were treated as sequences. In addition, we chose XGBoost for its flexibility in handling outliers and missing data, its ability to represent complex non-linear functions, its robustness to data preprocessing, and its training efficiency.

Due to the natural hierarchy of the outcomes, we structured the predictor architecture as two complementary binary classification models. The first one produces the probability of any successful exit, $P\left(\text{Success} \mid X\right)$, and its complement, the probability of no exit, $P\left(\text{No Exit} \mid X\right) = 1 - P\left(\text{Success} \mid X\right)$. The second one produces the probability of an M&A and public listing exit, conditioned on one of those two events occurring:

$P\left(\text{M\&A} \mid X,\ \text{Success}\right)$ and $P\left(\text{Public Listing} \mid X, \text{Success}\right)$. The unconditional successful exit type probabilities can then be calculated as $P\left(\text{Exit Type} \mid X\right) = P\left(\text{Success} \mid X\right) \times P\left(\text{Exit Type} \mid X,\ \text{Success}\right)$. One of the primary benefits of this structure is that it enables greater model explainability than a single multi-class model does, as shown in the Prediction Drivers charts for eligible company profiles on the PitchBook platform.

Each model's hyperparameters were tuned using five-fold cross validation with both grid search and Bayesian optimization. Cross validation is a data splitting process

used to select the "best" set of hyperparameters, in which the training data is split multiple times (in this case, five) to create additional validation sets. Each fold is used once as a validation set, while the remaining folds are combined to train the model. Just like the training and test sets, observations were assigned to each fold at the company level to avoid information leakage.

Evaluation

The structure of the prediction problem, with time-based observations and an undefined forward window during which company outcomes become known, creates challenges for evaluating model performance. Typically, time-based problems are evaluated using historical backtests—train the model only on data available over several historical dates, then make predictions on "future" observations. However, in this case, many outcomes for those future observations remain unknown while companies are still VC-backed. This produces a biased evaluation set because companies that ultimately exit successfully tend to remain VC-backed longer. Therefore, we present two evaluation methods in this section: standard train-test splitting regardless of when the observations were taken, and two historical backtests, each with its own caveats.

The evaluation set for the first method contains about 22,000 round-level observations from more than 10,000 unique companies. Given that we care about company-level performance, consistent with how the model is deployed in real time, a company's observations are weighted inversely proportional to its total number of observations, so that each company is weighted equally. The model achieved an accuracy of 68.4% on three-class predictions and 71.0% on two-class predictions of no exit versus an M&A or IPO exit. While accuracy is easy to interpret, it can be misleading and should be viewed in the context of the outcome distribution. A good way to do this is to look at the confusion matrix, which, in this case, is a 3x3 matrix where the rows represent

Normalized confusion matrix

		True outcome			
		No exit	Merger	Public listing	Total
Predicted	No exit	41.1%	15.8%	0.4%	57.3%
	Merger	12.6%	26.0%	1.8%	40.3%
	Public listing	0.3%	0.7%	1.4%	2.4%
	Total	54.0%	42.4%	3.6%	100.0%

Source: PitchBook • Geography: Global • As of June 28, 2026

the predicted outcome and the columns represent the true outcome. The values are normalized by the total sample size.

Diagonal entries represent correct predictions, whereas off-diagonal entries represent different types of errors. For example, the entry in the first row and second column (15.8%) is the percentage of companies predicted not to exit, but that were instead acquired. Two summary metrics related to the confusion matrix that provide additional perspective are precision and recall. Precision is the accuracy with which the model predicts a specific outcome, and recall is the percentage of observations with a specific outcome that the model correctly identifies. Lastly, we included the area under the receiver operating characteristic curve (ROC AUC), a ranking metric that can be

Select evaluation metrics

	Precision	Recall	ROC AUC	Outcome proportion	Number of companies
No exit	71.8%	76.1%	0.78	54.0%	5,603
Merger	64.4%	61.2%	0.76	42.4%	4,403
Public listing	58.0%	39.0%	0.92	3.6%	370

Source: PitchBook • Geography: Global • As of June 28, 2026

interpreted as the probability that a randomly selected observation from one class has a higher predicted class probability than a randomly selected observation from another class.

Evaluating the model based on the confidence of its predictions is another interesting angle. Probabilities naturally convey confidence—the higher the probability of the predicted class, the more confident the model is, given data availability and signal clarity. This also aligns with how we see investors using it in practice, since it's infeasible to evaluate and invest in every company in the universe. Investors can use the predictions to prioritize high-confidence successful predictions and deprioritize high-confidence no-exit predictions. For roughly 25% of predictions in the test set where the probability of the predicted class is above 75%, accuracy improves to 82.8%, with precision for all three classes above 80%. The outcome distribution of the high-confidence predictions is roughly in line with that of the broader dataset.

A potential drawback of the model evaluation presented thus far is that the data was not separated by time. While we have excluded any forward-looking information from the features for an individual company, the training data contained some observations that had not yet occurred for some observations in the test data. Therefore, the predictions made for the test data are not a true backtest (that is, the model output could not have been replicated on the prediction date). Setting up a backtest for this problem is challenging because it requires balancing the need for enough data to both train and evaluate the model. We need to go back far enough so the companies for which predictions were made have a chance to mature and exit. However, if the backtest date is too early, there will not be a sufficiently large sample of VC-backed companies with known outcomes to adequately train the model.

With these limitations in mind, we chose the beginning of 2019 and 2023 as the start dates for two separate backtests. For the 2019 backtest, approximately 30% of the data was available to train the model, and around 61% of the 33,572 companies eligible for a prediction on that date have a known outcome for evaluation. The second backtest, starting in 2023, coincides with the launch of PitchBook's VC Exit Predictor. While we refer to it in the context of model evaluation as a backtest, this is the first possible "live" model test, given that the model existed on this date. This resulted in roughly double the training data compared to the 2019 backtest, but only about 30% of the 58,245 companies eligible for a prediction on that date have a known outcome.

While the full backtests cannot be completed at this time because many companies in the evaluation set are still VC-backed, the predictions that can be evaluated also pose challenges for performance analysis. The primary concern is that companies with known outcomes are not a random sample because the timing of outcomes varies. In particular, we observe a higher-than-expected share of companies with outcomes deemed "no exits" because companies that exit successfully tend to raise more funding rounds and remain VC-backed longer. In the 2019 and 2023 backtests, the no-exit rates for companies with known outcomes were 61.7% and 69.8%, respectively, compared with 54.0% in the training data. To correct for this sample bias, we performed an ex post calibration to the predictions to align the average probability of each class with its proportion in the evaluation set. Therefore, the performance metrics (excluding AUC, which is unaffected by monotonic calibration) presented below are marked with an asterisk, but we believe this is a fair way to evaluate an incomplete backtest.

Select evaluation metrics for the 2019 backtest

	Precision	Recall	ROC AUC	Outcome proportion	Number of companies
No exit	72.9%	82.6%	0.74	61.7%	12,650
Merger	57.2%	44.5%	0.71	34.8%	7,133
Public listing	47.1%	39.3%	0.92	3.5%	713

Source: PitchBook • Geography: Global • As of June 28, 2026

Select evaluation metrics for the 2023 backtest

	Precision	Recall	ROC AUC	Outcome proportion	Number of companies
No exit	80.6%	89.0%	0.81	69.8%	12,162
Merger	62.1%	45.7%	0.80	28.5%	4,964
Public listing	33.1%	38.9%	0.94	1.7%	288

Source: PitchBook • Geography: Global • As of June 28, 2026

Overall, the accuracies for 2019 and 2023 were 67.8% and 75.8%, respectively. Compared with the non-time-based evaluation, other performance metrics were mixed. For the 2019 results, precision, recall, and AUC were slightly lower across the board. While the accuracy for 2023 was aided by the higher majority-class rate, AUC was also higher across all three outcomes.

Opportunity Score

The scoring process maps the outcome probabilities from the classification model to a naïve expected return from the perspective of an investor in a company's next VC financing round, based on historical returns by series derived from cap table information. Scoring serves two main benefits: 1) it creates a single value for each company, which is necessary for the final rankings, and 2) it quantifies the benefit of investing early in successful startups and/or exiting them via the public markets.

Average returns and holding periods by series

	Average return		Average holding period (years)	
	Merger	Public listing	Merger	Public listing
Series A	36.7%	47.8%	5.3	5.9
Series B	31.0%	37.9%	4.7	4.3
Series C	28.0%	34.4%	4.4	3.7
Series D+	20.0%	30.0%	3.7	3.0

Source: PitchBook • Geography: Global • As of June 28, 2026

The expected return for an individual startup is a weighted geometric average of historical returns, based on the upcoming VC financing round. For example, if a startup had last raised a Series B, the relevant returns would be for Series C investments. The weights are taken to be the probabilities of each exit outcome from the classification model. The tables below show average annualized startup returns by series and type, as well as the average holding period.[8] For simplicity, we assume that a failure results in a total loss at all stages.

For example, consider a startup that recently raised a Series B with failure, merger, and public listing exit probabilities of 50%, 30%, and 20%, respectively. The annualized geometric expected return would be calculated as follows:

$$\bar{r} = \left(1.28^{4.40} \times 0.3 + 1.34^{3.74} \times 0.2\right)^{\left(\frac{1}{4.40 \times 0.6 + 3.74 \times 0.4}\right)} - 1 = 10.0\%$$

Finally, the return figures are normalized as a percentile ranking across all eligible VC-backed companies. A percentile ranking of 100 represents the most attractive company, while a ranking of 0 represents the least attractive.

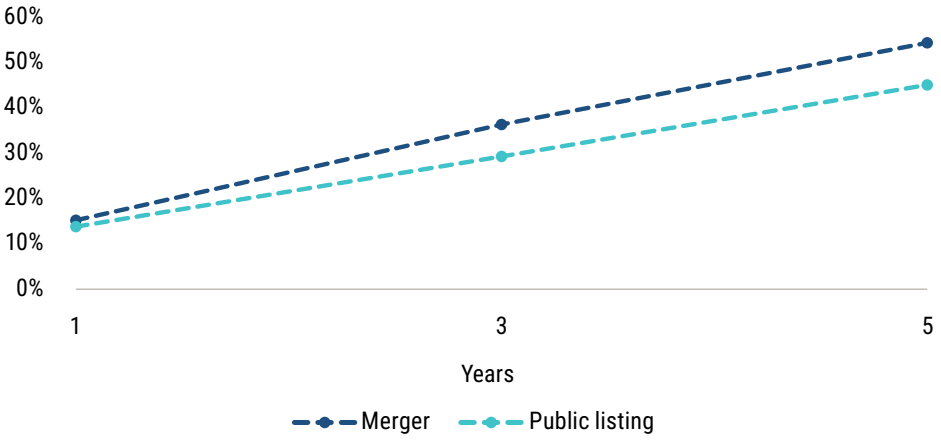
VC Time to Exit Predictor

The exit timing model complements the VC Exit Predictor by offering insights on when a successful exit is likely to occur. The exit timing predictions are conditioned on a

successful exit, answering the question: “If this company does go public, what is the probability it will happen within the next t years?” Due to differences in exit timing, M&A and public listing exits are modeled separately. This model design feature was driven by two factors: 1) we do not have reliable data on when VC-backed companies fail, making it impossible to model as a competing event, and 2) the events that each model predicts are independent, allowing the output from the two to be easily combined. Multiplying the probability that a company will ever undergo a successful exit by the probability of whether that exit will occur within a certain time, assuming that it will occur at some point, yields the probability that the event will happen within that time.

The underlying data source and methodology are the same as the VC Exit Predictor. However, because the outcomes are defined within specific forward windows (1, 3 and 5 years), some data from companies that are still VC-backed can be used for training and evaluation. This does create its own challenges because it requires determining

Cumulative exit rates by type



Source: PitchBook • Geography: Global • As of June 28, 2026

which VC-backed companies will succeed, since, by definition, timing predictions are conditional on success. The exit type predictions are therefore inherently useful to the timing model by providing weights for companies that are still VC-backed and have uncertain outcomes. This effectively leads to soft censoring of observations that have not yet experienced a successful exit, as discussed in more detail in the following section.

Modeling

The overall model architecture aims to output the cumulative probability that an exit will occur by time t, conditioned on the exit occurring at some point. This can be expressed formally in the following term, where S_k is a binary indicator of a successful exit of type k

$$h_{k,j} = P\left(T \leq t_j \mid T > t_{j-1},\ S_k = 1\right)$$

$$S_k\left(t\right) = \prod_{j=1}^{J(t)} \left(1 - h_{k,j}\right)$$

$$F_k\left(t\right) = 1 - S_k\left(t\right)$$

(IPO or M&A) and X is the feature matrix: $P(T \leq t \,|S_k = 1, X)$. This output is produced by a series of discrete-time hazard models applied to manually defined time intervals in set J of 1, 3, and 5 years. Each model produces hazard rates, $h_{k,\,j}$, over the window $[t_{j-1}, t_j)$, which are then chained to produce the survival function $S_k(t)$–the probability that a company remains VC-backed. The cumulative probability of an exit, $F_k(t)$, is simply the inverse of the survival function.

The benefits of this model structure include a statistically valid cumulative probability function as well as working seamlessly with the exit type predictions. Given that an eventual exit of type k and the conditional timing of that exit are complementary events, we can calculate the unconditional probability of a type k by time t as:

$$P(S_k = 1 \cap T \leq t) = P(S_k) \times F_k(t)$$

Additionally, we can calculate the conditional probability of either an IPO or M&A exit by time t as:

$$P(T \leq t \mid S = 1) = \frac{P(S_{IPO}) \times F_{IPO}(t) + P(S_{M\&A}) \times F_{M\&A}(t)}{P(S_{IPO}) + P(S_{M\&A})}$$

The main drawback of this modeling approach is that the training data, which includes both companies that have already exited and those still VC-backed (censored), is not fully defined. This stems from the fact that the exit timing predictions are conditioned on $S_k = 1$, which is unknown for censored observations. To mitigate this issue, we introduced the concept of "soft censoring," which assigns a sample weight to each observation during training as $P(S_k)$.[9] Observations for companies with $S_k = 1$ receive a full sample weight of 1.

For the exit type k hazard model covering the time window $w \in [t_{j-1}, t_j)$, the model training data consists of the at-risk set: observations for companies that had not experienced exit type k by t_{j-1} and were not censored before t_j.[10] Observations for companies where $S_k = 0$ (that is, those that experienced a different exit type, including failures) are excluded from the training data. Finally, the training labels are defined as $y_i = 1$ if company i exited within $[t_{j-1}, t_j)$, and $y_i = 0$ otherwise. The hazard rates for each independent time window were generated using a set of independent XGBoost binary classifiers that underwent a hyperparameter tuning process similar to that of the VC Exit Predictor models.

Top-down calibration

While training the models on longitudinal data is necessary for them to understand exit timing distributions across different environments, this approach embeds the incorrect assumption that cross-sectional predictions at a single point in time are independent. The most important counterpoint to the independence assumption is that demand for exits (that is, available capital) for a cohort of VC-backed companies is less sensitive to macro conditions than expected. When applied cross-sectionally, the exit timing predictions tend to overestimate the impact of macro conditions on both the upside and the downside. The calibration process also protects the model from making poor extrapolations on macro conditions it did not see during training.

To correct for this overconfidence, we implement a top-down calibration process that adjusts individual company predictions so that the average predicted hazard rates equal the expected aggregate hazard rates derived from a top-down macro model. For the initial hazard model of exit probability within one year, we model the log odds of the expected aggregate exit hazard rate at time t, $\bar{h}_t$, using a linear regression with macro covariates.[11] These covariates include the number of VC-backed IPO exits in the past six months,[12] the change in that number relative to the prior six-month period, and the return of the Morningstar US Large Cap Index over the past six months. For a set of N VC-backed companies at time t over time window $w \in [0,\ 1)$, $\bar{h}_{t,w}$ represents the percentage of companies expected to exit within the next year.[13,14]

$$ln\left(\frac{\bar{h}_{t,w}}{1-\bar{h}_{t,w}}\right) = \alpha + X_t\beta$$

For the remaining hazard models that condition on a company remaining VC-backed for an additional period, current macro conditions have not been useful predictors of aggregate hazard rates. For these models, we set the calibration parameter $\bar{h}_{t,w}$ as the historical average hazard rate observed across cohorts. The empirical data used to determine the calibration parameters were semi-annual VC-backed company cohorts beginning in 2001. For the cohort at time t with N companies, the cumulative exit rate over time t_0 to t_j is calculated as:

$$F\left(t_j\right) = \frac{\sum_{i=1}^{N} y_i wt_i}{\sum_{i=1}^{N} wt_i}$$

where $y_i = 1$ if company i exited by time t_j and $wt_i = P\left(S = 1\right)$ if company i is currently still VC-backed, and 1 if the company has already exited. The observed hazard rate over time window $w \in [t_{j-1}, t_j)$ can then be derived using the survival function:

$$h_{t,w} = \frac{S\left(t_{j-1}\right) - S\left(t_j\right)}{S\left(t_{j-1}\right)}$$

The next step in the calibration process is to set the weighted average predicted hazard rate for each time window to the target $\bar{h}_{t,w}$ derived in the prior step. This is done via a mean-shift process in which the predicted hazard rate logits, denoted as γ, are updated using:

$$\hat{\gamma}_{t,w} = \alpha + \hat{\gamma}_{t,w}$$

The scaling parameter is determined by solving the following equation:

$$\bar{h}_{t,w} = \frac{\sum_{i=1}^{N} \sigma\left(\alpha\ +\ \hat{\gamma}_i\right) wt_i}{\sum_{i=1}^{N} wt_i}$$

Evaluation

We evaluated the timing model using overlapping annual historical backtests from the beginning of 2018 to the start of 2025. This approach is more straightforward than evaluating exit-type predictions because all outcomes are observed within a defined forward window. However, as the backtest date approaches the present, not all exit time horizons can be evaluated. For example, only the one-year horizon can be evaluated for models as of 2024 and 2025, since the three-year forward window extends into the future.

We found that the models performed better at shorter time horizons and when predicting the timing of public listings. These results are somewhat intuitive, as the closer a company is to an exit, the more relevant the data available today is. Predictive signals for the timing of merger exits were limited, and the model was rarely confident that one would occur within the next year. As with the VC Exit Predictor, we found that high-confidence predictions were much more useful for identifying likely exit candidates. For example, companies in the top decile for 1-year public listing predictions were, on average, 3.3x more likely to go public than the remaining 90%.

Select exit timing model performance metrics

	Public listing				Merger			
	Accuracy	Precision	Recall	ROC AUC	Accuracy	Precision	Recall	ROC AUC
1-year	90.5%	58.9%	27.6%	0.73	89.6%	51.6%	3.7%	0.63
3-year	69.6%	47.8%	32.3%	0.66	65.2%	41.3%	17.2%	0.59
5-year	61.3%	60.4%	51.0%	0.66	56.1%	60.9%	42.1%	0.60

Source: PitchBook • Geography: Global • As of June 28, 2026

References

1: Data and performance metrics were generated on June 26, 2026.

2: The data frequency of observations used for model training and evaluation differ from that used for model inference. The outcome probabilities and scores shown on the PitchBook Platform are updated daily.

3: Due to better data coverage, deal size is used as a proxy for valuation.

4: The concept of eigenvector centrality was famously used in Google's PageRank algorithm. For more information on network centrality, see: "Network Centrality: An introduction," arXiv, Francisco Aparecido Rodrigues, January 22, 2019.

5: Information leakage occurs when the test set contains information from the training set that can cause overfitting and optimistic performance evaluation on the test set. This form of information leakage is known as identity confounding because the model learns identities (that is, companies) as well as features.

6: "*SMOTE*: Synthetic Minority Over-Sampling Technique." *Journal of Artificial Intelligence Research*, N.V. Chawla, et al., June 1, 2002.

7: "XGBoost: A Scalable Tree Boosting System," arXiv, Tianqi Chen and Carlos Guestrin, June 10, 2016.

8: Holding periods of less than one year are not annualized. In addition, outlier returns of more than 350% are excluded from the average calculations.

9: To improve model training efficiency, observations for companies with $P(S_k) < 0.1$ are discarded.

10: Observations are censored due to insufficient time past relative to the training date. In these cases, the labels are not yet known. The censoring time in years for observation i is: $c_i = (d_{\text{train}} - d_i) \,/\, 365$.

11: When the beginning time in the hazard model time window is 0, the aggregate exit hazard rate is equal to the percentage of companies expected to exit within that time window.

12: Both IPO-related variables are normalized by the trailing three-year average number of IPOs across monthly rolling six-month periods.

13: The calibration process is performed separately for M&A and IPO exits. The exit type k notation is dropped for simplicity in this section.

14: The predicted aggregate hazard rates in nominal terms $\bar{h} \in [0,1]$ are adjusted using a probit approximation to account for the fact that the expected value of a logit-transformed hazard rate is not equal to the logit of the expected hazard rate due to Jensen's inequality. Additionally, the residual variance of the regression is calculated using an effective sample size adjusted for autocorrelation arising from overlapping cohort data.

PitchBook provides actionable insights across the global capital markets.

Nizar Tarhuni
Executive Vice President of Research and Market Intelligence

Daniel Cook, CFA
Senior Director, Global Head of Quantitative Research and Market Intelligence

---

Learn more about PitchBook's Institutional Research team.

Click here for PitchBook's report methodologies.

PitchBook Insights is an online compendium of in-depth data, news, analysis, and perspectives that shape the private capital markets.

PitchBook subscribers enjoy exclusive access to a comprehensive suite of private market insights, including proprietary research, news, data, tools, and more on the PitchBook Platform.