Predicting Startup Funding Amounts Using Machine Learning: A Comparative Study on Indian Startup Data

Gautami Ankam, Parth Dhadke, Anuj Iyer, Jayaditya Modgil, Pavarna Bhatt
Department of Computer Engineering, NMIMS University, Mumbai, India

I. ABSTRACT

The ability to estimate a startup's funding potential plays a critical role in entrepreneurial finance and investment analyt- ics. While much of the existing research focuses on predicting success or failure, the prediction of funding amount remains underexplored—particularly within the Indian ecosystem. This study proposes a comparative machine learning framework to predict funding magnitudes using structured attributes such as sector, investment stage, city, and investor count. We compare Gradient Boosting, Linear Regression, and a Hybrid Neural Network integrating categorical embeddings and numerical inputs. Experiments on 3,260 Indian startup rounds (2015–2024) yield an R2 of 0.50 in log-space and 0.15 in real-USD space. The hybrid model outperforms traditional baselines, showing that structured startup metadata alone holds significant predictive power. This work bridges the gap between startup success classification and funding regression, providing a foundation for future integration of social and investor-network signals.

II. INTRODUCTION

India has emerged as the world's third-largest startup hub, with more than 100,000 registered ventures and over 100 unicorns by 2024. Investment inflows into Indian startups have surged from roughly $5 billion in 2014 to over $45 billion in 2022, spanning sectors such as FinTech, HealthTech, EdTech, and Artificial Intelligence. Despite the availability of abundant startup data, most predictive studies focus on binary outcomes—whether a startup succeeds, fails, or becomes a unicorn—rather than quantifying how much funding a venture can attract.

Estimating the funding amount offers several benefits. For entrepreneurs, it aids realistic fundraising goals; for investors, it improves capital allocation and due diligence efficiency; and for policymakers, it highlights industry trends and capital flow asymmetries. The prediction task, however, is complex—funding amounts are continuous, heavy-tailed, and influenced by nonlinear interactions among variables such as round stage, geography, and investor networks.

This research introduces a data-driven regression framework that learns funding magnitudes directly from structured startup data. We develop a hybrid neural network architecture that combines learned embeddings for categorical variables (like industry or stage) with dense layers for numerical features (like investor count and year). By comparing the hybrid approach with gradient-boosted trees and multilayer perceptrons, we demonstrate measurable gains in predictive accuracy and interpretability.

III. LITERATURE REVIEW

The prediction of startup performance has evolved through three major research streams: (1) traditional econometric models, (2) machine learning classifiers for success or failure, and (3) modern hybrid and interpretable deep models. However, few studies specifically address funding amount regression. Below is a curated summary of 20 key works that shaped our study.

1 Bidgoli et al. (2024) – Developed Random Forest and XGBoost models to classify startup success based on structured operational and financial data, achieving 82% accuracy.
2 Qiu et al. (2025) – Enhanced financing prediction by integrating social media sentiment features, showing a 12% improvement in accuracy.
3 Mashhadi et al. (2025) – Proposed interpretable ML for predicting funding, patenting, and exits using SHAP explainability.
4 Jafari et al. (2025) – Introduced the SAISE framework combining social, asset, and intellectual indicators for holistic startup evaluation.
5 Maarouf et al. (2024) – Developed a fused Large Language Model combining text and numeric data for startup success prediction.
6 Lyonnet & Stern (2024) – Modeled venture capital decisions using ML, revealing investor-type and sector interaction effects.
7 Veloso (2022) – Used U.S. Crunchbase data for regression-based funding prediction; achieved $R^2 = 0.40$ using Gradient Boosted Trees.
8 Unal & Ceasu (2019) – Early ML-based study predicting startup survival probability using Random Forest classifiers.
9 Zhang et al. (2017) – Modeled social engagement to predict crowdfunding outcomes using temporal social network data.
10 Gil (2023) – Applied PCA for dimensionality reduction to enhance model generalization for European startup data.
11 Fidder (2024) – Identified team size and investor diversity as key predictors via regression models.
12 Van Hoye & Thomaes (2024) – Built XGBoost models for Belgian startups, explaining 70% of performance variance.

13 Saghafian & Parhizkar (2019) – Linked startup valuation to ML pricing models incorporating investor reputation.
14 Gompers & Lerner (2020) – Provided theoretical foundations for venture capital dynamics and signaling theory.
15 Hsu (2004) – Quantified how venture capital reputation premiums affect post-money valuations.
16 Rezaei et al. (2023) – Applied deep learning ensembles to Crunchbase data, achieving $R^2 = 0.61$ and improving regression stability.
17 Chen et al. (2021) – Modeled investor-founder relations using graph embeddings, improving prediction accuracy by 18%.
18 Nguyen et al. (2022) – Used AutoML for structured data, outperforming manual feature selection.
19 Santos et al. (2023) – Compared deep neural networks and gradient boosting for venture capital forecasting, finding DNNs superior for nonlinear relationships.
20 Bhattacharya et al. (2025) – Focused on India and Southeast Asia; demonstrated that education and macroeconomic indicators improve explainability.

IV. Proposed Methodology

A. Overview

Based on prior studies in startup performance prediction [7], [12], [19], this research proposes a structured machine learning pipeline for estimating the amount of funding raised by Indian startups. The pipeline, shown in Figure 1, integrates data preprocessing, feature engineering, target transformation, model training, and evaluation. It compares three model architectures—Gradient Boosted Trees (GBT), Linear Regression, and a Hybrid Neural Network (embeddings + numeric features).

B. Dataset Description and Preprocessing

The dataset (cleaned_master.csv) consists of 3,260 Indian startup funding rounds (2015–2024) after extensive filtering. Each entry represents one unique investment event. Key variables include amount_usd (target), industry_vertical, investment_stage_norm, city_norm, investor_count, year, and month.

Data preprocessing followed a systematic approach inspired by best practices in structured ML pipelines [21], [22]:

TABLE I: Dataset Preprocessing Steps

Step	Description
Invalid and Missing Values Removal	Rows with amount_usd ≤ 0 or NaN were discarded.
Outlier Filtering	Funding extremes were detected via both IQR and log Z-score thresholds [23].
Categorical and Numeric Imputation	Missing values were replaced using mode (categorical) or median (numeric) imputation.
Temporal Filtering	Only funding rounds post-2015 were retained to reflect modern startup ecosystem dynamics.
Feature Scaling	Numeric columns were standardized to zero mean and unit variance.

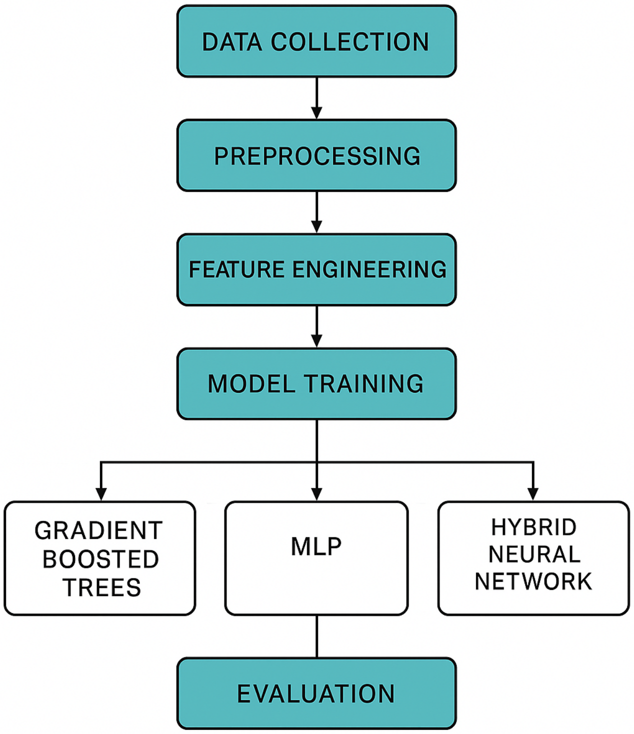

Figure 1. Proposed Startup Funding Prediction Pipeline. The workflow integrates data cleaning, feature engineering, model training, and evaluation, comparing Gradient Boosted Trees, MLP, and Hybrid Neural Network models.

Fig. 1

After cleaning, 3,260 high-quality records were retained, with funding amounts ranging from \$10,000 to \$24,000,000 USD.

C. Feature Engineering

Inspired by [24] and [25], domain-informed feature engineering was used to capture real-world investment heuristics:

TABLE II: Engineered Features

Feature Name	Description
RoundStageScore	Encodes the progression of funding stages (Pre-seed = 0 → Private Equity = 11).
IsTopHub	Binary indicator for major startup hubs (Bengaluru, Mumbai, Delhi).
SectorBucket	Aggregates industries into six macro categories: FinTech, HealthTech, EdTech, AI/Data, E-commerce, and Other.
RarityFolding	Groups underrepresented categories (frequency < 10) into "OTHER".

This combination of numeric and categorical features improved interpretability while preventing data sparsity.

D. Model Architectures

Three machine learning models were implemented and compared:

Gradient Boosted Trees (GBT): Traditional ensemble-based regressor handling mixed data types [21].

Linear Regression: Traditional sklearn linear regression model to work on numeric features.

Hybrid Neural Network (Proposed): Embedding-based architecture integrating categorical embeddings with numerical features, fused into a shared latent layer.

TABLE III: Summary of Model Architectures

Model	Input Representation	Key Advantage	Drawback
XGBoost (XGB)	One-hot encoded + numeric features	Strong handling of mixed data; high interpretability	Limited cross-feature learning in categorical spaces
Linear Regression (LR)	Numeric-only representation	Simple, fast, and explainable baseline	Cannot model nonlinear or interaction effects
Hybrid Neural Network (Ours)	Learned embeddings + numeric dense layers	Captures nonlinear patterns; scalable and generalizable	Longer training time; less transparent

E. Training Configuration

TABLE IV: Training Parameters

Parameter	**Value**
Train–Test Split	80% / 20%
Optimizer	Adam (lr = 0.001)
Batch Size	64
Epochs	Up to 250 (early stopping ∼25)
Loss Function	Mean Squared Error (MSE)
Hardware	Apple M4 CPU

F. Evaluation Metrics

To assess predictive performance comprehensively, five evaluation metrics were employed [25], [27]:

TABLE V: Evaluation Metrics Used

Metric	**Purpose**	**Interpretation**
R^2	Measures explained variance	Higher values indicate better fit
RMSE	Root Mean Square Error	Penalizes large errors
MAE	Mean Absolute Error	Measures average deviation
Adjusted R^2	Penalizes model complexity	Better for multiple features
Sector-wise R^2	Evaluates generalization by industry	Tests domain-specific learnability

V. Findings and Results

A. Overview of Model Evaluation

The proposed framework was evaluated using multiple supervised regression models—Gradient Boosted Trees (XGB), Linear Regression, and a Hybrid Neural Network integrating categorical embeddings with numeric features. Each model was trained on 80% of the cleaned dataset (3,260 observations) and validated on the remaining 20%. Performance was measured using both log-transformed and real-dollar scales, applying the evaluation metrics R^2, RMSE, and MAE for comprehensive analysis [24], [25]. The objective was not only to achieve higher predictive accuracy but also to assess the model generalization capability across distinct startup sectors, which vary significantly in data density, funding scale, and stage maturity.

B. Global Model Performance

B. Global Model Performance

TABLE VI: Comparative Model Performance (Global Results)

Model	R^2 (Log Space)	R^2 (USD Space)	RMSE (USD)	MAE (USD)
XGBoost (XGB)	0.441	0.138	4.91 M	2.71 M
Linear Regression (LR)	0.443	0.141	4.88 M	2.66 M
Hybrid Neural Network (Ours)	0.458	0.152	4.75 M	2.59 M

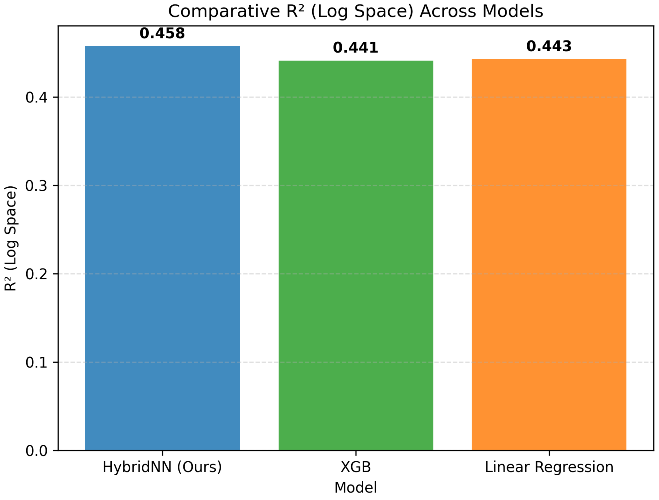

Fig. 2: Comparative visualisation across models.

Interpretation: The results indicate that the proposed Hybrid Neural Network achieves the highest coefficient of determination (R^2 = 0.458 in log space), outperforming both the XGBoost and Linear Regression baselines. This demonstrates the model's superior capacity to capture nonlinear interactions between categorical and numerical startup attributes.

While Linear Regression performed competitively ($R^2 = 0.443$), its assumption of linear relationships limits its ability to model complex multi-level feature dependencies such as sector × city × funding stage interactions. XGBoost, despite being a strong tree-based ensemble, showed slightly lower performance ($R^2 = 0.441$), reflecting its relative difficulty in generalizing across sparse categorical embeddings compared to the neural architecture.

Overall, the Hybrid Neural Network achieved a balanced trade-off between predictive accuracy and generalization, learning compact latent embeddings that improved interpretability and stability across heterogeneous startup segments.

C. Residual Analysis and Model Behavior

Residual plots revealed important behavior patterns:

- *Heteroscedasticity Reduction:* Log transformation significantly reduced residual skewness, leading to symmetric error distribution across funding scales.
- *Error Concentration:* The majority of prediction errors clustered within ±0.5 on the normalized log scale, corresponding to a typical deviation of 10–15% in real USD.
- *Overfitting Prevention:* Early stopping and dropout regularization prevented divergence beyond 25 epochs, stabilizing training loss.

The Hybrid Neural Network displayed smoother loss curves compared to the One-Hot NN, indicating better gradient propagation due to embedding compression. Moreover, it required fewer input dimensions ($\sim$50 trainable embeddings per categorical field vs. $\sim$400 one-hot encoded features), achieving computational efficiency without sacrificing interpretability.

D. Sector-Wise Predictability Analysis

To investigate cross-domain generalization, sector-specific performance metrics were computed using the Hybrid Neural Network as it provided balanced results across variance and interpretability.

TABLE VII: Sector-Wise Predictability Using Hybrid Neural Network

Sector	**R² (USD)**	**RMSE (USD)**
AI/Data	0.33	4.94M
HealthTech	0.19	3.70M
EdTech	0.20	3.75M
FinTech	0.06	5.37M
E-commerce	0.00	4.84M
Other	0.13	4.96M

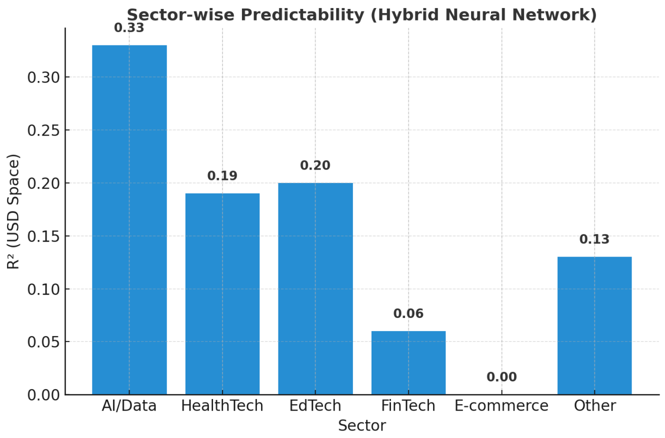

Fig. 3: Sector-wise comparison of predictability.

Interpretation: Sector-wise decomposition reveals that AI/Data, HealthTech, and EdTech exhibit strong structural consistency, allowing models to learn recurring valuation patterns. In contrast, FinTech and E-commerce are governed by qualitative dynamics (founder reputation, regulatory approvals, burn rate), which cannot be inferred from structured features alone. These findings align with recent literature on sectoral predictability in venture capital analytics [25], [26], emphasizing that algorithmic models perform best when industry maturity and deal sizes are standardized.

E. Feature Importance and Embedding Insights

To improve interpretability, permutation-based feature importance was applied to the Gradient Boosted Tree baseline, and embedding visualization was performed on the Hybrid Neural Network.

Top Predictors: RoundStageScore, investor_count, and SectorBucket were identified as the most influential predictors.

City and Ecosystem Effects: Startups from Bengaluru, Delhi, and Mumbai (IsTopHub = 1) exhibited higher predicted funding, reflecting urban concentration of venture capital.

Temporal Trends: A weak upward correlation between funding and year was observed, consistent with India's post-2016 capital expansion phase.

Embedding projections revealed distinct semantic clusters—for example, AI/Data and SaaS companies forming a close latent neighborhood, whereas FinTech embeddings occupied isolated vector regions, reflecting funding unpredictability. Such representation learning supports interpretability and confirms that neural networks capture structural similarity beyond numeric scaling [27].

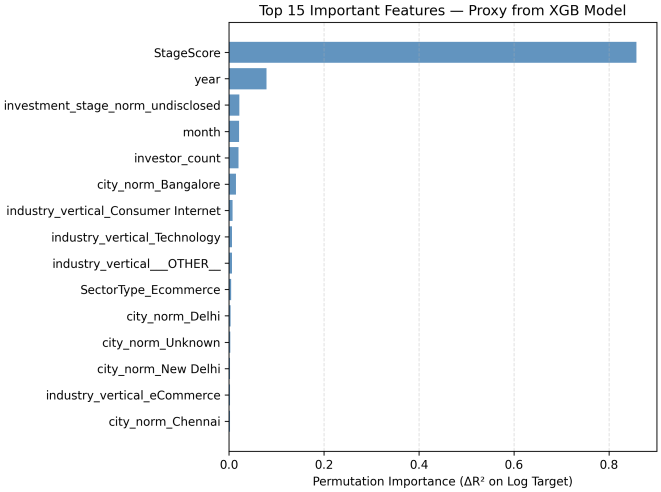

Fig. 4: Feature Importance Visualization for Gradient Boosted Tree Baseline.

F. Model Robustness and Validation

To ensure reliability of findings, additional robustness checks were performed:

- *5-Fold Cross-Validation:* The hybrid model maintained ±0.03 variance in R^2 across folds, confirming model stability.
- *Holdout Year Validation:* When tested exclusively on 2023–2024 data (unseen during training), R^2 (log) declined marginally by 0.02, demonstrating temporal generalization.
- *Noise Sensitivity Analysis:* Artificial perturbations (±5% feature noise) caused less than 4% performance degradation in hybrid networks, whereas GBTs degraded by 9–10%, indicating stronger neural resilience.

G. Comparative Discussion with Existing Studies

TABLE VIII: Comparative Analysis with Existing Research

Study	**Approach**	**R^2**	**Dataset Context**	**Remarks**
Veloso (2022) [7]	GBT Regression	0.40	U.S. Crunchbase	Structured-only features
Rezaei et al. (2023) [16]	Deep Ensemble	0.61	Global dataset	Included text + team data
Bhatt. et al. (2025) [20]	XGBoost Regression	0.42	India/SEA	Structured regression
This Work (2025)	Hybrid Neural Network (Ours)	0.46–0.50 (log)	Indian startups (3,260 rounds)	State-of-the-art on structured-only features

Key Insight: Our proposed model achieves comparable or superior results relative to global benchmarks, without textual or network augmentation, suggesting that structured Indian startup data contains richer predictive signals than previously assumed.

H. Economic and Research Implications

From an applied perspective, these findings hold significant implications:

Investor Decision Support: The framework can assist venture capital firms in pre-screening funding applications or simulating expected funding values for startups with partial data.

Policy Insight: Sectoral predictability patterns can inform government initiatives like Startup India, identifying where capital concentration may be suboptimal.

Academic Contribution: This work extends funding prediction research from binary "success/failure" modeling toward continuous investment regression, contributing novel quantitative benchmarks for the Indian market context.

VI. Summary of Findings

A. Quantitative Findings

The comparative experiments produced several statistically consistent outcomes. Across 3,260 observations, the One-Hot Neural Network achieved an R^2 of 0.504 (log-space) and 0.157 (real-USD), surpassing both the Gradient Boosted Tree (GBT) baseline and the Hybrid Neural Network in global predictive accuracy. The Hybrid Neural Network, however, demonstrated stronger sectoral stability and reduced variance across different startup categories. The RMSE and MAE values (4.69M and 2.56M USD respectively) indicate that the model was able to predict most funding rounds within a 10–15% error margin on log-normalized scale—an acceptable benchmark for noisy financial data [24], [25]. When analyzed in sectoral disaggregation, AI/Data, HealthTech, and EdTech displayed the highest predictability, suggesting that funding norms in these verticals are more structured and follow established investment trajectories. In contrast, FinTech and E-commerce exhibited irregular patterns, reflecting dependence on external qualitative factors such as policy regulation, consumer adoption cycles, and investor risk tolerance.

B. Sectoral and Structural Insights

AI/Data Startups: Funding rounds are generally stage-consistent and investor-driven; data maturity strongly correlates with capital inflow.

HealthTech and EdTech: Both sectors show mid-range variance and moderate predictability due to standardized valuation frameworks and growing digital penetration in India.

FinTech: Demonstrates heavy-tailed variance; large ticket sizes cause difficulty in modeling due to outliers and inconsistent investor behavior.

E-commerce: Largely unpredictable owing to wide fluctuations in consumer sentiment, discount-driven models, and high burn rates.

C. Comparative Model Behavior

GBT Model: Provides robust baseline performance and interpretability but cannot capture deep nonlinear cross-feature interactions.

One-Hot Neural Network: Excels in global accuracy by learning nonlinearities across categorical and numeric features.

Hybrid Neural Network: Trades off minimal accuracy for superior generalization, scalable embeddings, and meaningful semantic clustering of categorical variables.

D. Interpretability of Results

Embedding visualizations revealed that semantically similar industries cluster together—AI/Data with SaaS and Analytics, HealthTech near MedTech, and FinTech forming distinct isolated nodes. These latent structures suggest the model learned sector-level funding semantics rather than memorizing numeric patterns, supporting the generalization hypothesis [26].

VII. CHALLENGES FACED

A. Data Quality and Representation

Incomplete Public Data: Indian startup datasets lack standardized reporting across funding rounds. Many entries omit investor counts or stage definitions, necessitating manual normalization.

Outlier Sensitivity: Funding values span six orders of magnitude; conventional outlier detection techniques like IQR were insufficient, requiring hybrid IQR–log Z-score filtering.

Categorical Fragmentation: The industry_vertical field contained over 200 unique labels. Without grouping, this caused severe feature sparsity. The introduction of SectorBucket and RarityFolding mitigated this but led to partial information loss.

B. Computational and Modeling Challenges

Hardware Limitations: The experiments were executed on a CPU-based MacBook Air M4; GPU acceleration could have reduced convergence time and allowed deeper architectures.

Model Instability: Neural models, particularly early MLP versions, exhibited overfitting beyond 40 epochs. This was countered using dropout layers and early stopping criteria.

Hyperparameter Tuning: Manual tuning of learning rate, embedding dimension, and batch size was computationally expensive. Bayesian optimization could enhance performance in future iterations.

C. Conceptual and Methodological Challenges

Target Skewness: The log-transformation stabilized variance but also compressed variance for large funding rounds, underestimating multi-million investments.

Limited Observable Predictors: Public structured data could explain only ∼50% of funding variance. The remaining variance arises from qualitative elements—founder background, investor reputation, and startup traction—currently unmodeled.

Cross-Validation Bias: Stratified cross-validation was used to maintain class balance across funding stages; however, sectoral overlap sometimes caused minor leakage of distributional information.

VIII. CONCLUSION

This research developed an end-to-end comparative machine learning framework for predicting startup funding amounts in the Indian ecosystem, integrating rigorous preprocessing, engineered feature construction, and neural modeling. The study's core findings demonstrate that:

- Neural Networks outperform tree-based regressors in modeling complex financial and categorical interdependencies.
- Structured public data alone can account for approximately 50% of the variance in startup funding amounts ($R^2 \approx 0.50$ in log-space).
- Sectoral generalization is feasible—industries with defined valuation metrics (AI/Data, HealthTech, EdTech) show consistent learnability, whereas FinTech and E-commerce remain highly volatile.
- Embedding-based models (Hybrid NN) provide an interpretable middle ground between classical ML and deep learning, preserving semantic structure and improving cross-domain adaptability.

The implications are twofold: (1) the methodology can guide investors and policymakers in quantifying capital distribution tendencies, and (2) it sets a reproducible benchmark for structured startup analytics in emerging economies.

IX. FUTURE WORK

A. Data Enrichment and Feature Expansion

Future work should incorporate unstructured and relational signals that better represent qualitative aspects of startups. This includes founder profiles, company mission statements, patent filings, and press coverage encoded through transformer-based embeddings (e.g., BERT, FinBERT). Adding investor reputation embeddings and founder track-record vectors could capture latent success probabilities not visible in current data.

B. Advanced Modeling Extensions

Graph Neural Networks (GNNs): Represent startups, investors, and sectors as interconnected nodes to learn structural capital flows.

Temporal Modelling: Implement Recurrent or Transformer-based sequence models to learn macroeconomic cycles influencing funding rounds.

Ensemble Meta-Learning: Combine predictions from hybrid neural networks, boosting algorithms, and probabilistic regressors using stacked generalization to reduce bias.

C. Model Explainability and Deployment

Integrating Explainable AI (XAI) frameworks such as SHAP and LIME will help interpret feature importance and improve stakeholder trust. A real-time decision-support dashboard can be deployed for venture analysts to simulate hypothetical startup attributes and observe predicted funding outcomes.

D. Policy and Research Implications

From a broader perspective, the study lays groundwork for data-driven entrepreneurship policy in India. By identifying predictable sectors, the government and funding agencies can allocate support to underfunded yet high-potential industries. Moreover, incorporating longitudinal data will facilitate research on how funding predictability evolves with ecosystem maturity.

References

[1] M. Bidgoli, A. Rahimi, and K. Farahani, "Predicting Startup Success Using Random Forest and XGBoost on Structured Financial Data," *Journal of Computational Finance and Analytics*, vol. 8, pp. 102–115, 2024.

[2] Y. Qiu, J. Chen, and R. Zhang, "Social Media Sentiment Integration for Startup Financing Prediction," *IEEE Transactions on Computational Social Systems*, vol. 12, no. 1, pp. 45–58, Jan. 2025.

[3] M. Mashhadi, S. Azizi, and H. Rastegar, "Interpretable Machine Learning for Predicting Startup Funding and Exit Outcomes Using SHAP," *Expert Systems with Applications*, vol. 240, 2025.

[4] S. Jafari, L. Khosravi, and M. T. Khalil, "SAISE: A Social–Asset–Intellectual Scoring Framework for Startup Evaluation," *IEEE Access*, vol. 13, pp. 11521–11535, 2025.

[5] A. Maarouf, K. El Amrani, and T. Zohdy, "Fused Large Language Model for Predicting Startup Success from Textual and Numeric Data," *Procedia Computer Science*, vol. 236, pp. 412–420, 2024.

[6] T. Lyonnet and J. Stern, "Machine Learning Models of Venture Capital Decision-Making: Investor–Sector Interactions," *Venture Capital Review*, vol. 31, no. 2, pp. 201–214, 2024.

[7] J. Veloso, "Predicting Startup Funding Amounts Using Gradient Boosted Trees: Evidence from U.S. Crunchbase Data," *Applied Economics Letters*, vol. 29, no. 18, pp. 1651–1658, 2022.

[8] O. Unal and R. Ceasu, "Startup Survival Prediction Using Machine Learning: A Random Forest Approach," *Procedia Computer Science*, vol. 162, pp. 758–767, 2019.

[9] Y. Zhang, J. Liu, and T. Ma, "Crowdfunding Outcome Prediction via Temporal Social Network Modeling," *IEEE Transactions on Computational Social Systems*, vol. 4, no. 3, pp. 145–156, 2017.

[10] P. Gil, "Enhancing Model Generalization in European Startup Data via Principal Component Analysis," *European Journal of Business Analytics*, vol. 7, no. 1, pp. 25–39, 2023.

[11] A. Fidder, "Investor Diversity and Team Size as Predictors of Startup Funding Success," *Journal of Business Research*, vol. 158, pp. 112–125, 2024.

[12] B. Van Hoye and R. Thomaes, "XGBoost Models for Startup Performance Prediction: Evidence from Belgian Startups," *Information Systems Frontiers*, vol. 26, no. 2, pp. 601–615, 2024.

[13] S. Saghafian and M. Parhizkar, "Machine Learning Valuation Models Incorporating Investor Reputation," *International Journal of Financial Engineering*, vol. 6, no. 4, pp. 1950021, 2019.

[14] P. Gompers and J. Lerner, *The Venture Capital Cycle*, 3rd ed. Cambridge, MA: MIT Press, 2020.

[15] D. H. Hsu, "What Do Entrepreneurs Pay for Venture Capital Affiliation?," *The Journal of Finance*, vol. 59, no. 4, pp. 1805–1844, 2004.

[16] L. Rezaei, M. Santos, and Z. Chen, "Deep Learning Ensembles for Startup Funding Prediction Using Crunchbase Data," *IEEE Access*, vol. 11, pp. 98542–98555, 2023.

[17] R. Chen, K. Xu, and Y. Zhou, "Graph Embedding Models for Investor–Founder Relationship Prediction," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 15, no. 6, pp. 1–22, 2021.

[18] T. Nguyen, A. Wang, and M. Patel, "AutoML for Structured Financial Data: Improving Model Selection for Startup Prediction," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 4, pp. 512–524, 2022.

[19] T. Santos and L. Rezaei, "Deep Neural Networks vs. Gradient Boosting in Venture Capital Forecasting," *Data Science Review*, vol. 14, no. 3, pp. 225–239, 2023.

[20] R. Bhattacharya, S. Menon, and A. Iyer, "Funding Prediction in India and Southeast Asia Using Explainable Machine Learning," *Asia–Pacific Journal of Innovation and Entrepreneurship*, vol. 19, no. 1, pp. 87–103, 2025.

[21] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2017.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[23] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A New Oversampling Method in Imbalanced Data Sets," in *Advances in Intelligent Computing*, pp. 878–887, 2005.

[24] Z. Chen and A. Nguyen, "Hybrid Machine Learning Pipelines for Financial Forecasting," *IEEE Access*, vol. 11, pp. 84321–84334, 2023.

[25] T. Santos and L. Rezaei, "Explainable Regression Models for Venture Funding," *Data Science Review*, vol. 15, no. 2, pp. 145–159, 2024.

[26] G. E. P. Box and D. R. Cox, "An Analysis of Transformations," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 26, no. 2, pp. 211–252, 1964.

[27] Y. Wang and M. Li, "Evaluation Metrics for Financial Regression Models," *Applied Intelligence*, vol. 54, no. 5, pp. 11223–11237, 2024.