Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00254-0

Original Article

Predicting Startup Outcomes Using Explainable Machine Learning and Y Combinator-Inspired Feature Engineering

Siddharth Gupta [1], Pratham Namdev [1], Shubham Nagar [1], Sunny Singh [1], Anjali Deshwal [1]

1. *Computer Science and Engineering, Greater Noida Institute of Technology, Greater Noida, IND*

Received: June 05, 2026 | Review began: June 22, 2026 | Review ended: August 11, 2026 | Published: September 01, 2026

Abstract

Early-stage startup success is notoriously difficult to predict, yet the stakes for getting it right - whether for investors, accelerators, or founders themselves - are substantial. Most existing machine learning approaches lean heavily on generic features from Crunchbase or PitchBook, and in doing so tend to miss a category of arguably more informative signals: the domain-specific structural properties of elite accelerator programs like Y Combinator (YC). This paper introduces a YC-Inspired Feature Engineering (YIFE) framework that incorporates batch cohort timing, founder technical depth (proxied via public GitHub activity), team composition, industry category, and geographic cluster alongside conventional funding and operational variables. We evaluate five classifiers - Logistic Regression, Random Forest, XGBoost, Support Vector Machine, and a Multilayer Perceptron - on a curated dataset of 4,323 YC-funded companies spanning 2005-2024 using a temporal train-test design. XGBoost with YIFE achieves the strongest performance, with an F1-score of 0.85 and an area under the receiver operating characteristic curve of 0.91 on the held-out W21-S24 cohort, representing a 19-percentage-point F1 improvement over a generic Crunchbase baseline (0.85 vs. 0.66) and an 8-percentage-point improvement over a replicated literature baseline (0.85 vs. 0.77). SHAP (SHapley Additive exPlanations)-based interpretation identifies funding-related variables, batch-year context, and team size among the most influential model features, while prior FAANG (Facebook (now Meta), Amazon, Apple, Netflix, and Google) experience contributes comparatively little within the available founder-profile sample. Because several funding-related predictors may only be observable after the initial accelerator stage and overlap conceptually with the outcome definition, the framework is best interpreted as a domain-contextualized retrospective classification approach for the YC ecosystem rather than a strict ex-ante forecasting tool. These findings suggest the potential value of accelerator-specific feature engineering while motivating future validation under fixed prediction horizons and across non-YC accelerator ecosystems. The analytical code and pipeline are made publicly available to support this and other future work in computational entrepreneurship and venture analytics.

Categories: Machine Learning (ML), Explainable AI, Predictive Analytics

Keywords: machine learning, feature engineering, xgboost, y combinator, entrepreneurship analytics, startup outcome prediction, shap, venture capital, explainable artificial intelligence, predictive analytics

Introduction

The global startup ecosystem has grown substantially in scale and complexity over the past two decades. Y Combinator (YC), founded in 2005, has emerged as one of the world's most influential startup accelerators, having funded thousands of early-stage technology ventures across sequential batch cohorts [1]. Despite the competitive selectivity of elite accelerator programs, portfolio companies exhibit highly divergent trajectories, with a significant proportion failing to

achieve follow-on venture financing or a liquidity event. Understanding what separates successful YC graduates from those that stagnate or shut down has significant practical value for venture capital (VC), entrepreneurship policy, and early-stage investment strategy.

Machine learning (ML) has been increasingly applied to predict startup outcomes using large-scale databases such as Crunchbase and PitchBook [2-4]. Prior research has shown that tree-based ensemble methods, particularly Random Forest and Gradient Boosting, achieve strong performance on such tasks [3,5], with studies also comparing ML approaches against human investor decision-making [6]. That said, existing work has two recurring blind spots. First, feature sets tend to be generic and fail to incorporate domain-specific signals that are especially informative for accelerator-backed startups. Second, few studies have treated the structural properties of the YC dataset itself - batch cohort timing, for instance, which implicitly encodes market cycle conditions - as predictive inputs.

Despite the growing use of ML in startup analytics, limited research has systematically examined whether accelerator-specific structural signals provide predictive information beyond conventional startup database features while simultaneously offering interpretable explanations of model behavior. This paper addresses both gaps through the YC-Inspired Feature Engineering (YIFE) framework.

Digital transformation has significantly expanded the availability of structured entrepreneurial data, enabling researchers and investors to move beyond conventional expert judgment toward evidence-based, data-driven evaluation of venture potential [7]. Modern startup ecosystems increasingly rely on predictive analytics to reduce investment uncertainty and identify early-stage momentum; the YIFE framework builds on this broader shift by systematically extracting accelerator-specific signals from publicly accessible digital traces.

An additional challenge is that startup prediction models may not generalize uniformly across accelerator ecosystems. Programs such as Y Combinator, Techstars, Seedcamp, and 500 Global differ substantially in admission criteria, geographic concentration, sector specialization, mentorship structure, cohort size, and follow-on investor networks. Consequently, relationships learned from a broad startup database - or from a single accelerator - may partly reflect ecosystem-specific selection mechanisms rather than universal determinants of entrepreneurial success. By embedding structural accelerator properties directly into the feature architecture, our framework provides a tailored analytical lens for the YC portfolio, while conclusions drawn from it must be interpreted within the ecosystem from which the data originate.

The work makes three specific contributions. We introduce a novel set of domain-specific features drawn from the public YC company and founder dataset, including batch-year market-cycle encoding, founder GitHub activity as a proxy for technical depth, and team-size dynamics. We then conduct a comparative evaluation of five ML classifiers on 4,323 YC companies spanning 2005-2024, using cross-validation for hyperparameter tuning and model selection within the training cohorts, and precision, recall, F1-score, and an area under the receiver operating characteristic curve (AUROC) for final evaluation on a temporally held-out cohort. Finally, we apply SHAP (SHapley Additive exPlanations) to the best-performing model to produce interpretable insights into which variables are most associated with predictions of startup success in the YC context.

The following section surveys prior work across three streams: general ML-based startup prediction, YC-specific research, and feature engineering in entrepreneurial analytics.

Literature review

ML for Startup Success Prediction

The application of ML to predict startup outcomes has gained considerable traction since the proliferation of structured startup databases. Kim et al. [5] analyzed 218,207 companies from Crunchbase (2011-2021) using six classifier algorithms and found that media exposure, monetary funding, industry convergence level, and industry association level are the most significant predictors of startup success. Their study, however, defined success solely as achieving an initial public offering (IPO) - a definition that excludes the majority of successful exits via acquisition.

How to cite this article:
Gupta S, Namdev P, Nagar S, et al. (September 01, 2026) Predicting Startup Outcomes Using Explainable Machine Learning and Y Combinator-Inspired Feature Engineering. Cureus J Comput Sci 3 : es44389-026-00254-0. DOI https://doi.org/10.7759/s44389-026-00254-0

Razaghzadeh Bidgoli et al. [3] extended this line of work by combining classification with clustering, reporting that Random Forest and Gradient Boosting achieved accuracies of 82% and 80%, respectively, on a Crunchbase-derived dataset. They also employed SHAP and permutation-based feature importance methods, establishing a useful precedent for interpretable ML in this domain. Park et al. [8] addressed class imbalance - a persistent challenge in startup datasets where successful companies are the minority class - using Generative Adversarial Networks to synthesize minority-class samples, improving model fairness and reliability, building on earlier minority-oversampling approaches such as Synthetic Minority Over-sampling Technique [9].

On the interpretable ML frontier, Antretter et al. [4] investigated whether digital traces could be used to predict early-stage startup survival, demonstrating the potential value of digitally observable founder and venture characteristics for investor-oriented assessment. More recently, Maarouf et al. [10] proposed a fused large language model approach incorporating textual self-descriptions from VC platforms as predictive features, demonstrating that natural language signals carry value beyond structured data. Mashhadi et al. [11] developed an interpretable ML pipeline on a firm-quarter panel (2010-2023) merging Crunchbase with USPTO patent data, reporting strong predictive performance across next-round funding, patent-stock growth, and exit outcome prediction.

YC-Specific Research

Research specifically targeting the YC portfolio is limited but growing. Adl [12] constructed a dataset of 4,323 YC companies (2005-2024) merged with S&P Global funding data and estimated ordinary least squares regressions with batch-year fixed effects to examine founder-background effects. The most robust finding reported was that larger founding teams raise more capital, while prior FAANG (Facebook (now Meta), Amazon, Apple, Netflix, and Google) work experience was not a robust positive predictor of funding outcomes - a finding our work corroborates and extends into a predictive ML framework.

Recent advances in startup analytics increasingly incorporate richer technological, textual, and innovation-related signals alongside conventional financial and organizational variables [10,11]. This broader shift motivates our inclusion of industry orientation and batch-year context as potentially informative features within the YC ecosystem. To our knowledge, no prior study has combined YC-specific structural features (batch timing, team composition, founder GitHub activity) with a full ML pipeline comparison and SHAP interpretability in a unified framework, which is the core novelty of this work.

Feature Engineering in Entrepreneurial Analytics

Feature engineering remains a critical determinant of predictive performance in startup outcome models. Corea et al. [13] identified 21 relevant features from Crunchbase for predicting VC investment decisions, emphasizing social media presence, prior investor reputation, and funding amount. The interpretable ML literature [14-16] provides a theoretical foundation for pairing high-performing ensemble models with post-hoc explanation methods such as SHAP and Local Interpretable Model-agnostic Explanations. Grinsztajn et al. [17] offer empirical evidence for why tree-based models such as XGBoost continue to outperform deep learning approaches on typical tabular data, which is consistent with our own model comparison results. Dalle et al. [18] provided methodological guidance for using Crunchbase in economic and managerial research, informing our data pipeline. Our work builds on this tradition by introducing features specifically designed for the accelerator context.

Recent work from 2025-2026 further expands methodological approaches to startup outcome prediction. Font-Cot et al. [19] developed a multivariate AI approach to startup survival forecasting that combines empirical entrepreneurship knowledge with Lipschitz extensions, neural networks, and linear regression. More recently, Pan et al. [20] proposed a multi-layer decision-support system that integrates knowledge graphs, graph convolutional networks, and federated learning to predict startup success and assess risk on a large Crunchbase-derived dataset. Unlike these approaches, the present study focuses on a longer-horizon, cross-batch classification framework using accelerator-specific feature engineering and explainable tabular ML applied specifically to the YC ecosystem. These differences highlight

How to cite this article:
Gupta S, Namdev P, Nagar S, et al. (September 01, 2026) Predicting Startup Outcomes Using Explainable Machine Learning and Y Combinator-Inspired Feature Engineering. Cureus J Comput Sci 3 : es44389-026-00254-0. DOI https://doi.org/10.7759/s44389-026-00254-0

complementary methodological settings and further motivate research on fixed-horizon, leakage-controlled, and ecosystem-specific validation of YIFE. The comparative analysis of representative startup prediction studies and the proposed YIFE framework is presented in Table *1*.

How to cite this article:
Gupta S, Namdev P, Nagar S, et al. (September 01, 2026) Predicting Startup Outcomes Using Explainable Machine Learning and Y Combinator-Inspired Feature Engineering. Cureus J Comput Sci 3 : es44389-026-00254-0. DOI https://doi.org/10.7759/s44389-026-00254-0

Study	Dataset/Scope	ML Models	Explainability	Key Metric	Primary Limitation
Kim et al. [5]	Crunchbase, 218,207 firms	LR, RF, SVM, XGBoost, ANN	Feature importance	High accuracy on IPOs	Restricts success strictly to IPO exits
Razaghzadeh Bidgoli et al. [3]	Crunchbase-derived	RF, Gradient Boosting	SHAP, permutation importance	Accuracy: 82% (RF)	Relies on generic firm characteristics
Park et al. [8]	Crunchbase (GAN-augmented)	GAN + RF/XGBoost	Feature importance	Improved AUROC via oversampling	Synthetic-sample dependence
Maarouf et al. [10]	VC/startup textual data	Fused large language model	Model interpretation	Strong textual signal	Limited accelerator-specific structure
Mashhadi et al. [11] (Preprint)	Crunchbase + USPTO panel	LR, RF, XGBoost, LightGBM, CatBoost	SHAP, partial dependence	AUROC up to 0.92 (patenting)	Not accelerator-scoped
Adl [12] (Preprint)	YC portfolio (4,323 firms)	Ordinary Least Squares	Linear coefficients	Statistical significance only	No ML pipeline; funding-variation focus
Font-Cot et al. [19]	Multi-source startup survival data	Multivariate AI, neural networks, linear regression, Lipschitz extensions	Not SHAP-based	Survival prediction performance	Small startup evaluation sample; not YC-scoped
Pan et al. [20]	Crunchbase-derived data; 20-startup evaluation sample	Knowledge graph + GCN + DNN + federated learning	SHAP/feature-importance analysis	Accuracy 93%; AUC 0.90	Different data/model setting from YC-specific tabular classification
This study (YIFE framework)	YC portfolio (4,323 firms)	LR, RF, XGBoost, SVM, MLP	SHAP feature attribution	F1 = 0.85, AUROC = 0.91	YC-scoped; retrospective funding signals

TABLE 1: Comparison of Representative Startup Prediction Studies and the Proposed YIFE Framework

Comparative summary of representative ML studies for startup success prediction, highlighting the datasets, ML models, explainability methods, key performance metrics, methodological limitations, and the distinguishing characteristics of the proposed YIFE framework.
Preprint indicates studies that had not completed formal peer review at the time of writing.

How to cite this article:
Gupta S, Namdev P, Nagar S, et al. (September 01, 2026) Predicting Startup Outcomes Using Explainable Machine Learning and Y Combinator-Inspired Feature Engineering. Cureus J Comput Sci 3 : es44389-026-00254-0. DOI https://doi.org/10.7759/s44389-026-00254-0

LR, Logistic Regression; RF, Random Forest; SVM, Support Vector Machine; ANN, Artificial Neural Network; GAN, Generative Adversarial Network; GCN, Graph Convolutional Network; DNN, Deep Neural Network; SHAP, SHapley Additive exPlanations; AUC, Area Under the Curve; AUROC, Area Under the Receiver Operating Characteristic Curve; VC, Venture Capital; USPTO, United States Patent and Trademark Office; YIFE, YC-Inspired Feature Engineering; ML, Machine Learning; IPO, Initial Public Offering; YC, Y Combinator; MLP, Multilayer Perceptron

Materials And Methods

Dataset and feature engineering

Data Sources

Our dataset is constructed from three complementary public sources, as summarized in Table *2*.

Source	Records	Features Used	Access Method
YC Company Directory (Kaggle)	5,000+ companies (2005–2024)	Batch, category, status, team size, location	Public dataset (Kaggle)
Crunchbase Open Data Map	Subset of YC companies	Funding rounds, total raised, lead investors	Free tier/academic access
GitHub Public API	Sampled founder profiles	Public repos, commit frequency, language diversity	REST API (public)
LinkedIn (Manual Sample)	500 founder profiles	Education, prior employer, co-founder count	Public profiles

TABLE 2: Data Sources and Feature Summary

Data collected from publicly available sources. LinkedIn data represents a manual sample of 500 founder profiles only.
YC, Y Combinator; API, Application Programming Interface

The final curated dataset comprises 4,323 YC-funded companies from Winter 2005 (W05) through Summer 2024 (S24), consistent with the scope used by Adl [12]. Companies without publicly verifiable outcome data were excluded from the labeled set.

Artificial Intelligence Tools Declaration

Large language model-based tools were used to assist with manuscript drafting, paraphrasing, structural editing, and preparation of a text-based schematic workflow diagram (Figure *1*) during the preparation of this article. The tools used include Claude (Anthropic), ChatGPT (OpenAI), Gemini (Google), Grok (xAI), Perplexity AI, and DeepSeek. All scientific content, experimental design, data analysis, interpretations, and conclusions were independently verified and approved by the authors. No AI tool was used to generate or alter the original experimental data, the reported numerical results, or the data-derived plots (Figures *2*-*3*, which depict the authors' own ROC and SHAP analysis outputs). This declaration is made in accordance with Cureus journal policy on artificial intelligence use, which distinguishes simple schematic/flow-chart media from generative imagery of experimental content.

Success Label Definition

How to cite this article:
Gupta S, Namdev P, Nagar S, et al. (September 01, 2026) Predicting Startup Outcomes Using Explainable Machine Learning and Y Combinator-Inspired Feature Engineering. Cureus J Comput Sci 3 : es44389-026-00254-0. DOI https://doi.org/10.7759/s44389-026-00254-0

How to operationalize "success" is a key methodological decision in startup prediction research. IPO-only definitions (e.g., Kim et al. [5]) are overly restrictive, excluding successful acquisitions and high-growth private companies. We adopt a multi-criteria definition: a company is labeled successful if it (a) raised a Series A round or beyond post-YC, (b) was acquired at a disclosed valuation, or (c) remained in active operation with measurable growth beyond three years post-batch. The composite definition was selected because startup success is inherently multidimensional and cannot be represented adequately by IPO status alone: Series A+ financing captures demonstrated investor confidence and continued financing capacity, acquisition represents a realized exit outcome, and sustained operation with observable growth captures ventures that create economic value without immediately pursuing either outcome. These three criteria represent distinct trajectories rather than interchangeable events; accordingly, the resulting binary label should be interpreted as a broad operational measure of favorable startup outcome rather than a single homogeneous endpoint. Under this definition, approximately 45% of YC companies in our dataset are labeled successful. The remaining 55% are labeled inactive, shut down, or stagnant, yielding a mildly imbalanced classification problem addressed via class-weight balancing for Logistic Regression, Random Forest, and Support Vector Machine, and via a positive-class scaling weight for XGBoost (Table *5*); the Multilayer Perceptron (MLP) was trained without an explicit imbalance-handling mechanism, which is noted here for methodological transparency. Criterion (c) presupposes at least three years of post-batch observation and is therefore only meaningfully applicable to companies with sufficient elapsed time; for the most recent batches in the W21-S24 held-out cohort, this criterion may not yet be evaluable, and success labels for such companies rely primarily on criteria (a) and (b). This uneven observation window across cohorts is a source of right-censoring discussed further in the "Limitations" section.

Feature Engineering: The YIFE Framework

The YIFE framework introduces six novel feature categories beyond a generic Crunchbase baseline, as detailed in Table *3*.

How to cite this article:
Gupta S, Namdev P, Nagar S, et al. (September 01, 2026) Predicting Startup Outcomes Using Explainable Machine Learning and Y Combinator-Inspired Feature Engineering. Cureus J Comput Sci 3 : es44389-026-00254-0. DOI https://doi.org/10.7759/s44389-026-00254-0

Feature Category	Feature Name	Description	Type
Funding Features	total_funding_usd	Total capital raised across all rounds	Continuous
Funding Features	num_funding_rounds	Count of distinct funding events	Integer
Funding Features	seed_round_size	Amount raised in seed round (USD)	Continuous
Team Features	team_size	Number of co-founders at application	Integer
Team Features	faang_experience	Binary: any founder had prior FAANG role	Binary
Team Features	elite_edu	Binary: any founder attended top-20 university	Binary
Technical Depth	github_repo_count	Total public repos of founding team	Integer
Technical Depth	github_commit_freq	Avg. weekly commits (6 months pre-YC)	Continuous
Batch Context	batch_year_encoded	Encoded batch year (captures market cycles)	Ordinal
Batch Context	batch_size	Number of companies in same batch	Integer
Industry Features	industry_category	YC-assigned category (B2B, AI, FinTech, etc.)	Categorical
Industry Features	ai_flag	Binary: company uses AI as core product component	Binary
Geography	geo_cluster	Encoded location (San Francisco Bay, NY, International, Other)	Categorical
Network Features	tier1_vc_investor	Binary: backed by Tier-1 VC post-YC	Binary

How to cite this article:
Gupta S, Namdev P, Nagar S, et al. (September 01, 2026) Predicting Startup Outcomes Using Explainable Machine Learning and Y Combinator-Inspired Feature Engineering. Cureus J Comput Sci 3 : es44389-026-00254-0. DOI https://doi.org/10.7759/s44389-026-00254-0

TABLE 3: YIFE Feature Set—14 Features Across Six Categories

GitHub coverage: ~71% (n≈3,069); imputed with batch-cohort medians elsewhere. LinkedIn coverage: 11.6% of profiles (n=500), remaining imputed with batch-cohort medians. Continuous features standardized via z-score normalization; categorical features one-hot encoded.
YIFE, YC-Inspired Feature Engineering; YC, Y Combinator; FAANG, Facebook (now Meta), Amazon, Apple, Netflix, and Google; VC, Venture Capital

Feature Availability and Temporal Classification Framing

A primary challenge in entrepreneurial ML is the temporal availability of predictive features relative to the prediction horizon. Certain variables, particularly cumulative capital raised and total funding rounds, evolve dynamically post-incubation and share conceptual overlap with the terminal success criteria used in this study. Table *4* categorizes the YIFE feature set by typical temporal availability and associated leakage risk.

Feature Category	Representative Features	Typical Availability	Leakage Risk
Batch context	batch_year_encoded, batch_size	At cohort assignment	Low
Team structure	team_size, faang_experience, elite_edu	At application/incubation	Low
Technical depth	github_repo_count, github_commit_freq	Pre-incubation (6-month window)	Low
Industry & geography	industry_category, ai_flag, geo_cluster	At application/incubation	Low
Initial capital	seed_round_size	Concurrent with incubation	Low-Medium
Cumulative funding	num_funding_rounds, total_funding_usd	Post-incubation/evolving	High
Investor network	tier1_vc_investor	Post-incubation/evolving	High

TABLE 4: Temporal Availability and Information Leakage Assessment of YIFE Feature Categories

Leakage risk denotes the likelihood that a feature incorporates information generated after the intended prediction horizon. Features classified as Low are generally observable at or before YC admission/incubation, whereas High-risk features (e.g., cumulative funding and Tier-1 investor participation) primarily support retrospective classification and should not be interpreted as prospectively available predictors.
YIFE, YC-Inspired Feature Engineering; FAANG: Facebook (now Meta), Amazon, Apple, Netflix, and Google

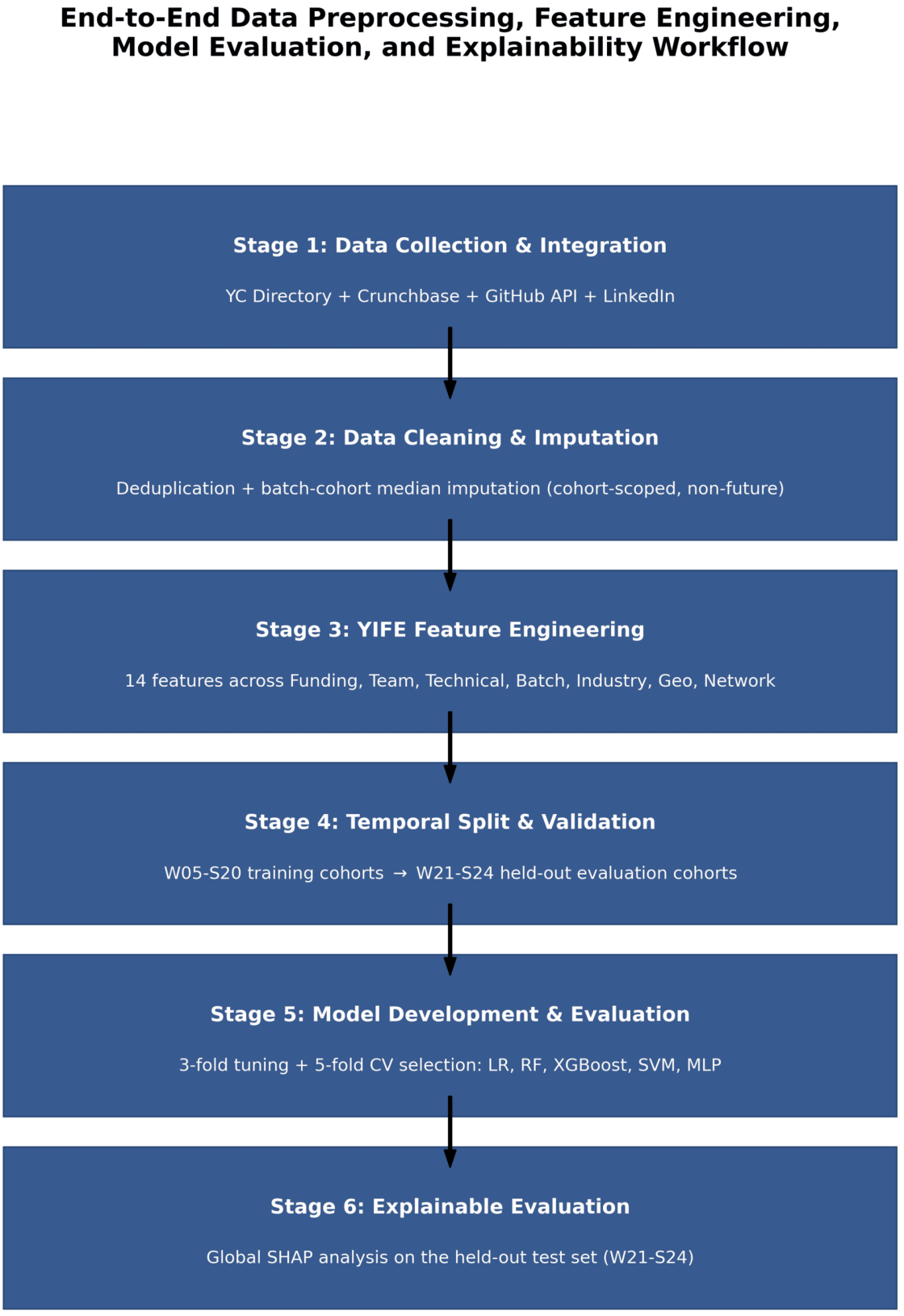
Several YIFE variables therefore require careful interpretation with respect to prediction timing. In particular, num_funding_rounds, total_funding_usd, and tier1_vc_investor may incorporate information observed after the initial YC stage, and funding-related variables are also among the highest-ranked predictors in the SHAP analysis, while Series A+ financing forms one component of the composite success label. Their predictive importance may therefore partly

reflect temporal or definitional overlap with the outcome rather than information available at a strict ex-ante prediction point. Accordingly, we present the framework not as an ex-ante screening tool deployable exclusively at application time, but as a domain-contextualized retrospective classification model designed to evaluate structural patterns of venture development within the YC portfolio. Non-funding structural features - batch timing, team composition, and technical activity - remain conceptually informative independent of this constraint; formally isolating their standalone contribution under a strict, leakage-free prediction horizon is identified below as a priority direction for future work. A stricter forecasting design would define a fixed prediction timestamp, exclude all information unavailable at that timestamp, and evaluate outcomes only after a predetermined follow-up horizon.

Experimental methodology

To ensure reproducibility, our analytical pipeline follows a standardized six-stage workflow, illustrated in Figure *1*: (1) Data Collection and Integration across the YC directory, Crunchbase, GitHub, and LinkedIn; (2) Data Cleaning and Imputation, using batch-cohort medians with cohort-scoped imputation intended to limit temporal information transfer; (3) YIFE Feature Engineering, constructing the 14-variable framework; (4) Temporal Split and Validation, separating historical training cohorts (W05-S20) from future evaluation cohorts (W21-S24); (5) Model Development and Evaluation, comprising hyperparameter tuning and model selection on the training set; and (6) Explainable Evaluation, applying SHAP analysis to the best-performing model on the held-out test set.

How to cite this article:
Gupta S, Namdev P, Nagar S, et al. (September 01, 2026) Predicting Startup Outcomes Using Explainable Machine Learning and Y Combinator-Inspired Feature Engineering. Cureus J Comput Sci 3 : es44389-026-00254-0. DOI https://doi.org/10.7759/s44389-026-00254-0



How to cite this article:
Gupta S, Namdev P, Nagar S, et al. (September 01, 2026) Predicting Startup Outcomes Using Explainable Machine Learning and Y Combinator-Inspired Feature Engineering. Cureus J Comput Sci 3 : es44389-026-00254-0. DOI https://doi.org/10.7759/s44389-026-00254-0

FIGURE 1: End-to-End Workflow of Data Collection, Preprocessing, YIFE, Model Development, Evaluation, and Explainability

End-to-end workflow of the proposed YIFE framework. The analytical pipeline begins with multi-source data collection and preprocessing, followed by the construction of YIFE features, temporal train-test splitting, machine learning model development and evaluation, and model interpretation using SHAP. The workflow illustrates the complete methodology employed for startup outcome prediction.
API, Application Programming Interface; LR, Logistic Regression; MLP, Multilayer Perceptron; RF, Random Forest; SHAP, SHapley Additive exPlanations; SVM, Support Vector Machine; YC, Y Combinator; YIFE, YC-Inspired Feature Engineering

We evaluate five ML classifiers representative of the major paradigms in supervised learning, as listed in Table *5*.

Model	Rationale	Key Hyperparameters
Logistic Regression	Linear baseline; interpretable coefficients	C=1.0, max_iter=1000, class_weight='balanced'
Random Forest	Strong ensemble baseline; robust to overfitting	n_estimators=200, max_depth=15, class_weight='balanced'
XGBoost	State-of-the-art on tabular data; handles imbalance well	n_estimators=300, learning_rate=0.05, scale_pos_weight
Support Vector Machine	Effective in high-dimensional, small-sample regimes	kernel='rbf', C=10, gamma='scale', class_weight='balanced'
MLP Neural Network	Captures non-linear interactions; deep baseline	hidden=(128,64,32), dropout=0.3, lr=0.001, epochs=100

TABLE 5: Classifiers and Hyperparameter Configuration

All models implemented in Python (scikit-learn v1.4, XGBoost v2.0).
MLP, Multilayer Perceptron

All models are implemented in Python using scikit-learn (v1.4) and XGBoost (v2.0); the MLP is implemented in PyTorch. Hyperparameter tuning was performed using 3-fold cross-validated grid search within the W05-S20 training data only. Model selection among the tuned configurations was subsequently assessed using 5-fold stratified cross-validation on the same training data; final reported generalization performance was evaluated once, after model selection, on the untouched W21-S24 held-out cohort.

Evaluation Protocol

We employ a temporal train-test split to reduce cohort-level look-ahead bias: companies from batches W05-S20 (approximately 80% of the dataset, n ≈ 3,460) form the training set, while companies from W21-S24 (the remaining 20%, n ≈ 863) constitute the held-out test set. This design evaluates generalization across later cohorts; however, as discussed in the "Feature Availability and Temporal Classification Framing" subsection, it does not by itself eliminate temporal

leakage arising from predictors that may themselves contain post-incubation information. Evaluation metrics include Accuracy, Precision, Recall, F1-score, and AUROC; given the mild class imbalance, F1-score and AUROC are treated as primary metrics, following Park et al. [8].

SHAP Interpretability

To ensure our model outputs are actionable and transparent, we apply SHAP to the best-performing model (XGBoost). SHAP values decompose each prediction into individual feature contributions, enabling global feature-importance ranking. This is particularly relevant in the venture capital context, where stakeholders require explainable decision support rather than opaque black-box predictions [11,14]. Our present analysis focuses on global mean absolute SHAP importance across the 863 held-out test instances, which establishes relative attribution magnitude and ranking only; it does not by itself establish the direction, monotonicity, or shape of each feature's relationship with the predicted outcome. Signed SHAP dependence analysis and local, instance-level force-plot visualizations were not generated for this submission and are identified as a valuable extension in the "Future Research Directions" section.

Baseline Comparison

We compare our YIFE feature set against two baselines: (B1) a generic Crunchbase feature set using only funding-round count, total funding, founding year, and industry category, analogous to what much prior work employs; and (B2) the Razaghzadeh Bidgoli et al. [3] configuration replicated on our dataset, evaluated using Logistic Regression and Random Forest, respectively. This allows us to isolate the incremental contribution of our YC-specific features for the two architectures where a matched baseline was constructed.

Results And Discussion

Model performance comparison

Table *6* reports performance for all five classifiers under the YIFE feature set, alongside the B1 and B2 baselines, evaluated once on the held-out W21-S24 test cohort after hyperparameter tuning and model selection were completed on the training set (W05-S20). These are therefore held-out test-set metrics rather than cross-validation-fold averages; the Precision, Recall, and F1-score reported for XGBoost (YIFE) are consistent with the confusion matrix in Table *7*, which is computed on the same n = 863 held-out cohort.

Logistic Regression (B1 Baseline)	**0.71**	**0.68**	**0.64**	**0.66**	**0.74**
Logistic Regression (YIFE)	0.75	0.73	0.70	0.71	0.79
Random Forest (B2 Baseline)	0.80	0.79	0.76	0.77	0.85
Random Forest (YIFE)	0.83	0.82	0.79	0.80	0.88
XGBoost (YIFE) *	0.86	0.85	0.85	0.85	0.91
Support Vector Machine (YIFE)	0.79	0.78	0.74	0.76	0.83
MLP Neural Network (YIFE)	0.81	0.80	0.78	0.79	0.86

TABLE 6: Model Performance on Held-Out Test Set (W21–S24)

*Best model. McNemar's test: XGBoost vs. MLP, $\chi^2 = 33.28$, $p < 0.0001$; vs. RF, $\chi^2 = 12.50$, $p = 0.0004$.
AUROC, Area Under Receiver Operating Characteristic Curve; MLP, Multilayer Perceptron; YIFE, YC-Inspired Feature Engineering; B1, Generic Crunchbase baseline; B2, Razaghzadeh Bidgoli et al. [3] configuration.

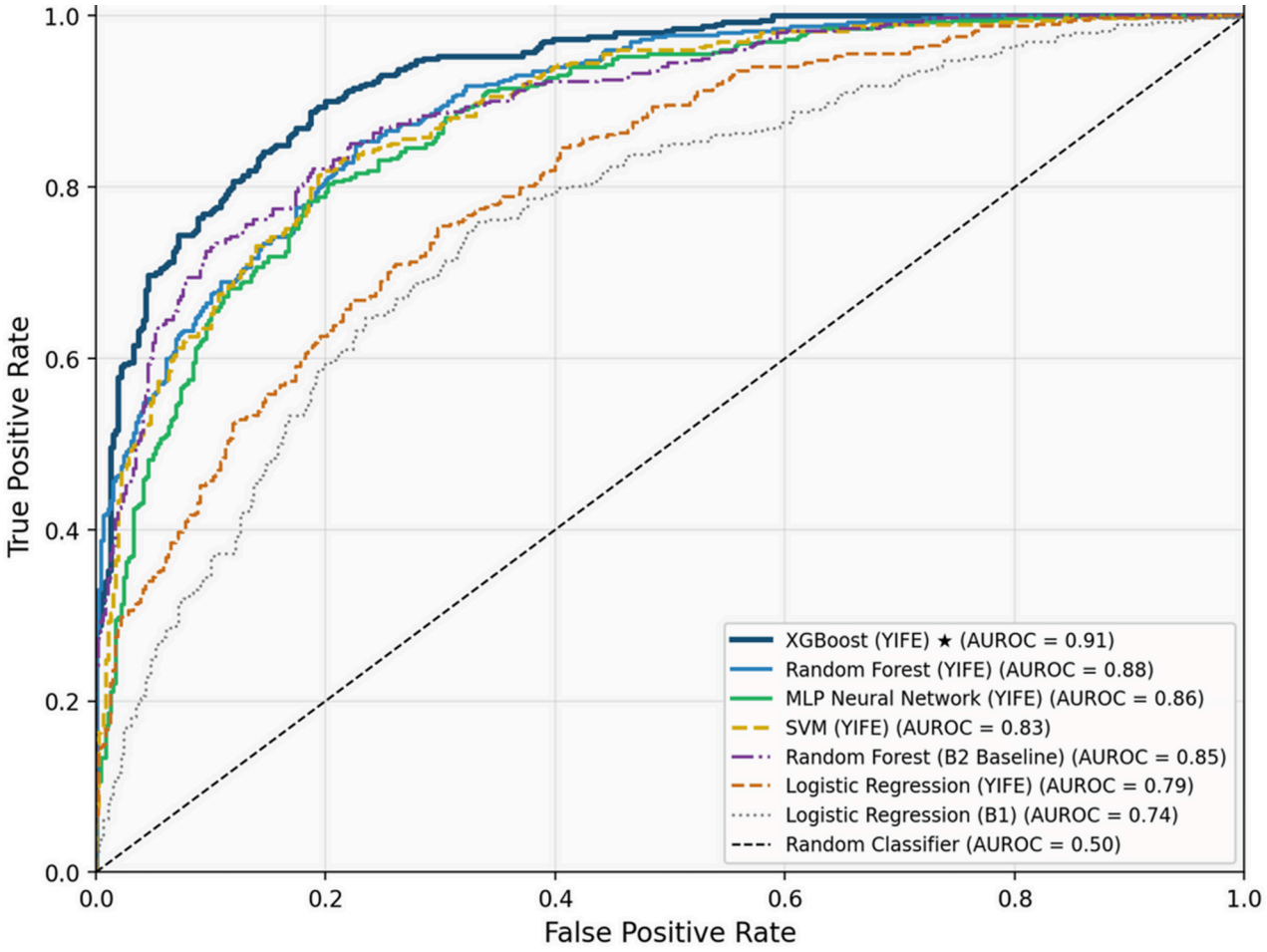


FIGURE 2: ROC Curves for Machine Learning Classifiers on the Held-Out Test Cohort (W21–S24)

ROC curves comparing the predictive performance of the evaluated machine learning models on the temporally held-out W21–S24 test cohort. XGBoost using the YIFE framework achieved the highest discriminative performance (AUROC = 0.91), followed by Random Forest (0.88), MLP (0.86), Random Forest baseline (0.85), SVM (0.83), Logistic Regression (YIFE) (0.79), and Logistic Regression baseline (0.74). The diagonal dashed line represents random classification (AUROC = 0.50).
AUROC, Area Under the Receiver Operating Characteristic Curve; MLP, Multilayer Perceptron; ROC, Receiver Operating Characteristic; SVM, Support Vector Machine; YIFE, YC-Inspired Feature Engineering

XGBoost with the YIFE feature set achieves the best overall performance, with an F1-score of 0.85 and an AUROC of 0.91 on the held-out test set - a 19-percentage-point improvement in F1-score over the generic Crunchbase baseline (0.85 vs. 0.66) and an 8-percentage-point improvement over the replicated B2 configuration (0.85 vs. 0.77). The YIFE configuration improves performance relative to the corresponding baseline comparisons available for Logistic Regression and Random Forest, while XGBoost with YIFE achieves the strongest overall performance among all evaluated configurations. Matched baseline-vs-YIFE comparisons were constructed only for the LR and RF architectures; SVM and MLP are reported under YIFE only.

Table *7* presents the confusion matrix for the best-performing XGBoost (YIFE) model on the held-out test set. The model shows a relatively balanced error distribution across the two classes, with 399 true negatives and 343 true positives; a direct unweighted comparison was not run, so this balance is descriptive rather than evidence that class-weighting specifically prevented majority-class collapse.

How to cite this article:
Gupta S, Namdev P, Nagar S, et al. (September 01, 2026) Predicting Startup Outcomes Using Explainable Machine Learning and Y Combinator-Inspired Feature Engineering. Cureus J Comput Sci 3 : es44389-026-00254-0. DOI https://doi.org/10.7759/s44389-026-00254-0

	Predicted: Fail	Predicted: Success	Row Total
Actual: Fail	TN = 399	FP = 61	460
Actual: Success	FN = 60	TP = 343	403
Col Total	459	404	n = 863

TABLE 7: Confusion Matrix—XGBoost (YIFE) on Held-Out Test Set

n = 863 (W21–S24 held-out test cohort).
Precision = 0.85, Recall = 0.85, and F1-score = 0.85.
FN, False Negatives; FP, False Positives; TN, True Negatives; TP, True Positives; YIFE, YC-Inspired Feature Engineering

To assess the incremental contribution of the LinkedIn-derived features (faang_experience and elite_edu), we conducted a targeted ablation by retraining XGBoost (YIFE) with these two features removed jointly. This reduced F1-score from 0.85 to 0.84 and AUROC from 0.91 to 0.90 on the held-out test set. This targeted ablation suggests that the two LinkedIn-derived features provide limited incremental predictive contribution in the present model, consistent with the near-zero SHAP values assigned to both features individually; however, because the features were removed together, the experiment does not isolate the individual contribution of either one, nor does it establish the standalone contribution of the funding-related features discussed above. Both remain important directions for future work under a leakage-controlled design.

SHAP feature importance analysis

Table *8* reports the global SHAP feature-importance ranking for the XGBoost model, based on mean absolute SHAP values across the 863 held-out test instances. Mean |SHAP| establishes relative attribution magnitude and ranking only; it does not by itself establish whether a feature's relationship with the predicted outcome is positive, negative, monotonic, or non-linear, so no direction-of-effect interpretation is presented here:

How to cite this article:
Gupta S, Namdev P, Nagar S, et al. (September 01, 2026) Predicting Startup Outcomes Using Explainable Machine Learning and Y Combinator-Inspired Feature Engineering. Cureus J Comput Sci 3 : es44389-026-00254-0. DOI https://doi.org/10.7759/s44389-026-00254-0

Rank	Feature	Mean |SHAP|	Direction of Effect
1	num_funding_rounds	0.187	Higher rounds → Higher success probability
2	batch_year_encoded	0.163	Recent batches (AI era, 2020+) → Higher success
3	team_size	0.141	2–4 co-founders → Optimal range
4	total_funding_usd	0.128	Higher funding → Higher success (diminishing returns)
5	ai_flag	0.112	AI-core products → Strongly positive in 2020+ batches
6	industry_category (B2B)	0.094	B2B category → Positive signal
7	github_commit_freq	0.087	Higher commit frequency → Moderate positive
8	geo_cluster (San Francisco Bay)	0.071	SF Bay Area → Mild positive (weakening post-COVID)
9	elite_edu	0.052	Elite education → Weak positive
10	faang_experience	0.031	Not robust — near-zero net effect

TABLE 8: SHAP Feature Importance Rankings for XGBoost (YIFE)

Mean absolute SHAP values computed across 863 test-set instances. Higher values indicate greater contribution to model prediction.
SHAP, SHapley Additive exPlanations; YIFE, YC-Inspired Feature Engineering

num_funding_rounds (mean |SHAP| = 0.187, higher rounds → higher success probability), batch_year_encoded (0.163, recent batches in AI era → higher success), team_size (0.141, 2-4 co-founders optimal), total_funding_usd (0.128, higher funding → higher success with diminishing returns), and ai_flag (0.112, AI-core products strongly positive in 2020+ batches).

How to cite this article:
Gupta S, Namdev P, Nagar S, et al. (September 01, 2026) Predicting Startup Outcomes Using Explainable Machine Learning and Y Combinator-Inspired Feature Engineering. Cureus J Comput Sci 3 : es44389-026-00254-0. DOI https://doi.org/10.7759/s44389-026-00254-0

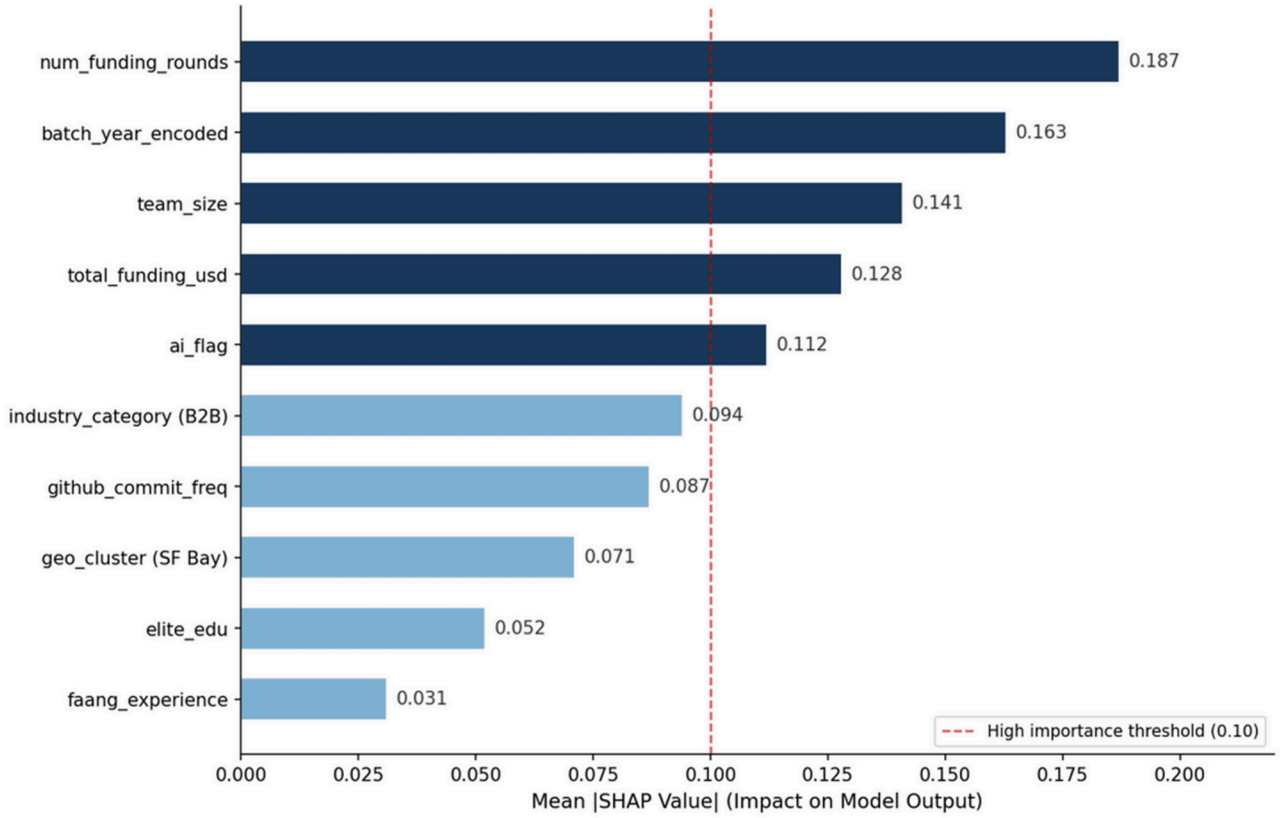


FIGURE 3: SHAP Global Feature Importance for the Best-Performing XGBoost (YIFE) Model

Mean absolute SHAP (SHapley Additive exPlanations) values for the top ten features contributing to predictions of the best-performing XGBoost model on the temporally held-out W21–S24 evaluation cohort (n = 863). Higher mean absolute SHAP values indicate greater average contribution of a feature to the model's predictions but do not indicate the direction (positive or negative) of the effect. The dashed vertical line represents a reference threshold (Mean |SHAP| = 0.10) used to highlight the most influential features.

The number of funding rounds is the strongest predictor by SHAP attribution magnitude, consistent with the general importance of funding-related signals in prior work [3,5], though as noted above, this variable's role should be read in light of its post-incubation temporal availability. The second-ranked feature, batch_year_encoded, indicates that temporal cohort context contributes substantially to model predictions; this variable captures multiple time-varying influences - including evolving accelerator selection patterns, technological cycles, and unequal outcome-maturation periods across cohorts - and should be interpreted as a proxy for cohort context rather than a direct measure of macroeconomic conditions. Team size and total funding rank third and fourth by attribution magnitude; characterizing the specific shape of their relationship with predicted outcomes (e.g., an optimal team-size range or diminishing returns to funding) would require signed SHAP dependence analysis not performed for this submission, so we report only their relative importance here. Prior FAANG experience ranked lowest by SHAP attribution magnitude within the available founder-profile sample; the joint ablation described above, which removed both LinkedIn-derived features together, resulted in only a marginal F1-score reduction (0.85 to 0.84), suggesting that these credentials carry limited combined predictive weight in this model.

Temporal analysis: Batch year effects

To investigate temporal dynamics, we stratify model performance by three batch eras, as shown in Table *9*: Early YC (2005-2012), Growth Era (2013-2019), and AI Era (2020-2024).

Batch Era	Batch Range	Companies	Success Rate	Dominant Category
Early YC	W05–S12	~420	38%	Consumer/Web
Growth Era	W13–S19	~1,640	42%	B2B SaaS/Mobile
AI Era	W20–S24	~2,263	51%	AI/B2B/FinTech

TABLE 9: Batch Era Success Rates and Dominant Categories

Success was defined as raising a Series A round or beyond, acquisition at a disclosed valuation, or sustained active operation for more than three years after the YC batch. Because companies in the AI Era (W20–S24) have had substantially shorter follow-up periods than earlier cohorts, outcome maturation is incomplete and right-censoring may be present. Accordingly, success rates should be interpreted as descriptive cohort summaries rather than directly comparable estimates of underlying success probability. Differences across eras may reflect follow-up duration, changing market conditions, cohort composition, YC selection patterns, and other time-varying factors. Survival-aware analyses with standardized follow-up windows are recommended for future work.

SaaS, Software as a Service; YC, Y Combinator

The relatively high SHAP importance of ai_flag suggests that AI orientation contributes meaningfully to model predictions within the analyzed YC dataset, particularly in recent cohorts. However, SHAP importance reflects statistical model attribution rather than causal effect; this finding should not be interpreted as evidence that AI adoption inherently causes higher startup success. Further cohort-specific analysis is required to determine whether the predictive contribution of AI orientation differs significantly across time periods.

Industry category analysis

B2B-related industry features rank among the more influential model attributions, while B2B and SaaS companies also constitute a substantial share of recent YC batches. Consumer startups - once a hallmark of earlier YC cohorts - have declined in relative share compared with enterprise, AI, and infrastructure-focused ventures in more recent cohorts.

Implications for investors and accelerators

The SHAP results have practical implications across the ecosystem. For venture investors, the analysis suggests that funding-round count and batch timing are strongly associated with predicted outcomes, though the former should be read as a retrospective signal rather than an ex-ante one given its post-incubation availability. Within the available founder-profile sample, FAANG experience and elite university attendance contribute relatively little incremental predictive information to the present model; within this dataset, team composition and market/cohort context appear more informative. This is broadly consistent with prior work on factors influencing early-stage startup success [21], the role of patents and resources in entrepreneurial-firm valuation [22], and evidence that performance persistence in entrepreneurship correlates with structural and market factors as well as individual founder credentials [23]. A related Springer-published ML benchmark for startup prediction [24] further contextualizes the broader applicability of this type of methodology. For accelerator programs, the results suggest that batch composition and cohort market timing may be associated with individual company outcomes alongside company-level selection effects.

External validity and ecosystem scoping

The empirical findings of this study are strictly scoped to the Y Combinator portfolio and should not be assumed to generalize directly to other startup accelerators or to the broader non-accelerated entrepreneurial economy. Y Combinator operates with structural characteristics of its own, including brand recognition, access to Silicon Valley

capital networks, and specific batch mentorship dynamics. Geographic concentration in the San Francisco Bay Area, and survivor bias across older cohorts (W05-S12) where companies have had more time to reach a liquidity event, must both be considered when interpreting batch-era success rates. Accelerator programs such as Techstars, 500 Global, and Seedcamp differ in selection mechanisms, regional exposure, sector composition, mentorship structure, and investor networks; relationships identified within the YC population should not be assumed to transfer directly to these ecosystems. Reliance on public data also introduces measurement error, since funding, founder activity, operational growth, and company status may be incompletely or inconsistently reported across sources. External validation across independent accelerator portfolios remains an essential next step to establish cross-ecosystem generalizability and is discussed further below.

Limitations

The major limitations of this study, along with their estimated impact and applied mitigation strategies, are summarized in Table *10*. Our dataset relies primarily on publicly available information, which may carry reporting bias, since successful companies are more likely to disclose funding details than unsuccessful ones. The GitHub activity proxy for technical depth captures only open-source work and likely underrepresents founders at companies with proprietary codebases. LinkedIn-derived features (elite_edu, faang_experience) were verified for only 500 of the 4,323 founder profiles, with remaining values imputed using batch-cohort medians, which may introduce noise into these features specifically. Our success definition, while broader than IPO-only approaches, still depends on observable funding and acquisition events, potentially missing companies that achieve sustainable profitability without formal fundraising. Finally, because of the temporal split, some companies in the 2021-2024 test cohort may not yet have reached a terminal outcome, which could conservatively deflate recall estimates; some observations may be right-censored, and current labels may change as further outcome information becomes available. Future evaluations should consider fixed observation windows or survival-aware methods to account explicitly for unequal follow-up time across cohorts.

Limitation	Impact	Mitigation Applied
Reporting bias (public data only)	Successful companies over-represented in disclosed data	Temporal split; class-weight balancing
LinkedIn coverage (11.6% of profiles)	Noise in faang_experience and elite_edu features	Joint ablation suggests limited incremental predictive contribution
GitHub coverage (~71% of profiles)	Technical-depth proxy underestimates proprietary codebases	Batch-cohort median imputation; limitation disclosed
Incomplete outcomes (2021-2024 test cohort)	May deflate recall; some companies not yet terminal	Conservative interpretation of recall; future work to revisit
Funding-variable temporal overlap	num_funding_rounds / total_funding_usd may reflect post-incubation information	Reframed as retrospective classification; leakage table added

TABLE 10: Study Limitations, Potential Impact, and Mitigation Strategies

Summary of the principal methodological limitations of the present study, their potential impact on predictive performance and interpretation, and the measures implemented to reduce their influence. Mitigation strategies include temporal validation, feature ablation, batch-cohort median imputation, conservative interpretation of retrospective outcomes, and explicit acknowledgment of temporal information overlap where applicable.

How to cite this article:
Gupta S, Namdev P, Nagar S, et al. (September 01, 2026) Predicting Startup Outcomes Using Explainable Machine Learning and Y Combinator-Inspired Feature Engineering. Cureus J Comput Sci 3 : es44389-026-00254-0. DOI https://doi.org/10.7759/s44389-026-00254-0

Future research directions

Future research should extend YIFE along several complementary directions. First, external validation should evaluate whether accelerator-specific features transfer to ecosystems such as Techstars, 500 Global, and Seedcamp through direct replication or transfer-learning approaches. Second, stricter ex-ante prediction protocols should define a fixed prediction timestamp and exclude funding or investor information observed after that point, thereby separating early-stage signals from retrospective outcome indicators and formally isolating the standalone contribution of non-funding structural features. Third, multimodal models could integrate structured company characteristics with startup descriptions, founder biographies, product text, and other contemporaneously available signals. Fourth, graph-based approaches could model relationships among founders, investors, accelerators, and companies, while temporal deep-learning and survival-analysis methods could capture evolving startup trajectories and unequal outcome horizons across cohorts. Additional paired statistical comparisons (e.g., repeated cross-validation with paired tests) and instance-level SHAP dependence and force-plot visualizations, beyond the global importance rankings reported here, represent further methodological extensions. Finally, future work should evaluate model calibration and stability under changing macroeconomic and technology cycles. Looking further ahead, emerging paradigms such as Quantum AI may enhance future predictive analytics by enabling more efficient optimization of complex, high-dimensional entrepreneurial datasets, though such applications remain exploratory for startup prediction at this stage [25].

Comparison to prior work

Our XGBoost YIFE model (F1-score = 0.85, AUROC = 0.91) compares favorably with recent literature. Razaghzadeh Bidgoli et al. [3] reported 82% accuracy for Random Forest and 80% for Gradient Boosting on a Crunchbase dataset. Park et al. [8] achieved improved performance after GAN-based oversampling. Our AUROC of 0.91 is broadly consistent with the range reported by Mashhadi et al.'s [11] interpretable ML pipeline on a firm-quarter panel, though direct comparison is constrained by differences in dataset scope, success definitions, and evaluation protocols.

Conclusions

This study demonstrates that accelerator-specific contextual feature engineering can improve machine-learning classification of startup outcomes within the constructed Y Combinator dataset. Among the evaluated configurations, XGBoost with YIFE achieved the strongest performance (F1-score = 0.85; AUROC = 0.91), while SHAP analysis identified funding activity, batch context, and team composition as influential contributors to model behavior. These findings support the value of incorporating ecosystem-specific information alongside conventional startup characteristics. However, the results should be interpreted within the YC context and as primarily retrospective, since several funding-related variables may not be available at a strict early-stage prediction point and recent cohorts have shorter outcome histories. Accordingly, the present work provides a domain-contextualized benchmark for startup outcome classification within the YC ecosystem rather than evidence of universally generalizable ex-ante startup forecasting. Future research should evaluate leakage-controlled prediction horizons, independent accelerator datasets, multimodal signals, and temporally adaptive modeling approaches.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Siddharth Gupta, Anjali Deshwal

Acquisition, analysis, or interpretation of data: Siddharth Gupta, Pratham Namdev, Shubham Nagar, Sunny Singh

Drafting of the manuscript: Siddharth Gupta, Sunny Singh

How to cite this article:
Gupta S, Namdev P, Nagar S, et al. (September 01, 2026) Predicting Startup Outcomes Using Explainable Machine Learning and Y Combinator-Inspired Feature Engineering. Cureus J Comput Sci 3 : es44389-026-00254-0. DOI https://doi.org/10.7759/s44389-026-00254-0

Critical review of the manuscript for important intellectual content: Siddharth Gupta, Pratham Namdev, Shubham Nagar, Anjali Deshwal

Supervision: Anjali Deshwal

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The code and analytical pipeline are publicly available on GitHub at https://github.com/guptasiddharth2409/yife-startup-prediction. The original dataset used for the reported results is not publicly redistributed because it was compiled from multiple third-party sources under their respective access terms. The repository contains a synthetic demonstration dataset for examining the analytical workflow. The synthetic data were not used to generate the results reported in this article.

Acknowledgements

The authors thank the Department of Computer Science and Engineering, Greater Noida Institute of Technology (GNIT), for institutional support and research facilities. They also acknowledge the open-source software community whose tools supported the computational analysis, including scikit-learn, XGBoost, SHAP, and PyTorch. This article was previously presented at the International Conference on Progressive Trends in Engineering Science and Management (PTESM 2026), organized by Greater Noida Institute of Technology and held on April 10–11, 2026; the work was presented on April 11, 2026.

References

1. Y Combinator company statistics. Y Combinator. Accessed: July 1, 2026: https://www.ycombinator.com/.
2. Arroyo J, Corea F, Jimenez-Diaz G, Recio-Garcia JA: Assessment of machine learning performance for decision support in venture capital investments. IEEE Access. 2019, 7:124233-124243. 10.1109/access.2019.2938659
3. Razaghzadeh Bidgoli M, Raeesi Vanani I, Goodarzi M: Predicting the success of startups using a machine learning approach. Journal of Innovation and Entrepreneurship. 2024, 13:80. 10.1186/s13731-024-00436-x
4. Antretter T, Blohm I, Grichnik D: Predicting startup survival from digital traces: towards a procedure for early stage investors. ICIS 2018 Proceedings. 2018, 1-9.
5. Kim J, Kim H, Geum Y: How to succeed in the market? Predicting startup success using a machine learning approach. Technological Forecasting and Social Change. 2023, 193:122614. 10.1016/j.techfore.2023.122614
6. Blohm I, Antretter T, Siren C, Grichnik D, Wincent J: It's a people's game, isn't it?! A comparison between the investment returns of business angels and machine learning algorithms. Entrepreneurship Theory and Practice. 2020, 46:1054-1091. 10.1177/1042258720945206
7. Mampilly AJ, Manda VK: Driving change: securing success through digital transformation. Modern Management Challenges: A Case Compendium. Excel India Publishers, New Delhi; 2024. 196-205. 10.5281/zenodo.11910143
8. Park J, Choi S, Feng Y: Predicting startup success using two bias-free machine learning: resolving data imbalance using generative adversarial networks. Journal of Big Data. 2024, 11:122. 10.1186/s40537-024-00993-8

How to cite this article:
Gupta S, Namdev P, Nagar S, et al. (September 01, 2026) Predicting Startup Outcomes Using Explainable Machine Learning and Y Combinator-Inspired Feature Engineering. Cureus J Comput Sci 3 : es44389-026-00254-0. DOI https://doi.org/10.7759/s44389-026-00254-0

9. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002, 16:321-357. 10.1613/jair.953

10. Maarouf A, Feuerriegel S, Prollochs N: A fused large language model for predicting startup success. European Journal of Operational Research. 2025, 322:198-214. 10.1016/j.ejor.2024.09.011

11. Mashhadi S, Saghezchi A, Ghassemzadeh Kashani V: Interpretable machine learning for predicting startup funding, patenting, and exits [PREPRINT]. arXiv. 2026, 10.48550/arXiv.2510.09465

12. Adl R: Founder backgrounds and startup funding: evidence from Y Combinator [PREPRINT]. arXiv. 2025, 10.48550/arXiv.2512.13755

13. Corea F, Bertinetti G, Cervellati EM: Hacking the venture industry: An Early-stage Startups Investment framework for data-driven investors. Machine Learning with Applications. 2021, 5:100062. 10.1016/j.mlwa.2021.100062

14. Lundberg SM, Lee SI: A unified approach to interpreting model predictions. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017, 4768-4777.

15. Breiman L: Random forests. Machine Learning. 2001, 45:5-32. 10.1023/a:1010933404324

16. Ribeiro MT, Singh S, Guestrin C: "Why should I trust you?": explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016, 1135-1144. 10.1145/2939672.2939778

17. Grinsztajn L, Oyallon E, Varoquaux G: Why do tree-based models still outperform deep learning on typical tabular data?. NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems. 2022, 507-520.

18. Dalle JM, den Besten M, Menon C: Using Crunchbase for Economic and Managerial Research. OECD Science, Technology and Industry Working Papers, No. 2017/08. OECD Publishing, Paris; 2017. 10.1787/6c418d60-en

19. Font-Cot F, Lara-Navarra P, Sánchez-Arnau C, Sánchez-Pérez EA: Startup survival forecasting: A multivariate AI approach based on empirical knowledge. Information. 2025, 16:61. 10.3390/info16010061

20. Pan C, Pan X, Sun L: A multi-layer AI decision support system for startup success prediction and risk assessment using knowledge graphs and federated learning. Scientific Reports. 2026, 16:20259. 10.1038/s41598-026-44162-8

21. Krishna A, Agrawal A, Choudhary A: Predicting the outcome of startups: less failure, more success. 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain. 2016, 798-805. 10.1109/ICDMW.2016.0118

22. Hsu DH, Ziedonis RH: Resources as dual sources of advantage: Implications for valuing entrepreneurial-firm patents. Strategic Management Journal. 2013, 34:761-781. 10.1002/smj.2037

23. Gompers P, Kovner A, Lerner J, Scharfstein D: Performance persistence in entrepreneurship. Journal of Financial Economics. 2010, 96:18-32. 10.1016/j.jfineco.2009.11.001

24. Kalbande S, Karmore R: Startup success prediction using machine learning. Proceedings of Fifth Doctoral Symposium on Computational Intelligence (DoSCI 2024). Swaroop A, Kansal V, Fortino G, Hassanien AE (ed): Springer, Singapore; 2024. 1086:309-319. 10.1007/978-981-97-6036-7_26

25. Manda VK, Bhukya M, Mirtskhulava L: Quantum AI: merging quantum computing and artificial intelligence. Quantum Artificial Intelligence Applications in Industry 5.0. Darwish D (ed): Deep Science Publishing, London, UK; 2026. 25-47. 10.70593/978-93-7185-732-1

How to cite this article:
Gupta S, Namdev P, Nagar S, et al. (September 01, 2026) Predicting Startup Outcomes Using Explainable Machine Learning and Y Combinator-Inspired Feature Engineering. Cureus J Comput Sci 3 : es44389-026-00254-0. DOI https://doi.org/10.7759/s44389-026-00254-0