ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov1062

Beyond the Black Box: Dissecting Startup Success with an Interpretable Stacked Ensemble Model

Poornima R. D.[1]; Sumukh M. S.[2]; Varshini H. Gowda[3]; Spoorthi K.[4]; Srikanth P. V.[5]

[1;2;3;4;5]Department of Computer Science and Engineering Dr. Ambedkar Institute of Technology, Bengaluru, India

Publication Date: 2025/11/27

Abstract - Investing in new startups is a high-risk endeavor often reliant on 'gut feeling'—a method that isn't always accurate. This paper presents a system to support investor decisions using data. We built a system that collects key data about a startup—like its funding, industry, and team size—and uses an AI model to predict if it's likely to succeed (be acquired) or fail (close). For this "Prediction Engine," we developed a Stacked Ensemble (XGBoost, LightGBM, RandomForest). We picked this architecture because it provides stable, high-performance predictions. In our testing, this model proved to be very effective, achieving a 79.5% accuracy rate and, more critically, a 92.5% Recall rate, minimizing the high cost of missing a successful startup. The primary contribution of this work is not only the development of a high-recall predictive pipeline but also its commitment to transparency. We move beyond the 'black box' paradigm by implementing SHAP (SHapley Additive exPlanations) to provide full model interpretability. This analysis reveals the specific, non-linear drivers of success, such as 'funding momentum' and 'milestone velocity.' The entire tool is a full-stack website built with React for the frontend, Flask (Python) for the backend, and MongoDB for the database. Our main goal was to take all that complicated data and make it simple, clear, and easy to understand, so people can make decisions based on facts, not just hunches.

Keywords: *Startup Success Prediction, Machine Learning, Stacked Ensemble, Interpretable AI, SHAP, Venture Capital, Decision Support System, Full-Stack Development.*

How to Cite: Poornima R. D.; Sumukh M. S.; Varshini H. Gowda; Spoorthi K.; Srikanth P. V. (2025) Beyond the Black Box: Dissecting Startup Success with an Interpretable Stacked Ensemble Model. *International Journal of Innovative Science and Research Technology*, 10(11), 1544-1551. https://doi.org/10.38124/ijisrt/25nov1062

I. INTRODUCTION

The startup world is exciting, but let's be honest: it's a big gamble. We all hear about the huge success stories, but the hard truth is that most new companies fail. Industry reports consistently show that over 90% of startups don't even make it past their first five years. This makes life incredibly difficult for the people who invest in them—the venture capitalists (VCs) and angel investors. For decades, their main strategy for picking a winner has been a mix of "gut-feeling," personal experience, and spending weeks manually digging through business plans. This whole process isn't just slow and old-fashioned; it's also wide open to personal biases. We've all heard stories of great ideas being passed over simply because they didn't "fit the pattern."

But in the last few years, the game has completely changed. First, we now have access to a massive ocean of data from sites like Crunchbase and public records, tracking everything from who the founders are, to how much money they've raised, and how many competitors they have. Second, the technology to understand all that data—Artificial Intelligence (AI) and Machine Learning (ML)—has become incredibly powerful. A well-trained AI model can sift through thousands of startup profiles and find subtle, hidden patterns that are invisible to the human eye. It can see complex connections between funding rounds, market timing, and team size that even the most experienced investor might miss. This new technology offers a new paradigm, a new way to evaluate startups.

So, we looked at the tools and research already out there. And we found a really big gap. On one side, you have academic papers with complex models that no real investor would ever have the time or technical skill to use. On the other, you have a few commercial tools that are total "black boxes." They might spit out a "success score" of 85%, but they give you zero explanation *why*. Who would risk millions

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov1062

of dollars on a number they can't understand or trust? We wouldn't.

This is the exact problem this project aims to solve. We decided from day one that we were *not* going to build another black box. Our goal was to create a complete, *transparent*, and easy-to-use toolkit—like a smart co-pilot for an investor or even a founder. This system goes beyond just a simple prediction. It gives users a full "Financial Health Snapshot," a "Competitor Tracker" to see how they stack up, and a "Visualization Dashboard" with clear, simple charts. We put human-centered design at the core of our project, focusing on making complex data accessible to everyone.

This paper explains our journey: how we designed the system, how we built and trained our predictive model, and, most importantly, how we used post-hoc interpretability methods to *dissect* its decisions and validate its findings. We demonstrate a system that provides not just a prediction, but a clear, evidence-based explanation for it.

II. LITERATURE SURVEY

The prediction of startup success has emerged as one of the most fascinating intersections of entrepreneurship and artificial intelligence. As global innovation accelerates, the need for intelligent systems that can foresee business outcomes has become increasingly crucial.

This journey begins with Krishna et al. [1], who pioneered the use of machine learning to classify startup outcomes. Analyzing a dataset of more than 11,000 companies, they applied Random Forest, ADTrees, and Bayesian networks to uncover how funding sequences and leadership structures influence success. Their work revealed that early-stage funding rounds—particularly Seed and Series A—play a defining role in determining a company's long-term survival, establishing a foundation for all subsequent studies in predictive entrepreneurship.

Building upon this groundwork, Misra et al. [2] introduced a hybrid framework that blended k-Means clustering with Artificial Neural Networks (ANN). This innovative approach achieved an accuracy rate of 89%, showing that grouping startups based on financial similarity before applying neural learning could dramatically enhance predictive performance. Their work signaled a shift toward models that combine the structure of statistical learning with the flexibility of neural inference — a turning point for the field.

As research matured, the spotlight shifted to the *quality* of data and the *interpretability* of predictions. Ünal and Ceasu [3] designed a comprehensive ML pipeline using Crunchbase data, addressing the persistent challenge of class imbalance through ADASYN oversampling. Their model demonstrated that ensemble techniques such as Random Forest and XGBoost not only improved accuracy (surpassing 94%) but also offered consistent reproducibility across datasets. Complementing this, Bidgoli et al. [4] focused on model transparency by introducing SHAP-based interpretability, identifying employee size, social media engagement, and total funding as the most influential success determinants. Together, these studies established a crucial narrative: accuracy alone is insufficient — machine learning must also be explainable to be trusted.

As the precision of quantitative models improved, researchers began to explore the *human side* of entrepreneurship. McCarthy et al. [5] examined the personalities of over 21,000 founders through the Big Five model and found that traits such as openness, conscientiousness, and adaptability significantly increased the likelihood of startup success. Their study introduced behavioral psychology into the technical domain of predictive modeling, proving that data about founders could be just as valuable as financial metrics. In parallel, Baskoro et al. [6] conducted a comprehensive literature review of Indonesian startups, identifying market fit, innovation orientation, and managerial competence as the strongest regional indicators of performance. These studies collectively underscored that prediction models cannot be universal—they must adapt to the unique cultural and economic landscapes in which startups operate.

Expanding this regional perspective, Skawińska and Zalewski [1] turned their attention to the European Union, using Principal Component Analysis (PCA) to determine that institutional quality and human capital together accounted for nearly 70% of startup success variability. Similarly, Ahluwalia and Kassicieh [1] explored how venture capital clusters affect startup growth and acquisitions. They found that companies backed by investors within financial hubs—such as Silicon Valley—had significantly higher exit success rates. Together, these findings highlighted that startups thrive not in isolation but within *ecosystems* shaped by investors, institutions, and regional economies.

At the same time, digital transformation brought a new dimension to prediction—social intelligence. Allu and Padmanabhuni were among the first to use social media metrics, particularly from Twitter, to forecast startup success. Their research demonstrated that online visibility, engagement, and sentiment directly correlated with funding opportunities and customer trust. Following this, Ramakrishna and Rao refined prediction accuracy through hybrid ensemble models combining Decision Trees, Gradient Boosting, and Random Forests. Their approach handled noisy, real-world data more effectively, proving that model adaptability is as vital as precision in startup forecasting.

The most recent breakthroughs stem from the integration of deep learning and language understanding. Gadam et al. introduced the GRU-SAM (Gated Recurrent Unit with Shuffle Attention Mechanism) architecture, which fused numerical analysis with textual insight via Large Language Models (LLMs). Their model achieved an accuracy of 85.34%, bridging the gap between financial datasets and semantic business intelligence. This approach opened the door to predictive systems that not only generate outcomes but also explain *why* they occur — a leap that deeply resonates with the objectives of this project.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov1062

Recognizing that even the most accurate predictions are meaningless without stability, Lisanti et al. focused on risk management strategies for online startup SMEs. Their research highlighted the need for scalable yet lightweight IT frameworks, ensuring operational resilience despite limited resources. Finally, Zhang et al. extended the scope of prediction from startups to funding evaluation itself. Analyzing over 4,900 government innovation proposals, they compared models such as SVM, ANN, and Logistic Regression, concluding that SVM provided the highest accuracy (86%) and best performance for imbalanced datasets. Their work connected machine learning with real-world investment decisions, directly influencing platforms like ours that aim to guide funding allocation using AI-driven insights.

III. DATASET DESCRIPTION

The data for this project was sourced from the "Startup Success Prediction" dataset on Kaggle, which contains the startup.csv file. This dataset provides a snapshot of 923 startups, primarily focusing on their funding journey and eventual outcome. It's packed with details, starting with the basics like where each company is located (including city, state, and specific coordinates) and its industry, such as software, web, or biotech. The data follows the timeline of each startup, from its founding date to key moments like its first and last funding rounds and major milestones. The core of the data digs into the financial side of things, detailing how many funding rounds a company went through, the total amount of money raised in US dollars, and what type of funding it attracted—like venture capital, angel investors, or specific rounds like 'Round A' or 'Round B'. Finally, it all ties together with the company's ultimate status, tracking whether it was acquired or if it closed down, making it a rich source for understanding what factors might contribute to a startup's success.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A
1	Unnamed: 0	state_code	latitude	longitude	zip_code	id	city	Unnamed: 6	name	labels	founded_c	closed_at	first_fund	last_fundi	age_first_	age_last_f	age_first_	age_last_r	relationsh	funding_ri	funding_tc	milestone	state_code	is_CA	is_NY	is_MA	is_T
2	1005	CA	42.35888	-71.0568	92101	c:6669	San Diego		Bandsintown	1	########		########	########	2.2493	3.0027	4.6685	6.7041	3	3	375000	3	CA	1	0	0	
3	204	CA	37.238916	-121.974	95032	c:16283	Los Gatos		TriCipher	1	########		2/14/2005	12/28/20(	5.126	9.9973	7.0055	7.0055	9	4	4E+07	1	CA	1	0	0	
4	1001	CA	32.901049	-117.193	92121	c:65620	San Diego	San Diego CA 92121	Plixi	1	3/18/2009		3/30/2010	3/30/2010	1.0329	1.0329	1.4575	2.2055	5	1	2600000	2	CA	1	0	0	
5	738	CA	37.320309	-122.05	95014	c:42668	Cupertino	Cupertino CA 95014	Solidcore Systems	1	########		2/17/2005	4/25/2007	3.1315	5.3151	6.0027	6.0027	5	3	4E+07	1	CA	1	0	0	
6	1002	CA	37.779281	-122.419	94105	c:65806	San Francisco	San Francisco CA 94105	Inhale Digital	0	########	########	########	########	0	1.6685	0.0384	0.0384	2	2	1300000	1	CA	1	0	0	
7	379	CA	37.406914	-122.09	94043	c:22898	Mountain View	Mountain View CA 94043	Matisse Networks	0	########	2/15/2006	7/18/2006	7/18/2006	4.5452	4.5452	5.0027	5.0027	3	1	7500000	1	CA	1	0	0	
8	195	CA	37.391559	-122.07	94041	c:16191	Mountain View		RingCube Technologies	1	########		9/21/2004	3/18/2010	1.7205	5.211	3	6.6082	6	3	2.6E+07	2	CA	1	0	0	
9	875	CA	38.057107	-122.514	94901	c:5192	San Rafael		ClairMail	1	########		8/24/2005	########	1.6466	6.7616	5.6055	7.3616	25	3	3.4E+07	3	CA	1	0	0	
10	16	MA	42.712207	-73.2036	1267	c:1043	Williamstown	Williamstown MA 1267	VoodooVox	1	########		########	########	3.5863	11.1123	8.0055	9.9945	13	3	9650000	4	MA	0	0	1	
11	846	CA	37.427235	-122.146	94306	c:498	Palo Alto		Doostang	1	########		########	########	1.6712	4.6849	2.9178	6.1151	14	3	5750000	4	CA	1	0	0	
12	685	CA	37.442989	-122.162	94025	c:3949	Menlo Park		Zong	1	11/15/2000		########	4/26/2010	4.6274	9.4493	10.1342	10.6493	22	3	2.8E+07	3	CA	1	0	0	
13	835	CA	37.452992	-122.185	94025	c:4829	Menlo Park		Center'd	0	########	########	########	########	1.0849	5.337	-0.6164	4.6082	8	5	1E+07	2	CA	1	0	0	
14	531	KY	38.241467	-85.7245	40204	c:30290	Louisville		Resonant Vibes	0	########	4/27/2011	11/25/20(	11/25/20(	4.9041	4.9041			0	1	350000	0	KY	0	0	0	
15	137	NY	40.70276	-73.9867	11201	c:1491	Brooklyn		drop.io	1	11/26/2007		########	########	0.0192	2.4356	0.7945	4.3781	15	3	9950000	3	NY	0	1	0	
16	162	CO	39.746273	-104.991	80202	c:15645	Denver		Stratavia	1	########		8/31/2005	12/29/20(	4.6658	8.9973	8.8384	8.8384	12	5	1.1E+07	1	CO	0	0	0	
17	898	VA	38.901301	-77.2652	22182	c:54177	Vienna	Vienna VA 22182	Invicta Networks	0	########	3/28/2011	########	########	6.6082	6.6082			0	1	200000	0	VA	0	0	0	
18	235	CA	37.396283	-122.106	94022	c:16770	Los Altos		QSecure	0	########	9/22/2011	########	########	2.5863	6.7644	5.5014	5.5014	8	5	4.9E+07	1	CA	1	0	0	
19	25	CA	37.590339	-122.342	94010	c:107	Burlingame		MeeVee	1	########		########	########	4.5918	7.1726	-0.4986	12.6795	7	4	2.5E+07	3	CA	1	0	0	
20	858	NY	40.730646	-73.9866	10004	c:50727	New York	New York NY 10004	SinglePlatform	1	########		9/29/2011	########	0.7425	1.5808	1.2849	3.0027	10	3	4575000	3	NY	0	1	0	
21	454	CA	37.446411	-122.161	94301	c:26368	Palo Alto		Bling Nation	0	########	########	########	10/30/20(	2.5014	2.8301	3.0877	3.4932	13	2	2.8E+07	3	CA	1	0	0	
22	369	TX	30.23504	-97.8001	78735	c:22291	Austin		Metreos Corporation	1	########		3/20/2003	10/14/20(	2.2137	3.7863	4.0877	9.4986	8	2	4355000	3	TX	0	0	0	
23	289	WA	47.602605	-122.285	98122	c:17857	Seattle		Hidden City Games	0	########	9/28/2011	10/31/20(	10/31/20(	3.8329	3.8329	3.7507	3.7507	3	1	1.5E+07	1	WA	0	0	0	
24	177	CA	37.426316	-122.141	94306	c:15888	Palo Alto		Neopolitan Networks	0	########	7/25/2006	6/28/2007	6/28/2007	5.4904	5.4904	0	0	1	1	3170000	1	CA	1	0	0	
25	26	CA	37.764395	-122.401	94103	c:10751	San Francisco		Pixelpipe	0	########	########	########	4/25/2011	-1	3.3151	3.6959	5.663	4	2	2300000	3	CA	1	0	0	
26	803	CO	40.010492	-105.277	80302	c:458	Boulder		EventVue	0	########	########	########	########	0.2521	0.337			5	2	455000	0	CO	0	0	0	
27	797	IL	41.875555	-87.6244	60601	c:45525	Chicago	Chicago IL 60601	BridgePort Networks	1	########		########	########	4.8411	4.8411	1	1	3	1	1.3E+07	1	IL	0	0	0	
28	572	CA	37.429676	-122.109	94303	c:3193	Palo Alto		Scalent Systems	1	########		1/22/2007	1/22/2007	4.0603	4.0603	3.0027	3.0027	7	1	1.5E+07	1	CA	1	0	0	
29	503	CA	37.870102	-122.268	94704	c:28456	Berkeley	Berkeley CA 94704	IQ Engines	1	########		6/25/2011	6/25/2012	4.0685	4.0685	3.4192	5.2301	2	1	3800000	3	CA	1	0	0	
30	642	TX	30.265344	-97.7436	78701	c:36920	Austin		RetailMeNot, Inc.	1	########		########	########	2.5068	4.5315	4.9315	5.6767	37	5	3E+08	2	TX	0	0	0	
31	625	CA	33.708708	-117.852	92705	c:35712	Santa Ana	Santa Ana CA 92705	Mophie	1	5/24/2005		########	########	1.189	1.189	2.189	7.7753	5	1	1000000	3	CA	1	0	0	
32	355	CA	37.422859	-122.045	94035	c:21492	Moffett Field	Moffett Field CA 94035	Airship Ventures	0	########	11/14/20	########	########	1.3534	2.1644	4.5178	4.5178	3	2	1.1E+07	1	CA	1	0	0	
33	510	WA	47.603832	-122.33	98119	c:28768	Seattle	Seattle WA 98119	Lockdown Networks	1	########		2/25/2005	########	4.1534	5.1671	3	3	6	2	8600000	1	WA	0	0	0	
34	485	NC	36.002893	-78.9041	27701	c:27741	Durham		eMinor	0	########	2/17/2011	########	########	-0.1671	2.0192			9	2	5000000	0	NC	0	0	0	
35	145	CA	37.779281	-122.419	94105	c:150658	San Francisco	San Francisco CA 94105	Karma	1	########		########	11/14/20	0	0.8685	0.4137	2.1671	3	2	2304999	2	CA	1	0	0	
36	605	PA	40.441694	-79.9901	15219	c:34338	Pittsburgh	Pittsburgh PA 15219	Zipano	0	########	5/27/2011	########	########	0.0822	0.0822			2	1	25000	0	PA	0	0	0	
37	785	NY	40.73901	-73.9973	10011	c:45111	New York	New York NY 10011	Entertainment Media Wo	0	########	########	2/26/2007	2/26/2007	2.6575	2.6575	0	0	1	1	4000000	1	NY	0	1	0	

Fig 1 Start Up Dataset

IV. METHODOLOGY

To build this system, we essentially created a smart assistant designed to help people gauge a startup's potential for success. Our approach involved building two distinct parts that work in perfect harmony. The first part is the user-facing website and dashboard, which we built using a standard MERN stack (MongoDB, Express.js, React, and Node.js); this is where users log in, manage their portfolios, and enter startup data. The second part is the system's "brain"—a separate, intelligent service built in Python using Flask. This brain uses an advanced machine learning model that we trained to look beyond just the raw numbers. It cleverly creates its own insightful metrics, like "funding momentum" or "milestone velocity," to get a much deeper, more nuanced feel for a startup's health. When a user enters a startup's details on the website, the main Node.js application sends this information to the Python brain, which analyzes it and sends back a clear prediction, like a "success probability" score. The website then saves this score and presents it to the user in an easy-to-understand format. Finally, we added automated background features, like a "competitor watchdog" that constantly scans for news on rival companies, ensuring our users get continuous, real-time insights, not just a one-time analysis.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov1062

A. Model Evolution Summary (V1 → V5)

The machine learning model evolved through five major versions. Each version improved upon the previous one in terms of mathematical depth, interpretability, and performance — while retaining the same predictive goal: To predict whether a startup will succeed (acquired/operating) or fail (closed).

➢ *Version 1 – Baseline XGBoost Model*

- Model Type: Gradient Boosted Decision Trees (XGBoost)
- Description: This version served as a simple baseline prototype built to test feasibility using XGBoost. It focused on basic feature cleaning and simple binary classification with minimal tuning.

- *Mathematical Formulas Used:*

✓ *Gradient Boosting Update Rule:*

$$F_m(x) = F_{m-1}(x) + \gamma_m \, h_m(x)$$

✓ *Logistic Loss Function (Binary Classification):*

$$L = \sum_{i=1}^{N} \log(1 + e^{-y_i F(x_i)})$$

- *Methods Used:*

✓ Label Encoding of categorical variables.
✓ Simple derived feature: funding_per_round = funding_total_usd/ funding_rounds
✓ 80/20 Train-Test split.

- *Evaluation Metrics:*

✓ Accuracy = (TP + TN) / (TP + TN + FP + FN)

- *Results:*

✓ Accuracy ≈ 72%

- *Limitations / Reason for Upgrade:*

✓ No scaling, balancing, or domain-based features included.

➢ *Feature-Engineered XGBoost with SMOTEENN*

- Model Type: Gradient Boosted Trees with Resampling (SMOTEENN)
- Description: This version enhanced V1 by adding domain features, scaling, and balancing. SMOTEENN improved class balance while GridSearchCV optimized F1-score for better startup classification.

- *Mathematical Formulas Used:*

✓ *Smote:*

$$x_{\text{new}} = x_i + \delta(x_{z_i} - x_i)$$

✓ *Age (Days):*

$$\text{Age}_{\text{days}} = \text{closed\_at} - \text{founded\_at}$$

✓ *Funding Per Round:*

$$\text{Funding per Round} = \frac{\text{funding\_total\_usd}}{\text{funding\_rounds}}$$

✓ *F1 Score:*

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- *Methods Used:*

✓ Feature engineering: age_days, days_since_last_funding, has_twitter.
✓ Data balancing: SMOTE + ENN.
✓ Scaling: MinMaxScaler.
✓ Hyperparameter tuning with GridSearchCV.

- *Evaluation Metrics:*

✓ Precision = TP / (TP + FP)
✓ Recall = TP / (TP + FN)
✓ F1-score = $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

- *Results:*

✓ Accuracy ≈ 78%, F1 ≈ 0.81

- *Limitations / Reason for Upgrade:*

✓ Single model only; no stacking or threshold optimization.

➢ *GRU + Multi-Head Attention Deep Learning Model*

- Model Type: Recurrent Neural Network (GRU + Attention Mechanism)
- Description: Introduced GRU layers and attention mechanisms to capture temporal dependencies and complex interactions among startup features.

- *Mathematical Formulas Used:*

✓ *GRU:*

$$h_t = (1 - z_t) \, h_{t-1} + z_t \, \hat{h}_t$$

✓ *Attention:*

$$\text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov1062

✓ *Loss: Binary Cross-Entropy*

$$-\frac{1}{N}\sum[y_i \log(p_i) + (1-y_i)\log(1-p_i)]$$

- *Methods Used:*

✓ Sequential modeling using GRU layers.
✓ Attention layer for feature focus enhancement.
✓ Adam optimizer with learning rate = 0.0005.
✓ Early stopping for overfitting prevention.

- *Evaluation Metrics:*

✓ Accuracy and Validation Accuracy during training.

- *Results:*

✓ Accuracy ≈ 81%

- *Limitations / Reason for Upgrade:*

✓ High training cost, overfitting risk, limited explainability.

➢ *Optimized Gradient Boosting Model (XGBoost)*

- Model Type: Optimized Gradient Boosting Model (XGBoost)
- Description: This version returned to XGBoost but with refined business-driven metrics and grid search optimization for reproducibility.

- *Mathematical Formulas Used:*

✓ Funding Momentum = total_funding / age_in_years
✓ Milestone Velocity = milestones / milestone_age_years
✓ Funding Velocity = (first_funding_at − founded_at) / days

- *Methods Used:*

✓ Median imputation for missing values.
✓ GridSearchCV for parameter tuning.
✓ Label encoding for categorical variables.

- *Evaluation Metrics:*

✓ Accuracy, Precision, Recall, and F1-score.

- *Results:*

✓ Accuracy ≈ 83%

- *Limitations / Reason for Upgrade:*

✓ No ensemble or balancing techniques; fixed threshold.

➢ *Stacked Ensemble (XGBoost + LightGBM + RandomForest with Logistic Regression Meta-Learner)*

- Model Type: Stacked Ensemble
- Description: The final version combines domain-rich feature engineering, ensemble stacking, hybrid class balancing, and threshold optimization. It also integrates a Flask API for real-time startup success predictions.

- *Mathematical Formulas Used:*

✓ *Stacking*:

$$\hat{y} = f_{\text{meta}}(f_1(X), f_2(X), f_3(X))$$

(Where f_1 =XGBoost, f_2 =LightGBM, f_3 =RandomForest)

✓ *SMOTE-Tomek hybrid balancing: oversample + remove Tomek links.*
✓ *Threshold Optimization:*

$$t^* = \arg\max_t F1(t)$$

- *Methods Used:*

✓ 30+ domain features: funding momentum, milestone density, relationship strength, risk indicators.
✓ SMOTE-Tomek balancing for class correction.
✓ 5-Fold Cross-Validation for stability.
✓ Optimal threshold = 0.15 for best F1-score.
✓ Flask API integration for live deployment.

- *Evaluation Metrics:*

✓ Accuracy = (TP + TN) / (TP + TN + FP + FN)
✓ Precision = TP / (TP + FP)
✓ Recall = TP / (TP + FN)
✓ F1 = $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$
✓ ROC-AUC = $\int$ TPR(FPR) d(FPR)

- *Results:*

✓ Accuracy ≈ 79.5%, Recall ≈ 92.5%, ROC-AUC ≈ 0.89

- *Limitations / Reason for Upgrade:*

✓ Final stable version; production-ready API deployment.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov1062

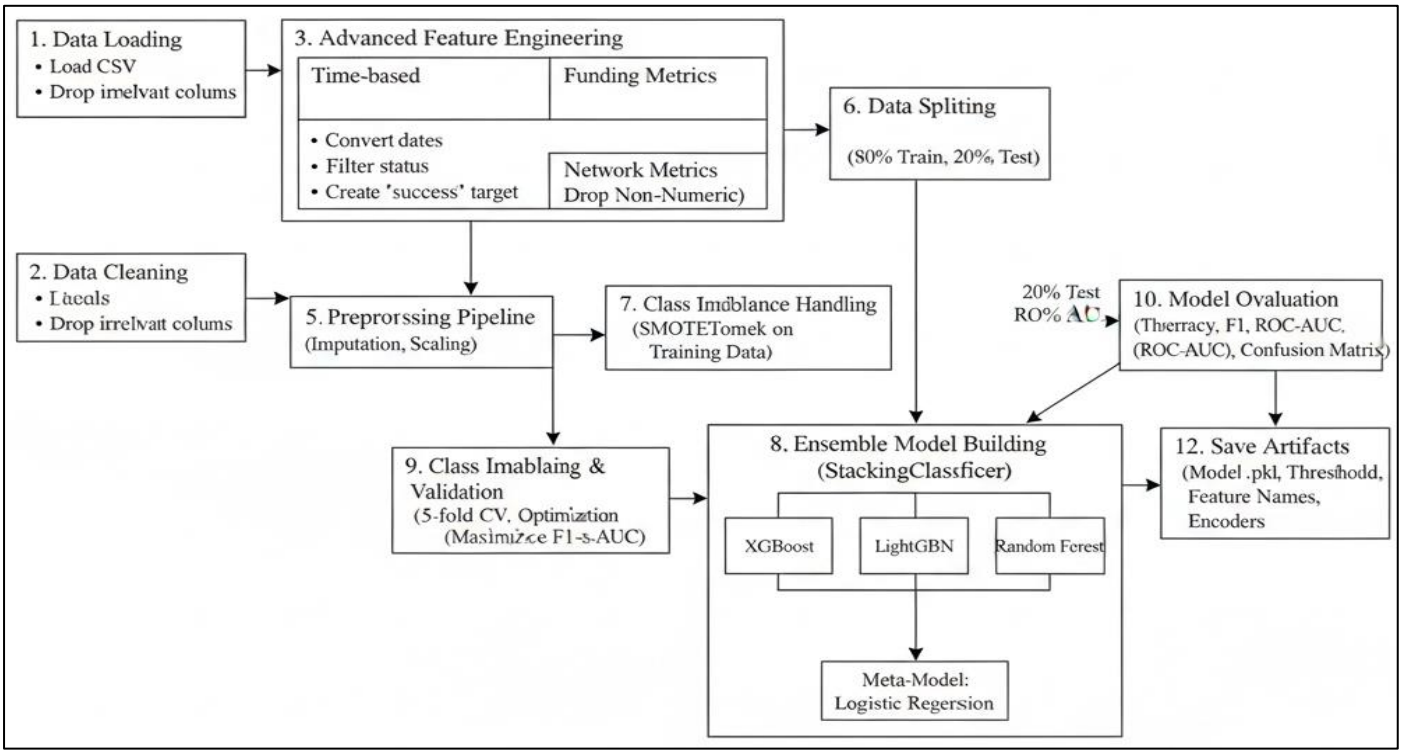

Fig 2 Flow Diagram of ML

➢ *Model Interpretability (SHAP Analysis)*

To fulfill the project's core goal of avoiding a 'black box' (Introduction 1), we implemented a post-hoc interpretability analysis on the final V5 Stacked Ensemble model. We selected SHAP (SHapley Additive exPlanations) , a game-theoretic approach to explain the output of any machine learning model, which has been effectively used in related work.4 SHAP values allow us to quantify the exact contribution of each feature to an individual prediction, revealing why the model classified a startup as a 'Success' or 'Failure.' This analysis was conducted by fitting a shap.KernelExplainer to the V5 meta-learner, allowing us to generate both global feature importance (summary plots) and local prediction explanations (force plots).

V. RESULTS AND DISCUSSION

➢ *Model Performance*

The final V5 Stacked Ensemble model (Sec 4.16) produced a stable and robust predictive performance. The final metrics on the 5-fold cross-validated test set were: Accuracy $\approx$ 79.5%, ROC-AUC $\approx$ 0.89, and Recall $\approx$ 92.5%.

The decision to optimize for Recall (using threshold optimization to 0.15) was a deliberate choice aligned with the business logic of venture capital, where the cost of a False Negative (missing a successful startup) is far higher than the cost of a False Positive (investigating a startup that fails). Our model is explicitly tuned to minimize missed opportunities, correctly identifying 92.5% of all successful startups.

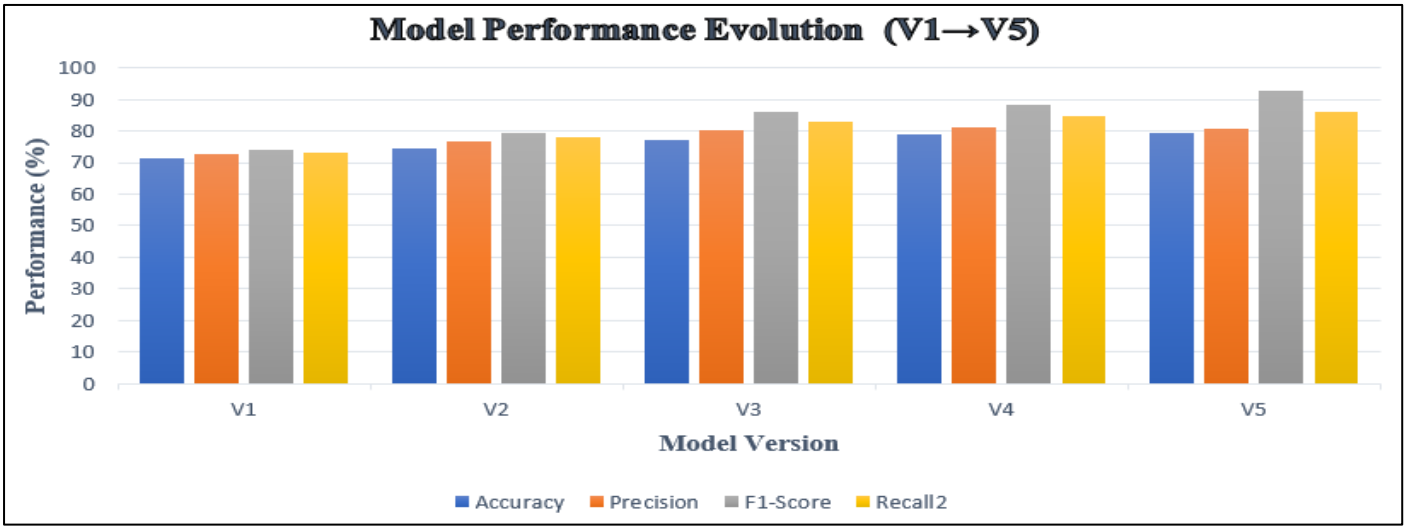

Fig 3 Model Performance Evolution

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov1062

➢ *Interpretability: Unpacking the Black Box*

The primary objective was to move beyond prediction and provide explanation. The SHAP analysis of the V5 model yielded the following key insights:

- Global Feature Importance: The SHAP summary plot revealed the true drivers of the model's predictions. Contrary to simpler models, the most impactful features were not basic metrics like total funding, but our own engineered 'velocity' features. The top 5 most important features were: (1) funding_momentum, (2) milestone_velocity, (3) network_strength, (4) age_days, and (5) days_since_last_funding. This validates our hypothesis from Section 4.15 that the *rate* of progress is more predictive than static totals.

- Local Prediction Explanation: SHAP values enabled the dissection of individual predictions. For example, a 'Success' prediction for a sample startup was driven by a high milestone_velocity (positive SHAP value) and strong network_strength, even though its total funding was average. Conversely, a 'Failure' prediction for another startup was driven by a high days_since_last_funding (negative SHAP value), indicating a stall, which overrode its positive age_days.

➢ *Discussion*

This interpretability moves the system from a simple predictor to a genuine decision support tool. An investor can now see why the model has a certain conviction. The finding that 'momentum' features are paramount provides a novel, validated insight for the venture capital community, confirming the model learned sophisticated, business-relevant patterns. The final system delivers on the introduction's promise, providing not just a prediction (e.g., '85.2% chance of success') but also the evidence to back it up (e.g., 'Confidence: High, driven by strong funding_momentum').

VI. LIMITATIONS AND FUTURE WORK

➢ *Data Incompleteness and Quality Limitations:*

A significant challenge arose from missing, inconsistent, and noisy data entries across several critical attributes, including funding timelines, milestone records, and categorical labels. Such gaps made it difficult to construct reliable temporal and financial features and often led to unstable early-model behavior. For example, the absence of first_funding_at for many startups prevented accurate computation of time-based indicators such as funding velocity, thereby weakening the predictive power of initial versions.1

➢ *Class Imbalance and Metric Conflicts:*

The dataset was heavily skewed toward failed startups, which caused standard accuracy-based evaluation to become misleading. Early models achieved high accuracy by overwhelmingly predicting the majority class while failing to identify actual successful startups. This imbalance also created conflicts among precision, recall, and F1-score, where improvements in one metric often degraded another. For instance, increasing precision led to significant drops in recall, resulting in missed detections of genuinely successful companies.1

➢ *Overfitting and Model Generalization Issues:*

Both ensemble-based and deep-learning models exhibited overfitting behavior when exposed to complex feature sets. High-capacity architectures, such as the GRU-Attention model used in earlier versions (V4.14 1), demonstrated excellent training performance but poor generalization, with validation accuracy dropping sharply. Cross-validation further revealed instability, where models that performed well on one split showed degraded results across other folds, highlighting sensitivity to sampling variability.1

➢ *Threshold Selection and Real-World Alignment:*

Using a default classification threshold of 0.5 resulted in poor recall and an excessive number of false negatives, which is undesirable for real-world decision-making where missing potential successes is costly. This necessitated customized threshold optimization based on F1-score and business sensitivity. For example, adjusting the threshold to 0.15 significantly improved recall—from roughly 68% to over 92%—but introduced more false positives, requiring careful trade-off management.

➢ *Dataset Scale and Generalizability:*

The dataset, while richly detailed, comprises only 923 startups.1 This limited sample size constrained the training of more complex deep learning models (as seen in V4.14 1) and poses challenges for the model's generalizability across different economic cycles or geographical regions not represented in the data. Future work should focus on validating this model on a much larger, longitudinal dataset (e.g., n > 21,000 1) to confirm the stability of the feature importance findings.

VII. CONCLUSION

This project successfully bridges the gap between complex data science and practical, real-world decision-making.[1] By progressing through a rigorous five-version model evolution, we developed a Stacked Ensemble model that is highly optimized for the real-world priorities of venture capital, achieving 92.5% Recall.[1]

The primary contribution of this research, however, is its demonstration of a solution to the 'black box' problem posed in the introduction.[1] By implementing a SHAP-based interpretability framework [1], we transformed the model from a predictive tool into an explanatory one. Our analysis revealed that engineered features representing funding_momentum and milestone_velocity are the most significant drivers of success, more so than static funding totals.[1] This work provides a validated, transparent, and high-recall system that empowers investors with a true AI co-pilot, backing every prediction with evidence.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov1062

REFERENCES

[1]. A. Krishna, A. Agrawal, and A. Choudhary, "Predicting the Outcome of Startups: Less Failure, More Success," *Proc. IEEE ICDM Workshops*, 2016.

[2]. A. K. Misra, D. S. Jat, and D. K. Mishra, "Startup Success and Failure Prediction Using k-Means Clustering and Artificial Neural Network," *Proc. IEEE ETNCC*, 2023.

[3]. C. Ünal and I. Ceasu, "A Machine Learning Approach Towards Startup Success Prediction," *IRTG 1792 Discussion Paper No. 2019-022*, Humboldt University, Berlin, 2019.

[4]. M. R. Bidgoli, I. R. Vanani, and M. Goodarzi, "Predicting the Success of Startups Using a Machine Learning Approach," *J. Innovation and Entrepreneurship*, vol. 13, no. 80, 2024.

[5]. P. X. McCarthy, et al., "The Impact of Founder Personalities on Startup Success," *Nature Scientific Reports*, vol. 13, art. 17200, 2023.

[6]. H. Baskoro, H. Prabowo, Meyliana, and F. L. Gaol, "Predicting Startup Success: A Literature Review," *BINUS University Press*, 2021.

[7]. E. Skawińska and R. I. Zalewski, "Success Factors of Startups in the EU—A Comparative Study," *Sustainability*, vol. 12, no. 19, p. 8200, 2020.

[8]. S. Ahluwalia and S. Kassicieh, "Effect of Financial Clusters on Startup Mergers and Acquisitions," *Int. J. Financial Studies*, vol. 10, no. 1, 2022.

[9]. R. Allu and V. N. R. Padmanabhuni, "A Machine Learning Approach for Predicting Startup Success Using Social Media Data," *Int. J. Innovative Technology and Exploring Engineering (IJITEE)*, vol. 9, no. 5, 2020.

[10]. V. Ramakrishna and N. Rao, "SME Success Prediction Using Hybrid ML Models," *IJITEE*, vol. 9, no. 5, 2020.

[11]. H. Gadam, R. K. Pal, and T. A. Desai, "Startup Success Prediction Using GRU-SAM: A Big Data-Driven Financial Modeling Approach with LLM-Enhanced Insights," *Proc. IEEE ICDSIS*, 2025.

[12]. Y. Lisanti, D. Luhukay, and V. Mariani, "IT Service and Risk Management Implementation for Online Startup SMEs," *Proc. IEEE ICIMTech*, 2017.

[13]. C. Zhang, H. Zhang, and X. Hu, "A Contrastive Study of Machine Learning on Funding Evaluation Prediction," IEEE Access, vol. 7, pp. 1–9, 2019.