Interpretable Machine Learning for Predicting Startup Funding, Patenting, and Exits

Saeid Mashhadi* Amirhossein Saghezchi[†]
Vesal Ghassemzadeh Kashani[‡]
October 13, 2025

Abstract

This study develops an interpretable machine learning framework to forecast startup outcomes, including funding, patenting, and exit. A firm-quarter panel for 2010-2023 is constructed from Crunchbase and matched to U.S. Patent and Trademark Office (USPTO) data. Three horizons are evaluated: next funding within 12 months, patent-stock growth within 24 months, and exit through an initial public offering (IPO) or acquisition within 36 months. Preprocessing is fit on a development window (2010-2019) and applied without change to later cohorts to avoid leakage. Class imbalance is addressed using inverse-prevalence weights and the Synthetic Minority Oversampling Technique for Nominal and Continuous features (SMOTE-NC). Logistic regression and tree ensembles, including Random Forest, XGBoost, LightGBM, and CatBoost, are compared using the area under the precision-recall curve (PR-AUC) and the area under the receiver operating characteristic curve (AUROC). Patent, funding, and exit predictions achieve AUROC values of 0.921, 0.817, and 0.872, providing transparent and reproducible rankings for innovation finance.

Keywords: Startup Finance, Interpretable Machine Learning, Innovation Prediction, Predictive Modeling, SHAP Analysis

^{*}LeBow College of Business, Drexel University, Philadelphia, PA 19104 (email: saeid.mashhadi@drexel.edu)

[†]LeBow College of Business, Drexel University, Philadelphia, PA 19104

[‡]University of Tehran, Tehran, Iran

1 Introduction

Innovation is central to value creation in entrepreneurship. For high-growth startups, the ability to generate and protect new knowledge shapes product-market paths, affects access to external finance, and influences the likelihood of timely exit. Patents are central to this process, as they secure appropriability and act as credible signals that reduce information frictions between founders and investors (Conti et al., 2013; Hsu and Ziedonis, 2013). Causal evidence shows that patent grants improve a venture's prospects by increasing financing access, growth, and the probability of IPO or acquisition (Farre-Mensa et al., 2020). At the ecosystem level, venture capital is associated with innovative output, reinforcing a cycle between invention, investor attention, and firm growth (Kortum and Lerner, 2000). The ability to forecast which startups will expand their intellectual property portfolios, raise additional capital, and reach an exit has direct value for investors, corporate acquirers, and policy stakeholders.

Predictive work in entrepreneurial finance and strategy generally follows two strands. One focuses on financing outcomes, such as the timing and likelihood of later rounds, based on observable histories of capital raised, investor breadth, and founder characteristics (Gompers et al., 2010; Eggers and Song, 2015; Gompers et al., 2020). The other uses innovation as an explanatory factor, showing that patent stocks and citations anticipate financing and exits (Hsu and Ziedonis, 2013; Farre-Mensa et al., 2020; Kogan et al., 2017). Recent machine learning (ML) studies use large platforms such as Crunchbase to predict survival, fundraising, or valuation by capturing non-linearities and high-dimensional interactions (Antretter et al., 2019; Dalle et al., 2017). Parallel work in corporate finance shows that interpretable machine learning can recover meaningful drivers while delivering competitive out-of-sample accuracy. For example, Kim et al. (2024) use a model zoo with imputation, oversampling, and SHAP-based explanations to predict activist targets.

Two gaps remain. First, finance- and innovation-focused predictions are often studied separately. This separation leaves open questions about the joint and relative roles of fi-

nancing recency and momentum versus intellectual property stocks in shaping near-term funding, medium-horizon patenting, and longer-horizon exit. Second, while ML can improve discrimination, many applications do not enforce leakage-safe time splits or provide transparent interpretability and reliability checks. Scalable and interpretable prediction of startup patenting, defined as whether a firm will expand its patent stock over two years, has received less systematic attention than fundraising or exit forecasting, despite its importance.

This paper. An interpretable ML pipeline is developed to forecast three outcomes on a leakage-safe firm-quarter panel: (i) next financing within 12 months, (ii) patent-stock growth within 24 months, and (iii) IPO or acquisition within 36 months. The panel was built from Crunchbase and merged to U.S. Patent and Trademark Office (USPTO) assignee data to track cumulative patent and citation stocks (Dalle et al., 2017; Kogan et al., 2017). All steps were scripted in Python for exact reproducibility (McKinney, 2010). To avoid lookahead, preprocessing (infinite-to-NaN, missing-value flagging, and median imputation) was fit on a development window (2010-2019) and applied without change to a holdout window (2020-2021) and a final window (2022-2023). Class imbalance was addressed inside development with inverse-prevalence weights and Synthetic Minority Oversampling Technique for Nominal and Continuous features (SMOTE-NC), and evaluation splits were not resampled or reweighted. A linear baseline (regularized logistic) was compared with tree ensembles (Random Forest (Breiman, 2001), XGBoost, LightGBM, CatBoost). Winners were selected by area under the precision-recall curve (PR-AUC) with area under the receiver operating characteristic curve (AUROC) as a tiebreaker. Models were diagnosed with interpretable artifacts: Tree-based SHapley Additive exPlanations (TreeSHAP) for gradient-boosted trees where feasible and permutation or impurity importance for Random Forests, along with calibration curves and partial dependence (Lundberg and Lee, 2017).

The unit of analysis is the firm-quarter. Outcomes were defined strictly forward from each quarter-end, and rows too close to panel boundaries for a full horizon were flagged as non-evaluable. The time splits are non-overlapping: 2010-2019 (development), 2020-2021 (holdout), and 2022-2023 (final). Predictor construction relied on signals highlighted in prior work on startup dynamics, including financing recency and momentum, cumulative exposure to investors and capital, and intellectual property stocks and citations (Gompers et al., 2010; Conti et al., 2013; Hsu and Ziedonis, 2013; Farre-Mensa et al., 2020). The development feature list is persisted to guarantee column alignment at scoring time, and deterministic seeds, filenames, and manifests are used to enable replication.

Three results summarize the contribution. First, startup patenting is highly predictable at the firm-quarter level. On the holdout window, the Random Forest trained with inverse-prevalence weights achieves AUROC = 0.921 and PR-AUC = 0.631 for predicting a two-year increase in the patent stock. Global importance and partial-dependence diagnostics show pronounced mean reversion with respect to existing IP stock and a negative gradient with time since last round, which is consistent with the view that younger, actively financed firms are more likely to expand their portfolios.

Second, fundraising within 12 months is dominated by recency and momentum. On the final window, a LightGBM model with inverse-prevalence weights attains AUROC = 0.817 and PR-AUC = 0.220. SHAP identifies days since last round and age as first-order drivers, with concave gains from cumulative capital and investor breadth. Reliability curves show optimism at higher scores. Isotonic recalibration is therefore recommended when calibrated probabilities, rather than rankings, are required.

Third, exits are associated with financing maturity. For 36-month exits, the Random Forest with inverse-prevalence weights delivers AUROC = 0.872 and PR-AUC = 0.559 on holdout. Importance and partial dependence indicate that time since last round and age, together with cumulative capital raised, are the leading drivers of exits, with investor breadth and cumulative rounds contributing secondarily. Together, these results show that interpretable ML can deliver ranking-useful precision on out-of-time cohorts while recovering mechanisms that align with economic priors. Fundraising behaves as a recency and momen-

tum phenomenon, patent growth reflects maturity and mean reversion, and exits load on financing depth.

Relation to the literature. The analysis complements and integrates three literatures. From entrepreneurial finance, roles for founder and firm maturity and for financing histories in shaping outcomes are confirmed (Gompers et al., 2010; Eggers and Song, 2015; Gompers et al., 2020). From the innovation literature, patent stocks and citations are used as predictive inputs rather than only as outcomes, consistent with patents both signaling quality and causally enhancing financing and growth (Conti et al., 2013; Hsu and Ziedonis, 2013; Farre-Mensa et al., 2020; Kogan et al., 2017). From ML in finance and innovation, an interpretable, leakage-averse pipeline is adopted in the spirit of Kim et al. (2024), targeted to the startup domain with quarter-level panels, non-overlapping time splits, and explicit reliability checks. The novelty is the combination of a patent-growth dependent variable at scale, unified benchmarking across funding, patenting, and exit horizons within one pipeline, and deployable ranked lists from strictly out-of-time scoring.

Contributions. The paper makes four contributions:

- An interpretable, leakage-safe pipeline for startup outcomes. A firm-quarter panel (2010-2023) with strict non-overlapping time splits, development-only preprocessing, and persisted feature lists is constructed, providing a reproducible benchmark that integrates funding, patenting, and exit horizons (Dalle et al., 2017; McKinney, 2010).
- Scalable prediction of patent-stock growth. Two-year patent expansion is shown to be predictable out of time (AUROC = 0.921, PR-AUC = 0.631), with interpretable drivers that align with economic priors on maturity and financing recency (Kogan et al., 2017; Farre-Mensa et al., 2020).
- Unified model benchmarking with transparent diagnostics. Logistic and mod-

ern tree ensembles are compared under imbalance treatments. Selection is by PR-AUC, and SHAP or importance, partial dependence, and calibration curves are reported to pair discrimination with explainability (Breiman, 2001; Lundberg and Lee, 2017; Kim et al., 2024).

• Deployable ranked target lists. Validated models are translated into out-of-time scored cohorts for screening and benchmarking, and the conditions for post hoc calibration are clarified for probability-based decisions.

The remainder of this study is structured as follows. Section 2 discusses related work. Section 3 details data collection (Crunchbase-USPTO merge), preprocessing, imbalance handling, and model families. Section 4 reports the results, including descriptive diagnostics, model selection, interpretability, calibration, and out-of-time scoring manifests. Section 5 concludes.

2 Literature Review

Research on predicting startup outcomes spans finance, entrepreneurship, and innovation, and has shifted from traditional statistical analyses to broader predictive frameworks. Early work identified observable correlates of startup success or failure, including the timing and size of funding rounds, founder experience, and market conditions. Venture investors have consistently emphasized founding-team quality as a main factor in investment decisions (Gompers et al., 2020). Empirical work supports this view, as founders with prior success tend to succeed again, while experience from failure provides a weaker benefit (Gompers et al., 2010; Eggers and Song, 2015). These results suggest that human capital and entrepreneurial persistence, together with early financing dynamics, are core elements behind heterogeneity in startup performance.

A second stream studies innovation outputs, especially patents and their quality, as predictors of venture success. Patents play a dual role by protecting intellectual property

and signaling technological quality, which helps reduce information asymmetries between entrepreneurs and investors (Conti et al., 2013). Evidence shows that startups holding patents, particularly highly cited ones, are more likely to secure additional financing and reach successful exits (Hsu and Ziedonis, 2013; Farre-Mensa et al., 2020). Quasi-experimental evidence provides causal support, as Farre-Mensa et al. (2020) exploit quasi-random examiner assignment and show that favorable patent grants are followed by greater growth, improved access to external capital, and higher chances of IPO or acquisition. These findings indicate that patents do more than correlate with success, as they also shape a venture's trajectory. At the aggregate level, venture capital investment is associated with higher innovative output, creating a reinforcing cycle between innovation, investor attention, and firm growth (Kortum and Lerner, 2000). Overall, innovation indicators appear central for predicting startup outcomes.

The availability of large-scale datasets such as Crunchbase and PitchBook has encouraged the use of ML to forecast startup outcomes. ML is well suited to this setting because it captures non-linearities and high-dimensional interactions that logit or probit models may miss. Common targets include next-round financing, survival, valuation milestones, and exit events. Evidence indicates that tree-based ensembles perform well on these tasks. Antretter et al. (2019) use Twitter-based measures of online legitimacy to predict five-year survival and show that social attention adds predictive content beyond standard covariates. Related work in corporate finance shows the value of interpretable ML, as Kim et al. (2024) develop interpretable random-forest models with SHAP-based explanations to predict activist fund targets and show that SHAP recovers economically meaningful drivers such as valuation and free float. Post hoc interpretability tools such as SHAP are used to attribute predictions in high-capacity models, which improves transparency and comparison with theory (Lundberg and Lee, 2017). Together, these studies suggest that ML can improve out-of-sample accuracy in high-dimensional settings while preserving interpretable economic insights.

Despite these advances, two limitations persist. Financial and innovation factors have

often been studied separately, as innovation-focused work emphasizes patenting outcomes while finance-oriented work focuses on funding dynamics, which leaves open questions about their joint and relative importance. In addition, few studies benchmark modern ML methods against econometric alternatives under strict out-of-sample validation while also requiring interpretability. The present study addresses these gaps by integrating financial and innovation features within a single ML framework, benchmarking ensemble-tree methods against baseline logistic models, and applying SHAP analysis to recover economically meaningful drivers. The goal is to show that interpretable ML can raise predictive accuracy and clarify why innovation signals and financial traction forecast funding, innovation, and exit outcomes.

3 Research Methodology

3.1 Data Collection

A U.S. startup financing panel was constructed from Crunchbase covering 222,126 funding rounds announced between 2010 and 2023.¹ The unit of observation is the *funding round*. Each record is identified by a unique funding-round identifier linked to an organization identifier and a standardized issuer name. Rounds with missing organization identifiers were dropped, and duplicate rows were removed by exact match on the funding-round identifier. For each round, the announcement date, investment type (e.g., seed, venture, grant, debt), and stage category (early-stage, mid-stage, late-stage, other) were observed. Investor participation was tracked using the number of investors per round and the set of lead investors when available. Exit outcomes were proxied by platform flags for initial public offerings (IPO) and acquisitions. The round-level data are transformed into a firm-quarter panel in §3.2.

To capture innovative activity, the Crunchbase panel was merged with USPTO data.

¹Crunchbase is a widely used, continuously updated directory of entrepreneurial activity. See, e.g., Dalle et al. (2017) for documentation and validation of its research use.

Organizations were linked to their granted patents using assignee information and standard name-matching procedures, which allowed observation of both patent counts and forward citations at the startup level. These variables serve as proxies for the quantity and quality of technological output, following prior work on patent-based measures of innovation (Kogan et al., 2017).

The dataset yielded a median round size of \$1.76 million (mean \$20.99 million; 95th percentile ≈\$63.5 million) and a median of one participating investor. California, Massachusetts, and New York account for the largest shares of issuers, while the most frequent industry labels are Consumer goods and services, Computer software and Internet, and Biotechnology and health care. Exit flags indicate that 6.64% of issuer-round pairs are associated with an IPO and 14.91% with an acquisition in the broader firm history. These patterns align with stylized facts of entrepreneurial finance, including skewed capital distributions, geographic concentration in a few hubs, and sparse binary exit outcomes.

3.2 Data Processing

Figure 1 (Box 1) provides an overview of the transformation from raw Crunchbase extracts to a leakage-safe firm—quarter panel for 2010—2023. All steps were scripted in Python and versioned for exact reproducibility (McKinney, 2010). Platform identifiers (org_uuid) and quarter-end dates were retained to preserve a lossless lineage (Dalle et al., 2017).

The panel was created by building a quarterly calendar for each firm, beginning at the first observed activity (or the founding date, when available) and ending at the earliest exit or December 31, 2023. Periods following an IPO or acquisition were excluded to prevent post-outcome contamination. Financing activity was aggregated to the quarter level, producing counts of rounds, participating investors, and total capital raised. To reduce heterogeneity in platform labels, investment types were classified into four categories: early-stage, mid-stage, late-stage, and other. Cumulative histories were constructed for total capital raised, number of rounds, and number of investors, together with momentum measures (e.g., funding raised

in the prior four quarters) and recency measures (days since the last round).

Innovation variables were merged from matched USPTO data. Patent counts and citation totals were aligned to the firm–quarter calendar and carried forward within firm histories (with no backfilling prior to a firm's first observation), creating cumulative stocks that serve as proxies for technological output (Kogan et al., 2017). Firm descriptors such as industry, geography, and founding date were attached, and firm age in years was computed.

Outcomes were defined strictly forward from the quarter boundary. Three dependent variables were created: (i) whether the firm raises another funding round within 12 months, (ii) whether its patent stock increases within 24 months, and (iii) whether it experiences an IPO or acquisition within 36 months. Observations that are too close to the panel boundary for a full horizon were marked as non-evaluable and excluded from the corresponding outcome. Temporal splits follow Figure 1: development (2010–2019), holdout (2020–2021), and final (2022–2023). All preprocessing parameters in Box 1 were estimated on the development period only and then applied forward without change.

Missingness was handled within Box 1 by replacing infinities with NaN, adding explicit indicators for features with $\geq 10\%$ NA in development plus a forced flag for days since last round, and applying median imputation fitted on development and reused unchanged on holdout and final. To ensure column alignment across splits, the development feature list was persisted and applied at scoring time.

Because outcomes are rare, Box 2 prepared alternative training matrices on the *development* slice only (evaluation splits were never modified): inverse-prevalence class weights, random oversampling (ROS), SMOTE-NC, Borderline-SMOTE, and ADASYN. For the main analysis, only the weights and SMOTE-NC variants were carried forward, and ROS and the Borderline-SMOTE/ADASYN fallbacks were retained for robustness in the appendix.

3.3 Methodological Approach

Each task was framed as a quarterly binary classification problem with non-overlapping time splits (Figure 1, Boxes 3–5). A Random Forest (Breiman, 2001) is used as the base-line estimator in the model zoo and is compared with regularized logistic regression and gradient-boosting algorithms (XGBoost, LightGBM, CatBoost). This setup allows comparison between a linear baseline and flexible tree ensembles that capture non-linearities and interactions.

Model development and selection (Box 3). For each dependent variable and each imbalance variant from Box 2, models were trained on the *development* period using the persisted feature list for column alignment. Infinities were set to NaN, and any residual missingness was filled with development-fitted medians. Class weights or SMOTE-NC were applied only during development, and evaluation splits were never resampled or reweighted. Model selection prioritized PR-AUC as the primary criterion, with AUROC as a tiebreaker. Brier score and precision@K ($K \in 30, 100, 500$) were also reported to quantify ranking usefulness. Evaluation was strictly out of time, with the *final* window used for 12-month funding and the *holdout* window used for 24-month patent growth and 36-month exits.

Model analysis (Box 4). Interpretability and reliability were computed on the evaluable out-of-time split (final or holdout, as applicable). Where feasible, TreeSHAP was computed on the winning tree model using capped samples (and optional forest thinning) to obtain global importance (mean |SHAP|) and dependence plots. If TreeSHAP was infeasible due to memory or runtime, permutation importance scored by average precision was used; for logistic baselines, coefficient magnitudes were reported. Each importance figure and table states the exact method used. Calibration was assessed with quantile-binned reliability curves and the Brier score. Where curves showed optimism, it was noted that post hoc isotonic scaling can improve probability calibration without affecting ranking.

Out-of-sample prediction (Box 5). The winning specification for each outcome was applied to the most recent cohorts consistent with the horizon (h=12/24/36 months). Columns were aligned via the persisted feature list, development medians were applied unchanged, and no reweighting or resampling was used on evaluation data. Predicted probabilities \hat{p} , integer ranks, and percentiles were produced, with one row per organization after deduplication. Random seeds, file paths, and manifests were fixed to guarantee exact replication.

4 Results

4.1 Descriptive Setup and First Signals

This section reports diagnostics from the learning panel (2010-2023) and a univariate screen for each outcome. The panel is split by calendar time into a development sample (through 2019), a holdout sample (2020–2021), and a final test sample (2022–2023). By construction, only the 12-month funding outcome is evaluable in the final window; the 24-month patent and 36-month exit outcomes are evaluated on the holdout window. In the final window, the evaluable base rate for next financing within 12 months is 8.8%.

Missingness is concentrated in two variables, days since last round ($\approx 24\%$) and firm age ($\approx 6\%$). Other predictors are essentially complete. Training rows typically contain at most one missing feature, which motivates the median imputation used in the pipeline (Refer to §3.2). Table 1 summarizes pairwise correlations with the three outcomes.

The univariate screen is computed on development rows that are evaluable for each outcome to avoid look-ahead. Signed AUC is used (so values > 0.5 indicate helpful discrimination regardless of direction), and PR-AUC magnitudes are reported; direction is indicated by the AUC sign (see Table 2).

Next financing (12 months). Signals related to recency and momentum dominate. Days since last round loads negatively and shows meaningful stand-alone discrimination (signed AUC ≈ 0.69 ; signed PR-AUC ≈ 0.27). Age is also negative (AUC ≈ 0.69). Short-horizon

momentum, rounds in the last four quarters and funding in the last four quarters, and accumulated exposure, cumulative investors and cumulative rounds/raised, load positively (AUCs $\approx 0.50-0.58$). Patent-stock variables have small negative associations.

Patent growth (24 months). The pattern is consistent with a change in stock rather than a short-run flow. Larger existing patent and citation stocks are negatively related to future increases, with relatively strong univariate lift (AUCs $\approx 0.68-0.70$). Younger firms are more likely to expand their portfolios (age: AUC ≈ 0.70 , negative sign). Financing-activity variables carry small and mostly negative loadings.

Exit (36 months). Exit likelihood rises with maturity. Cumulative investors, cumulative rounds, and cumulative capital raised are positive (AUCs ≈ 0.64 –0.66). Later-stage exposure, mid/late counts, tilts positive (AUCs ≈ 0.54 –0.55). Days since last round is negative (AUC ≈ 0.61). Intellectual-property stocks show mild positive associations in the descriptive screen. Takeaway. The screen aligns with economic intuition. Near-term fundraising is driven by recency and momentum, patent growth is mean-reverting with respect to existing IP stock and age, and exits move with accumulated financing maturity, with IP as a modest complement. These descriptive signals set priors for the model-based results that follow (§4.5).

4.2 Univariate Discrimination and Effect Sizes

Signal strength for each predictor is quantified using *signed* AUC and *signed* PR-AUC computed on *development* rows that are evaluable for each outcome to avoid look-ahead at panel edges.

Table 2 lists the five most informative features per outcome by signed PR-AUC; Table 1 provides complementary correlations. Because several financing-history variables are correlated, for example cumulative rounds, investors, and capital, magnitudes are not additive and should be read as stand-alone lift only.

Next financing (12 months). Days since last round and age are the two dominant stand-alone signals (AUCs ≈ 0.69 ; PR-AUCs ≈ 0.27 –0.31, both negative), followed by short-horizon momentum, rounds and funding in the last four quarters, and accumulated exposure, cumulative investors. This confirms that recent activity and firm youth capture most of the univariate lift for near-term fundraising.

Patent growth (24 months). Mean reversion is first-order. Larger patent and citation stocks today imply lower odds of expansion within two years (AUCs ≈ 0.68 –0.70, negative). Age enters with a strong negative sign, indicating that younger firms add patents, and cumulative rounds contributes additional negative discrimination. Short-run funding momentum plays little role.

Exit (36 months). Cumulative investors, cumulative rounds, and cumulative capital raised rank highest (AUCs ≈ 0.64 –0.66, positive), consistent with exits later in the financing lifecycle. Days since last round remains negative (AUC ≈ 0.61). Intellectual-property stocks show mild positive associations in the correlations (Table 1) but do not appear among the top-five predictors by signed PR-AUC (Table 2).

Overall, the diagnostics indicate that next-round fundraising is largely a recency and momentum phenomenon, patent growth is strongly mean-reverting and concentrated among younger firms, and exits are tied to accumulated financing maturity. In the next subsection, it is shown that multivariate models preserve these qualitative patterns while improving discrimination through non-linear interactions.

4.3 Leakage-Safe Imputation and Feature Curation

The modeling feature space was completed by estimating a simple, leakage-safe imputation step on the *development* period (2010–2019) and applying it unchanged to the *holdout* (2020–2021) and *final* (2022–2023) windows. The imputer is median-based, fit after replacing infinities with NaN, and it preserves all platform identifiers. To make missingness

informative, explicit indicators were added for any feature with at least 10% missing values in development, plus a forced flag for days since last round. Consistent with the diagnostics in §4.1, only days since last round exceeds this threshold, while age is about 6% and does not, so a binary days-since-last-round-is-missing indicator is included in the modeling set.

Predictors were chosen using the univariate screen in §4.2, supplemented by a small set of always-keep engineered variables that capture recency and capitalization. Table 3 lists the features passed to the models for each outcome, saved in the persisted feature list and used for column alignment across splits, including the missingness indicator noted above. This procedure fixes the set used in §4.5 and ensures that preprocessing parameters are estimated once on development and then reused without modification, which avoids target leakage.

4.4 Class Imbalance Handling

Class-imbalance variants were constructed only on the *development* period (2010–2019), and the *holdout* (2020–2021) or *final* (2022–2023) samples were never modified. This guards against leakage by design. For each dependent variable, five training versions were created: a weights-only baseline using inverse-prevalence weights, random oversampling (ROS), SMOTE-NC, Borderline-SMOTE, and ADASYN. For SMOTE-NC, the categorical mask includes only the missingness indicator for *days since last round*, and all other features are treated as numeric. All features in Table 3 are numeric, and no industry or geography dummies are included. Consequently, the only categorical input to SMOTE-NC is the binary NA-indicator for *days since last round*.

Table 4 reports class prevalence and whether each method runs on the panel. Training prevalence is moderate for funding and patent growth (17.6% and 21.9%) and low for exits (5.9%). SMOTE-NC balances each outcome to roughly 50:50 as intended, as shown in the "SMOTE_NC N (balanced)" column. Borderline-SMOTE and ADASYN fail on this dataset due to residual NaN values in a small number of engineered features and therefore revert to ROS, which is not used in the main analysis.

To keep the main analysis focused and reproducible, two imbalance treatments per outcome were carried forward into model development: the weights-only baseline and SMOTE-NC balanced 50:50, exactly as summarized in the rightmost column of Table 4. ROS and the Borderline-SMOTE or ADASYN fallbacks are used only for sensitivity checks outside the main text. This choice trades breadth for stability, predictable runtime, and a clean separation between preprocessing and modeling.

All outcomes are defined strictly forward from the quarter boundary with horizons $h \in 12, 24, 36$ months. Rows too close to panel edges for full-horizon evaluation are marked non-evaluable and excluded from that task. Temporal splits are non-overlapping, with *development* (2010-2019), *holdout* (2020-2021), and *final* (2022-2023). All preprocessing, infinite-to-NaN, NA indicators, and development-fitted medians, is estimated once on development and applied unchanged to later splits. Class weights and any resampling, SMOTE-NC or ROS, are confined to development, and evaluation splits are never reweighted or resampled.

4.5 Model Development and Selection

For each dependent variable, five model families were trained: logistic regression, Random Forest, XGBoost, LightGBM, and CatBoost. Each was estimated under two leakage-safe imbalance variants from §4.4, inverse-prevalence weights and SMOTE-NC.

Training uses the *development* period (2010–2019), with features and medians fixed in §4.3 through the persisted feature list and development-fitted medians. Evaluation is strictly out of time, the *final* window (2022–2023) when the label is observable, otherwise the *holdout* window (2020–2021). Headline metrics are PR-AUC as the primary criterion, AUROC, and Brier score, with sample sizes and base rates shown alongside results. By construction, only the 12-month funding outcome is evaluable in the final window. The 24-month patent and 36-month exit outcomes are evaluated on the holdout window.

Table 5 is the single source of truth for results. (i) Funding (12 months). On the final window, the LightGBM model trained with weights attains AUROC = 0.817, PR-AUC

= 0.220, and Brier = 0.144. (ii) Patent growth (24 months). With only the holdout window evaluable, the weights-only Random Forest delivers PR-AUC = 0.631 and AUROC = 0.921 (Brier = 0.074). (iii) Exit (36 months). On holdout, the weights-only Random Forest yields PR-AUC = 0.559 and AUROC = 0.872 (Brier = 0.032).

Within the pre-specified variants, weights and SMOTE-NC, Random Forest dominates gradient-boosting and logistic baselines on patent growth and exits, while *LightGBM with inverse-prevalence weights* is selected for near-term fundraising. These winners are carried forward to §4.6 for interpretability and calibration diagnostics and to §4.7 for out-of-time scoring summaries.

4.6 Model Analysis and Interpretability

The winning specification for each outcome is evaluated using three diagnostics: global feature importance; reliability via calibration curves and the Brier score; and partial dependence for the most influential predictors. All diagnostics are strictly out of time. Summary performance metrics are reported in Table 5, and interpretability artifacts and reliability are visualized in Figures 2–10.

For the funding winner, LightGBM with weights, *TreeSHAP* is computed on capped samples. For the patent and exit winners, Random Forest with weights, *feature_importances_* based on Gini importance and, where applicable, permutation importance scored by average precision are reported, and partial dependence is provided for top features. Calibration is assessed with quantile-binned reliability curves and the Brier score. Where curves show optimism, post hoc isotonic recalibration is recommended, with ranking unaffected.

Funding within 12 months (winner: LGBM, weights). On the *final* window, AU-ROC = 0.817, PR-AUC = 0.220, and Brier = 0.144 (Table 5). The SHAP profile in Figure 2 is led by *days since last round* and *age*, followed by *cumulative capital raised* and *cumulative investors*. Partial dependence in Figure 4 indicates that the probability of another round

decreases as time since the last round increases and with firm age, and increases with recent financing intensity and with cumulative funding and investor breadth, with concave gains at higher levels. The calibration curve in Figure 3 shows optimism across most of the score range, strongest at higher predicted probabilities, so post hoc isotonic scaling would improve probability calibration without affecting ranking.

Patent growth within 24 months (winner: RF, weights). On the holdout window, AUROC = 0.921, PR-AUC = 0.631, and Brier = 0.074 (Table 5). Global importance, Figure 5, identifies firm age, time since last round, and cumulative capital raised as the dominant predictors, with patent and citation stocks contributing at secondary magnitudes. The partial-dependence panels in Figure 7 show economically plausible associations: predicted patent growth declines with existing IP stock and with time since last funding, increases with cumulative and recent funding intensity, and is higher for younger firms. The reliability curve in Figure 6 is close to the 45° line across most of the range, with slight underconfidence in the top bin. Isotonic recalibration is advisable if calibrated probabilities are required.

Exit within 36 months (winner: RF, weights). On the holdout window, AUROC = 0.872, PR-AUC = 0.559, and Brier = 0.032 (Table 5). Global importance, Figure 8, indicates that time since last round, age, and cumulative capital raised are the leading predictors, with investor breadth, cumulative rounds, and IP stocks contributing at secondary levels. Partial dependence in Figure 10 shows that exit likelihood decreases as time since the last round increases and increases with cumulative capital raised and investor breadth; it also increases with age. IP variables exhibit weaker, partly non-monotonic but generally positive associations at higher levels. Calibration in Figure 9 is close to well calibrated at low to mid scores with mild optimism in the right tail, so scores are best used for ranking unless probabilities are re-calibrated.

Notes on importance. For the funding tree model, LightGBM, *TreeSHAP* is used. For the patent and exit winners, Random Forest, *feature_importances_* based on Gini importance

are reported, and where permutation or average-precision scoring is used, the figure and caption state it explicitly. Across outcomes, financing recency and momentum are consistently informative. Innovation stock is strongly predictive of subsequent patent growth and only modestly informative for near-term fundraising. Exit prediction moves with accumulated financing maturity. Ranking is useful, and calibrated probabilities can be obtained via post hoc isotonic scaling when needed.

4.7 Out-of-time scoring and ranked target lists

This step applies the winning specification for each outcome to the most recent cohort for which the prediction horizon is fully evaluable. Columns are aligned using the persisted feature list, infinities are replaced with NaN, and missing entries are filled with development-period medians. No resampling or reweighting is applied to evaluation data. Predictions \hat{p} are converted to integer ranks and percentiles. A single row per organization is retained after deduplication by keeping the highest score and the most recent quarter when ties occur. Table 7 summarizes winners, selection splits, and scored cohorts. Performance metrics used to select the winner come from Table 5, with capacity-style checkpoints in Table 6, and are not re-estimated here.

The ranked lists are intended for screening and benchmarking. Given that Section 4.6 documents discrimination and calibration, only a manifest of the scored cohorts is reported in the main text. If calibrated probabilities are required for decision-making, isotonic recalibration can be applied post hoc without changing ranking.

The scoring step translates the validated models into actionable target lists. Cohort choices reflect the horizon length, 2022 for financing within 12 months, 2021 for patent growth within 24 months, and 2020 for exits within 36 months. These outputs enable practical prioritization while preserving the leakage controls defined in Section 3.3.

5 Conclusion

This paper examined whether near-term funding, medium-horizon patent growth, and longer-horizon exits for startups can be predicted in a leakage-safe and interpretable way using routinely available information. A firm-quarter panel was built from Crunchbase and merged to USPTO assignee data. A pipeline was designed to train only on a development window and to evaluate strictly out of time on non-overlapping holdout and final windows. Model selection emphasized ranking performance under class imbalance, and post-estimation diagnostics paired discrimination with interpretability and reliability.

Three findings emerge. First, patent expansion over a two-year horizon was highly predictable at the firm-quarter level on the holdout window. A Random Forest with inverse-prevalence weights delivered strong discrimination, and interpretation based on feature importance and partial dependence indicated maturity and financing recency as central drivers, consistent with views of patents as protective assets and as credible signals that ease financing frictions (Conti et al., 2013; Hsu and Ziedonis, 2013; Farre-Mensa et al., 2020). Second, next funding within 12 months was best captured by a weighted LightGBM on the final window. Explanations indicate pronounced roles for time since last round, age, and cumulative financing intensity, consistent with evidence that venture outcomes reflect momentum in investor attention and organizational maturity (Gompers et al., 2010; Kortum and Lerner, 2000). Third, exit predictions at 36 months again favored Random Forests with weights, and the most important predictors relate to the depth and breadth of prior financing, with intellectual property stocks contributing positively but secondarily, consistent with trajectories in which capital structure and investor networks shape eventual liquidity events.

Methodologically, the paper contributes an interpretable ML template for innovation finance. Development-only preprocessing was used, including infinite-to-NaN, missing flags, and median imputation. Persisted feature lists ensured column alignment, and clean temporal splits avoided look-ahead. The comparative model zoo across linear and tree-based learners under inverse-prevalence weights and SMOTE-NC follows best practice in explain-

able prediction (Breiman, 2001; Kim et al., 2024). SHAP and partial dependence provided narratives that link scores to economically meaningful margins (Lundberg and Lee, 2017). Calibration analysis highlights when isotonic scaling is advisable before probabilities are used for decision thresholds, while ranking use cases remain robust.

Limitations suggest clear extensions. Platform coverage and survivorship in Crunchbase can introduce selection and measurement error (Dalle et al., 2017). Expanding sources or auditing against administrative records would test robustness. The focus here is on structured variables. Incorporating text from patents, company descriptions, and filings could enrich signals, provided interpretability is preserved. Although strict time splits mitigate leakage, regime shifts, including post-pandemic funding cycles, can alter base rates and feature distributions. Formal drift monitoring and periodic recalibration would support deployment. Finally, outcomes are coarse, defined as any funding, any patent growth, or any exit. Future work could model timing with survival methods, explore heterogeneity by sector or technology class (Kogan et al., 2017), and evaluate counterfactual policy experiments under explainable learners.

In sum, leakage-safe and interpretable ML was shown to yield ranking-useful predictions for startup funding, patenting, and exits, while recovering mechanisms that align with established theory and evidence. The pipeline's emphasis on transparency, through time-clean evaluation and economically legible explanations, makes it suitable for scholarly replication, investor screening, and policy monitoring. Extending the framework with richer data modalities and drift-aware maintenance promises greater practical relevance for innovation-finance research.

References

- Antretter, T., Blohm, I., Grichnik, D., and Wincent, J. (2019). Predicting new venture survival: A twitter-based machine learning approach to measuring online legitimacy. *Journal of Business Venturing Insights*, 11:e00109.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
- Conti, A., Thursby, M., and Thursby, J. (2013). Patents as signals for startup financing. Journal of Industrial Economics, 61(3):592–622.
- Dalle, J.-M., den Besten, M., and Menon, C. (2017). Using crunchbase for economic and managerial research. OECD Science, Technology and Industry Working Papers 2017/08, OECD Publishing.
- Eggers, J. P. and Song, L. (2015). Dealing with failure: Serial entrepreneurs and the costs of changing industries between ventures. *Academy of Management Journal*, 58(6):1785–1803.
- Farre-Mensa, J., Hegde, D., and Ljungqvist, A. (2020). What is a patent worth? evidence from the u.s. patent "lottery". *Journal of Finance*, 75(2):639–682.
- Gompers, P., Gornall, W., Kaplan, S. N., and Strebulaev, I. A. (2020). How do venture capitalists make decisions? *Journal of Financial Economics*, 135(1):169–190.
- Gompers, P., Kovner, A., Lerner, J., and Scharfstein, D. (2010). Performance persistence in entrepreneurship. *Journal of Financial Economics*, 96(1):18–32.
- Hsu, D. H. and Ziedonis, R. H. (2013). Resources as dual sources of advantage: Implications for valuing entrepreneurial-firm patents. *Strategic Management Journal*, 34(7):761–781.
- Kim, M., Benabderrahmane, S., and Rahwan, T. (2024). Interpretable machine learning models for predicting the next targets of activist funds. *The Journal of Finance and Data Science*, 10.
- Kogan, L., Papanikolaou, D., Seru, A., and Stoffman, N. (2017). Technological innovation, resource allocation, and growth. *Quarterly Journal of Economics*, 132(2):665–712.
- Kortum, S. and Lerner, J. (2000). Assessing the contribution of venture capital to innovation. *RAND Journal of Economics*, 31(4):674–692.
- Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (NeurIPS).
- McKinney, W. (2010). Data structures for statistical computing in python. In van der Walt, S. and Millman, J., editors, *Proceedings of the 9th Python in Science Conference (SciPy 2010)*, pages 51–56.

Figures

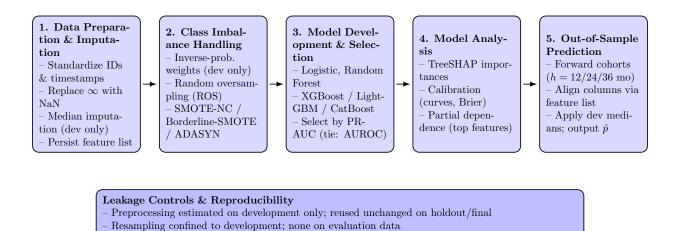


Figure 1: End-to-end data processing and modeling pipeline. Steps 1–5 summarize the sequential workflow from data preparation to out-of-sample prediction. The lower ribbon highlights leakage controls and reproducibility safeguards, including development-only preprocessing, resampling confined to training data, and deterministic artifacts for replication.

Deterministic filenames; saved leaderboards, figures, models

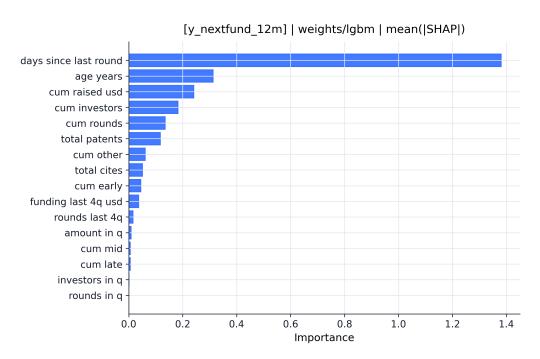


Figure 2: Funding (12m), LightGBM trained with inverse-prevalence weights. Bars show global feature importance measured by mean absolute SHAP values (mean(|SHAP|)) on the final evaluation window. Financing recency and firm age dominate, followed by cumulative capital raised and investor breadth.

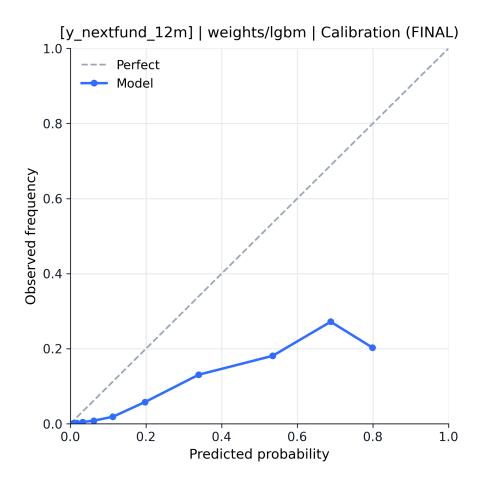


Figure 3: Funding (12m), LightGBM trained with inverse-prevalence weights. Quantile-binned calibration curve on the *final* window. The model shows optimism at higher predicted probabilities relative to the 45° reference line, suggesting that isotonic recalibration could improve probability calibration without affecting ranking.

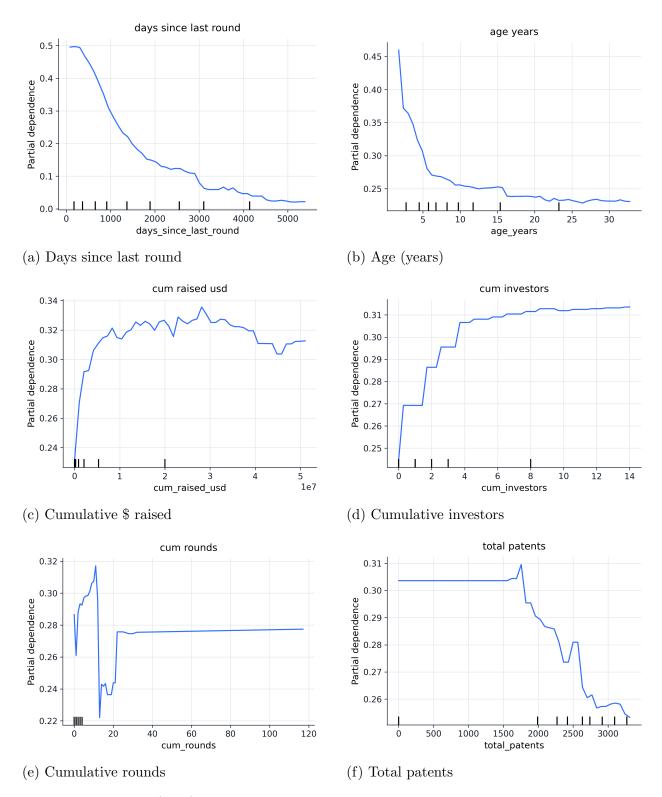


Figure 4: Funding (12m), LightGBM trained with inverse-prevalence weights. Partial dependence plots on the *final* window for the top six predictors. The probability of next-round funding decreases with time since the last round and firm age, and rises with cumulative funding, investor breadth, and round count. Higher patent stock shows a mild negative effect. *Note:* Uncalibrated probabilities appear overconfident at high scores, but ranking remains unaffected.

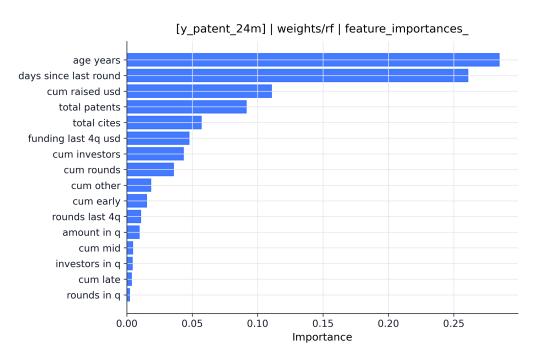


Figure 5: Patent growth (24m), Random Forest trained with inverse-prevalence weights. Bars show global feature importance based on mean decrease in impurity (MDI, Gini). Firm age, time since last round, and cumulative capital raised dominate the ranking, followed by patent and citation stocks. *Note:* MDI can overemphasize high-cardinality or correlated predictors.

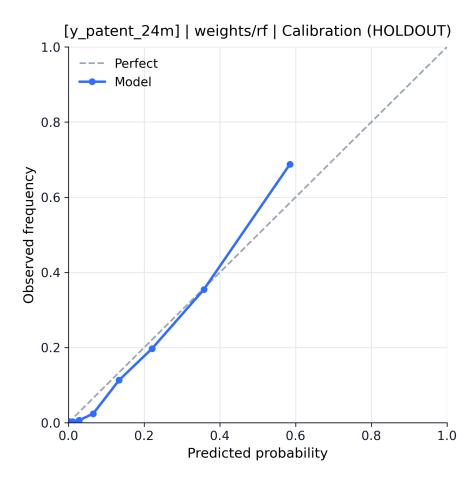


Figure 6: Patent growth (24m), Random Forest trained with inverse-prevalence weights. Quantile-binned calibration curve on the *holdout* window. The model aligns closely with the 45° reference line, showing well-calibrated predicted probabilities across most of the range.

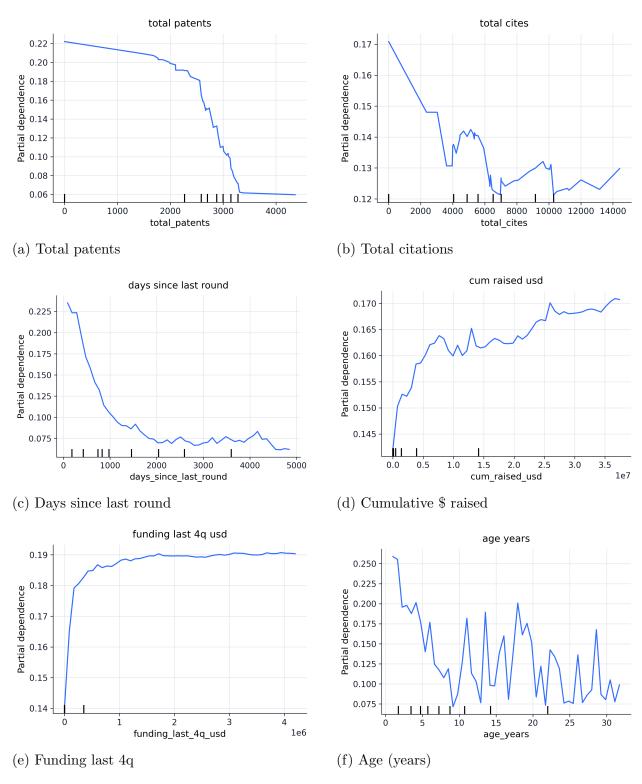


Figure 7: Patent growth (24m), Random Forest trained with inverse-prevalence weights. Partial dependence plots on the *holdout* window for six key predictors. The probability of patent expansion decreases with existing patent and citation stocks and with time since the last round, and increases with cumulative and recent funding. Younger firms show higher predicted probabilities, consistent with maturity and mean-reversion effects.

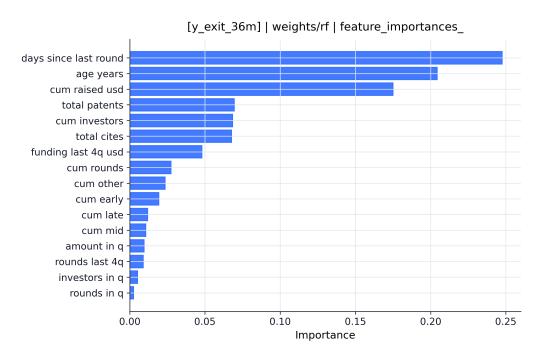


Figure 8: Exit (36m), Random Forest trained with inverse-prevalence weights. Bars show global feature importance based on mean decrease in impurity (MDI, Gini). Exit predictions are driven by financing maturity measures such as time since last round, firm age, and cumulative capital raised, with intellectual property variables contributing positively but secondarily. *Note:* MDI may overweight features with higher cardinality or correlation.

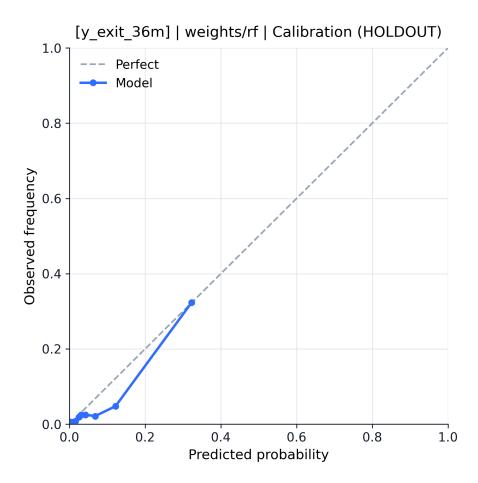


Figure 9: Exit (36m), Random Forest trained with inverse-prevalence weights. Quantile-binned calibration curve on the *holdout* window. The model follows the 45° reference line closely at low and mid probability ranges, indicating good calibration, with mild optimism at higher predicted probabilities.

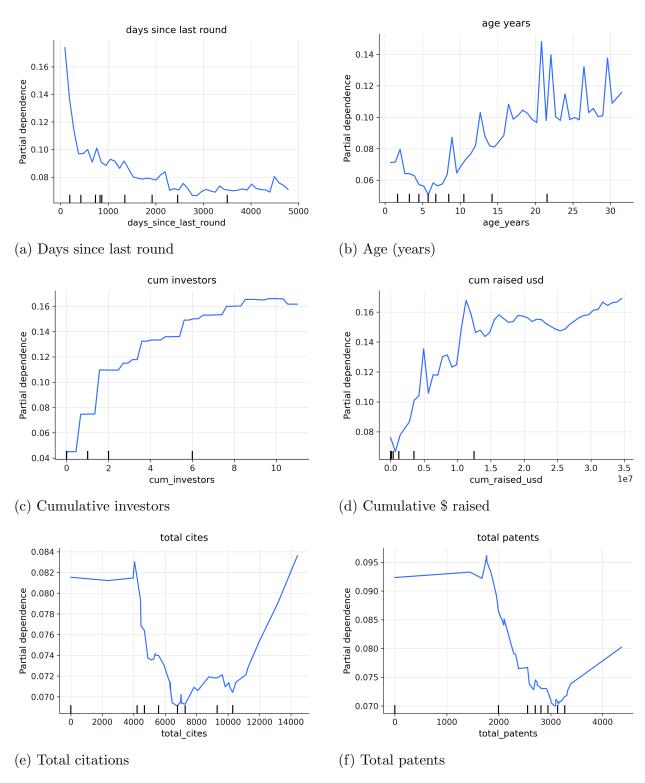


Figure 10: Exit (36m), Random Forest trained with inverse-prevalence weights. Partial dependence plots on the *holdout* window for six key predictors. Exit likelihood increases with cumulative investors and total capital raised, and decreases with time since last round. Age and intellectual property variables show weaker but consistent positive associations.

Tables

Table 1: Pairwise correlations between selected firm-level features and three startup outcomes on the development panel (2010–2019). Correlations are computed at the firm-quarter level after preprocessing. Positive values indicate features associated with a higher likelihood of the outcome, while negative values indicate inverse relationships. Patterns reflect financing recency and momentum for funding, mean reversion in patenting, and maturity effects for exits. Correlations are linear and may differ in sign from partial dependence shapes when effects are non-linear or interact with other variables (e.g., age in Exit).

Feature	Next funding (12m)	Patent growth (24m)	Exit (36m)
Total patents	-0.049	-0.294	0.078
Total citations	-0.056	-0.252	0.072
Days since last round	-0.222	-0.195	-0.071
Cumulative investors	0.114	-0.039	0.162
Rounds in last four quarters	0.163	0.006	0.078
Cumulative rounds	0.065	-0.138	0.132
Cumulative mid-stage rounds	0.064	-0.007	0.126
Cumulative late-stage rounds	0.052	-0.004	0.130
Cumulative early-stage rounds	0.051	-0.097	0.067
Cumulative other-stage rounds	0.025	-0.123	0.074
Capital raised in last 4 quarters	0.028	-0.002	0.040
Cumulative capital raised (USD)	0.029	-0.013	0.068
Capital raised in current quarter	0.014	-0.002	0.022
Rounds in current quarter	0.080	0.005	0.036
Investors in current quarter	0.060	0.010	0.045
Firm age (years)	-0.105	-0.090	-0.014

Table 2: Top univariate predictors for each outcome based on signed AUC and signed PR-AUC computed on training rows. Signs indicate the direction of association with the outcome (positive or negative). Reported values reflect stand-alone discrimination, not additive effects. The results show that fundraising depends on recency and momentum, patent growth is mean-reverting with respect to age and IP stock, and exits are linked to accumulated financing maturity.

Outcome	Feature	AUC (signed)	PR-AUC (magnitude)
Next financing (12m)	Age (years)	0.69(-)	0.31
	Days since last round	0.69(-)	0.27
	Rounds in last 4q	0.58 (+)	0.23
	Cumulative investors	0.54 (+)	0.22
	Funding in last 4q (USD)	0.57 (+)	0.22
Patent growth (24m)	Total patents	0.70 (-)	0.32
	Total citations	0.68(-)	0.31
	Age (years)	0.70 (-)	0.39
	Cumulative rounds	0.66(-)	0.30
	Days since last round	0.61(-)	0.27
Exit (36m)	Cumulative investors	0.66 (+)	0.12
	Cumulative rounds	0.64 (+)	0.10
	Cumulative capital raised (USD)	0.65 (+)	0.14
	Days since last round	0.61 (-)	0.08
	Late-stage rounds (cumulative)	0.54 (+)	0.08

Table 3: Final feature sets used for model training by outcome, derived from the development period (2010–2019) after leakage-safe preprocessing and univariate screening. Each column lists the variables included in the persisted feature list used for downstream modeling and column alignment across splits. Features capture firm maturity, financing recency and momentum, and innovation intensity.

Next financing (12 months)	Patent growth (24 months)	Exit (36 months)
age days since last round rounds in last four quarters cumulative investors	age total patents total citations cumulative rounds	cumulative capital raised (USD) cumulative investors cumulative rounds capital raised in last four quarters (USD)
capital raised in last four quarters (USD)	days since last round	cumulative late-stage rounds
cumulative capital raised (USD) cumulative rounds total citations total patents cumulative early-stage rounds rounds this quarter investors this quarter capital raised this quarter (USD) cumulative mid-stage rounds cumulative other-stage rounds cumulative late-stage rounds	cumulative capital raised (USD) cumulative other-stage rounds cumulative investors cumulative early-stage rounds rounds in last four quarters investors this quarter cumulative mid-stage rounds rounds this quarter cumulative late-stage rounds capital raised this quarter (USD) capital raised in last four quarters (USD)	days since last round cumulative mid-stage rounds cumulative other-stage rounds cumulative early-stage rounds rounds in last four quarters total citations total patents capital raised this quarter (USD) investors this quarter rounds this quarter age

Table 4: Class imbalance and resampling feasibility on the development period (2010–2019). The table reports sample sizes, positive-class prevalence, and balanced sample counts under SMOTE-NC for each outcome. Only inverse-prevalence weighting and SMOTE-NC were carried forward into the main analysis; other resampling variants (ROS, Borderline-SMOTE, ADASYN) were retained for robustness checks but not used in final models.

Outcome	Train N	Positives	Prevalence	SMOTE_NC N (balanced)	Methods carried forward
Next financing (12m) Patent growth (24m) Exit (36m)	2,495,444	439,009	0.176	4,112,870	Weights; SMOTE-NC
	2,495,444	547,553	0.219	3,895,782	Weights; SMOTE-NC
	2,495,444	147,395	0.059	4,696,098	Weights; SMOTE-NC

Note: Borderline-SMOTE and ADASYN error on NaNs and fall back to ROS; ROS saved but not used in main text. Weights-only positive class weights (uncapped): 4.69 (12m), 3.56 (24m), 15.93 (36m); capped at 50 if needed.

Table 5: Winning model and out-of-time performance for each prediction task. Each outcome is evaluated on its corresponding non-overlapping split, with the positive-class base rate shown for reference. Reported metrics include AUROC and PR-AUC, with PR-AUC used as the main selection criterion. All reported winners use development-only preprocessing and inverse-prevalence weighting.

Outcome	Split (N)	Base rate	Model (imbalance)	AUROC	PR-AUC	Brier
Next funding (12m)	Final 2022–2023 (281,326)	0.088	LGBM (weights)	0.817	0.220	0.144
Patent growth (24m)	Holdout 2020–2021 (644,765)	0.139	RF (weights)	0.921	0.631	0.074
Exit (36m)	Holdout 2020–2021 (273,147)	0.048	RF (weights)	0.872	0.559	0.032

Table 6: Precision at top-K predictions (P@K) for each outcome on the corresponding evaluation split. Values reflect the proportion of true positives among the top-ranked predictions, compared to the base rate in the evaluated sample. Rankings are computed globally with deterministic tie-breaks. Precision may vary with K.

Outcome (split)	P@30	P@100	P@500	Base rate
Next funding (12m) (Final)	0.200	0.290	0.282	0.088
Patent (24m) (Holdout)	0.933	0.800	0.816	0.139
Exit (36m) (Holdout)	0.967	0.980	0.978	0.048

Table 7: Summary of final out-of-time scoring results for each prediction task. Each model was applied to the most recent evaluable cohort, using features and preprocessing fixed on the development period. Reported metrics correspond to the split used for model selection. Output files include organization identifiers, descriptors, predicted probabilities, ranks, and percentiles, with one record per firm after deduplication. Rankings can be used directly, while calibrated probabilities require post hoc isotonic scaling. Selection metrics are repeated from Table 5; no re-estimation is performed. For exits, the evaluable holdout rows are concentrated in 2020 due to the 36-month horizon.

Outcome	Winner (imbalance; model)	Selection split and metrics	Scored cohort	Rows
Next funding (12 months) Patent growth (24 months) Exit (36 months)	LGBM (weights) RF (weights) RF (weights)	Final: PR-AUC 0.220, AUROC 0.817 Holdout: PR-AUC 0.631, AUROC 0.921 Holdout: PR-AUC 0.559, AUROC 0.872	2022 2021 2020	281,326 280,249 273,147