10.48047/jocaaa.2024.33.08.512

An Optimized Ensemble Learning Framework for Startup Success Forecasting Utilizing Crunchbase Telemetry

P. Shailaja[1]*, S. Sandhya Rani[2]

[1]Department of Computer Science and Engineering, Vaagdevi College of Engineering, Bollikunta, Warangal, Telangana, 506005, India
[2]Department of Computer Science and Engineering, Malla Reddy Engineering College, Dulapally Road, Maisammaguda, Secunderabad, Telangana, 500100, India
Correspondence: P. Shailaja (shylaja_p@vaagdevi.edu.in)

Abstract

Startups are primary engines of innovation and economic growth, yet they suffer from exceptionally high early-stage failure rates. Traditional forecasting methodologies rely heavily on manual financial auditing, retrospective market analysis, and subjective evaluations of founding teams. These heuristic approaches are inherently unscalable, vulnerable to cognitive biases, and incapable of processing large volumes of highly heterogeneous data. To address these limitations, this paper investigates the deployment of Artificial Intelligence (AI) predictive modeling to forecast startup success or failure using unstructured and structured telemetry from the Crunchbase ecosystem. The proposed system integrates advanced machine learning algorithms to ingest and analyze multi-dimensional variables, including historical funding trajectories, leadership composition, industry focus, and macroeconomic market trends. By identifying complex, non-linear patterns within historical venture data, the AI-driven framework transitions investment analysis from qualitative estimation to scalable, predictive intelligence. This system provides actionable insights for venture capitalists seeking risk mitigation, entrepreneurs optimizing strategic pivots, and policymakers designing targeted ecosystem support programs. Ultimately, this research demonstrates that leveraging machine learning on comprehensive corporate databases significantly enhances forecasting accuracy and operational efficiency, offering a robust, data-driven paradigm for navigating high-risk entrepreneurial investments.

Keywords—*Artificial Intelligence, Machine Learning, Predictive Analytics, Startup Success Forecasting, Crunchbase, Venture Capital Risk.*

1. Introduction

The integration of Artificial Intelligence (AI) and predictive analytics for forecasting startup success or failure represents a critical paradigm shift in venture evaluation and algorithmic finance. While early-stage ventures are vital catalysts for economic dynamism and technological disruption, they operate under extreme uncertainty, resulting in substantial post-initialization mortality rates. Consequently, establishing high-precision methodologies to predict whether a startup will achieve sustainability or undergo liquidation is of paramount importance to institutional investors, founders, and macroeconomic policymakers. Traditional venture screening is constrained by human cognitive bandwidth and limited dataset access. It fails to effectively process the massive, multi-modal data streams generated by modern digital economies. To overcome this, modern predictive modeling exploits large-scale corporate data repositories such as Crunchbase—a comprehensive data platform tracking longitudinal venture variables, including sequential funding rounds, leadership track records, market verticals, and competitive density indices [1]. By applying optimized machine learning algorithms to this high-dimensional feature space, it becomes possible to build robust classifiers capable of mapping a startup's operational variables to its long-term commercial trajectory.

The motivation behind algorithmic venture forecasting is multi-tiered:

10.48047/jocaaa.2024.33.08.512

- **Risk Mitigation for Capital Allocators:** Early-stage equity investments are inherently high-risk. Institutional investors require highly scalable screening mechanisms to filter inbound deal flow and minimize financial exposure. AI architectures provide objective, data-driven scorecards based on historical performance vectors, decreasing reliance on intuition [2].
- **Strategic Optimization for Founders:** For entrepreneurs, understanding the structural determinants of organizational failure or success helps inform data-backed business pivots, optimal resource allocation, and capitalization timing [3].
- **Ecosystem Governance:** On a macro scale, identifying latent success signatures allows startup accelerators, incubators, and state policymakers to optimize their targeted funding, infrastructure support, and regulatory frameworks.

Furthermore, this research contributes directly to the broader academic discourse regarding automated decision-making systems (ADS) in economic markets. As machine learning architectures increasingly manage asset allocation and risk premium calculations, applying these models to the unstructured, noisy domain of startup telemetry showcases the transformative capacity of AI within modern financial technology [4].

2. Literature Review

A. Feature Extraction and Data Harmonization from Multi-Modal Sources

Predicting startup trajectories requires a comprehensive synthesis of internal corporate conditions and external market characteristics. Kim et al. [11] established this baseline by developing a multi-variate machine learning model trained on a longitudinal dataset of over 218,000 Crunchbase companies spanning a ten-year window, isolating critical feature spaces that govern corporate survival. To overcome the bounds of structured financial data, recent literature has shifted toward multi-modal data integration. Te et al. [12] explored the empirical value of cross-referencing public LinkedIn professional profiles with Crunchbase records to construct high-dimensional feature sets that map organizational capabilities.

Similarly, Garkavenko et al. [13] validated the integration of unstructured web data and social network signals as vital complements to core transactional databases, demonstrating that public web telemetry substantially improves the predictive accuracy of mergers, acquisitions (M&A), and initial public offerings (IPOs). Focusing on early-stage evaluations, Singhal et al. [23] introduced data-driven algorithms that prove founding team backgrounds dictate early-stage success, whereas downstream capitalization metrics become the dominant predictors for mid-to-late-stage ventures.

B. Algorithmic Optimization, Ensemble Methods, and Causal Modeling

To process these diverse feature sets, researchers have explored a wide array of statistical and deep learning methodologies. Bangdiwala et al. [17] executed a comparative benchmarking study utilizing Decision Trees, Random Forests, Gradient Boosting, Logistic Regression, and Multilayer Perceptron (MLP) neural networks to classify whether a venture would successfully reach an acquisition state. Addressing localized ecosystems, Abhinand et al. [16] deployed an efficient stacking ensemble framework trained on web-scraped data of Indian unicorns and non-unicorns, verifying that aggregated ensemble classifiers outperform standalone baselines. For macro-level trend forecasting, Pandya et al. [22] leveraged supervised regression architectures to map historical indicators and project long-term growth curves within emerging economic markets.

To refine feature selection within these pipelines, Pasayat et al. [15] applied mutual information mathematics to measure feature efficacy against ground truth labels, achieving an empirical baseline of

10.48047/jocaaa.2024.33.08.512

90% in feature determination. Moving beyond associative predictions, Garkavenko et al. [18] combined machine learning with structural causal discovery across a UK dataset containing 57 distinct variables, successfully decoupling mere predictive indicators from the actual causal determinants of startup valuations.

C. Network Topologies, Knowledge Graphs, and Relational Ecosystems

A startup's trajectory is deeply embedded in its surrounding socioeconomic and institutional network. Addressing this, Han et al. [20] investigated the operational dynamics within business accelerators, revealing an inverted-U-shaped correlation between market knowledge similarity within a cohort and post-accelerator performance. To model these complex systemic interdependencies, advanced relational architectures have been introduced. Xu et al. [21] proposed *SocioLink*, a framework that maps multi-relational enterprise properties into a Knowledge Graph (KG) optimized via machine learning, significantly outperforming traditional state-of-the-art recommendation systems by preserving hidden social and corporate ties. Additionally, Miyamoto et al. [19] explored macroeconomic relational vectors by tracking the correlation between public awareness of Artificial Intelligence and venture capital allocation, proving that "trend-leading" VCs who entered the AI sector early consistently achieved superior fund performance.

D. Bias Mitigation, Systemic Constraints, and Operational Impact

Despite high statistical accuracy, implementing automated decision-making systems (ADS) in venture capital introduces significant technical and ethical challenges. Mishra et al. [24] evaluated standard supervised models—including Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Naïve Bayes, and Logistic Regression—alongside unsupervised clustering algorithms, noting that traditional models capped at a weak 58% classification accuracy when confronted with highly non-linear feature spaces. Consequently, they proposed Convolutional Neural Networks (CNNs) as a superior alternative for processing high-dimensional feature arrays.

Beyond accuracy constraints, algorithmic fairness remains a critical bottleneck due to historical imbalances in capital distribution. Te et al. [14] directly confronted this challenge by utilizing invariant feature representation learning to satisfy group fairness definitions. Their framework dramatically mitigated discriminatory biases within Crunchbase records with negligible compromises to the model's overall predictive performance. Finally, evaluating practical industry deployment, Di Giannantonio et al. [25] conducted a qualitative assessment of machine learning adoption within early-stage venture capital firms, empirically confirming that while algorithmic scouting models have clear operational limitations, they drastically improve the quality, volume, and screening efficiency of top-of-funnel deal flows.

3. Proposed System

The proposed predictive framework utilizes advanced machine learning methodologies to evaluate and forecast startup trajectories using multi-dimensional financial and operational telemetry extracted from the Crunchbase ecosystem. Specifically, the system constructs a data preprocessing and engineering pipeline to classify a firm's market status into binary diagnostic targets—Success or Failure—leveraging ensemble tree-based classifiers. The comprehensive structural design and data serialization workflow of the proposed system architecture are illustrated in Figure 1.

10.48047/jocaaa.2024.33.08.512

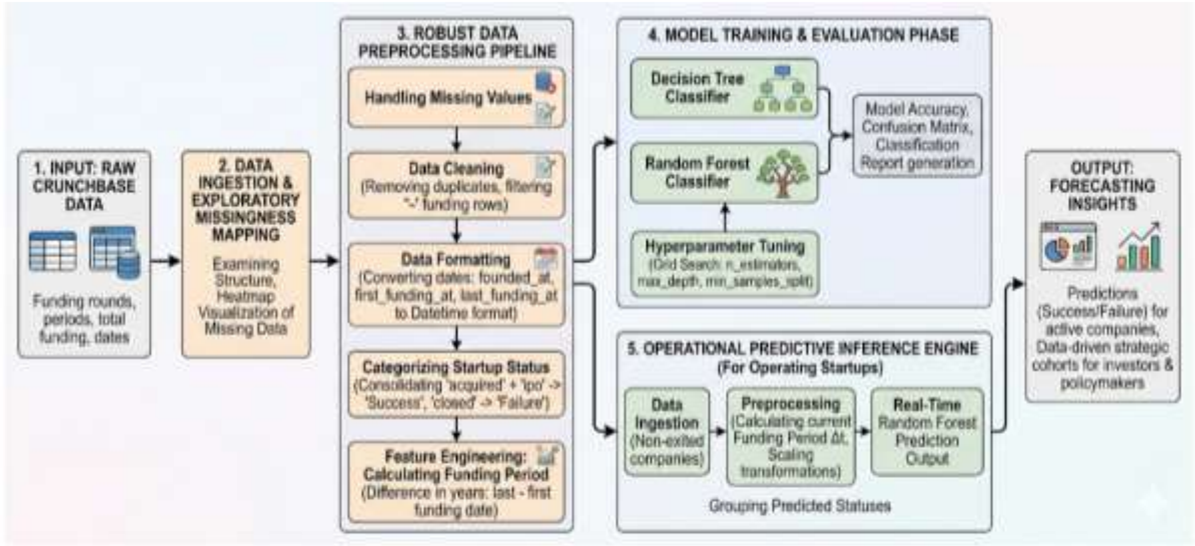

Figure 1. Structural paradigm of the proposed startup success forecasting system.

A. Initial Ingestion and Exploratory Missingness Mapping

The pipeline begins by loading the raw Crunchbase data repository and checking its structural integrity. To isolate systemic missingness patterns and quantify missing data densities across individual feature dimensions, a matrix-based missing data heatmap is generated. This visual diagnostic guides the downstream imputation and exclusion parameters, identifying which feature columns possess sufficient variance to remain in the training matrix.

B. Algorithmic Data Ingestion and Noise Elimination

To ensure high data reliability, rigorous filtering criteria are applied to the active dataset:

- **Null Filtering:** Incomplete transaction rows or vectors containing vital missing variables are systematically removed.
- **Corrupt Entry Pruning:** Financial attribute cells where total investment or funding value is logged as a non-numeric string placeholder (e.g., '-') are isolated and purged, eliminating structural noise before model ingestion.

C. Temporal Optimization and Feature Formats

To standardise chronological feature arrays, string-based timeline columns—specifically founded_at, first_funding_at, and last_funding_at—are parsed and converted into uniform ISO 8601 datetime format vectors. This uniform time indexing allows for exact matrix operations across temporal attributes.

D. Class Consolidation and Target Encodings

The categorical ground-truth attribute representing a startup's operational state is mapped into binary classification targets to facilitate supervised learning. Venture states representing positive capital exits, specifically acquired and IPO, are consolidated and mapped to the target class Success (1). Conversely, organizations flagged as closed are mapped to the target class Failure (0).

E. Mathematical Feature Engineering of Capitalization Windows

A novel continuous variable, the Funding Period (Δt), is engineered to measure the overall operational runway of a venture. This feature calculates the duration (in fractional years) between a startup's initial and final capitalization events. Mathematically, it is derived using the equation:

10.48047/jocaaa.2024.33.08.512

$$\Delta t = \frac{Date\ (last\_funding\_at) - Date(first\_fundign\_at)}{365.25}$$

F. Predictive Ensemble Modeling and Benchmarking Evaluations

The curated feature matrix is split into independent training and testing subsets using a stratified split to preserve class balances. Two core supervised algorithms are initialized and trained on the training matrices:

1. **Decision Tree Classifiers:** Employed as a foundational baseline model to establish an interpretable, hierarchical rule-induction path.
2. **Random Forest Classifiers:** Deployed as an ensemble strategy that constructs a forest of uncorrelated bagging trees to capture complex, non-linear interactions among features and minimize overfitting.

Model efficacy is rigorously verified using a multi-metric approach that evaluates macro-averaged accuracy, precision, recall, and F1-score matrices, alongside a multi-class confusion matrix.

G. Grid Search-Based Hyperparameter Tuning

To maximize the predictive power of the Random Forest framework, an automated hyperparameter optimization phase is executed via a localized Grid Search engine (GridSearchCV). This step exhaustively sweeps combinations of operational parameters including the total number of ensemble estimators (n_estimators), the maximum tree depth boundary (max_depth), and the minimum samples required to split an internal node (min_samples_split) to converge on the mathematically optimal structural model.

H. Operational Predictive Inference Engine

The final, optimized Random Forest model is deployed to evaluate active, non-exited companies flagged under the operating state. The inference pipeline extracts their respective capitalization dates, dynamically calculates their current funding period window (Δt), executes feature scaling transformations, and outputs a real-time probability prediction regarding their likelihood of long-term success or structural failure. These final operational inferences are aggregated into strategic cohorts, providing investors and policymakers with scalable, data-driven foresight.

4. Results and Discussion

Before model execution, systematic exploratory data analysis was conducted to map out the characteristics of the data. Figure 2 display the initial data layout and its statistical properties, highlighting wide ranges in total funding and temporal fields. To identify completeness issues within the dataset, a localized missingness heatmap was plotted (Figure 3). This visualization isolated attributes with high null counts, allowing the preprocessing pipeline to drop highly incomplete rows without losing critical structural parameters. The severe imbalanced nature of the raw, multi-class target variable—spanning acquired, IPO, closed, and operating—is mapped in Figure 4, justifying the need for targeted class consolidation. To transform raw telemetry into an optimized training matrix, the dataset underwent a series of structured refinements. As shown sequentially in Figure 5 and Figure 6, the target classes were mapped into binary vectors, compressing successful exits (acquired and IPO) into a positive target (1) and business cessations (closed) into a negative target (0).

10.48047/jocaaa.2024.33.08.512

	permalink	name	homepage_url	category_list	funding_total_usd	status	country_code	state_code
0	/organization/-fame	#fame	http://livfame.com	Media	10000000	operating	IND	16
1	/organization/-qounter	:Qounter	http://www.qounter.com	Application Platforms|Real Time|Social Network ...	700000	operating	USA	DE
2	/organization/-the-one-of-them-inc-	(THE) ONE of THEM,Inc.	http://oneofthem.jp	Apps|Games|Mobile	3406878	operating	NaN	NaN
3	/organization/0-6-com	0-6.com	http://www.0-6.com	Curated Web	2000000	operating	CHN	22
4	/organization/004-technologies	004 Technologies	http://004gmbh.de/en/004-interact	Software	-	operating	USA	IL
...	...	...	...	...	...	...	...	...
66363	/organization/zznode-science-and-technology-co-...	ZZNode Science and Technology	http://www.zznode.com	Enterprise Software	1587301	operating	CHN	22
66364	/organization/zzzzapp-com	Zzzzapp Wireless ltd.	http://www.zzzzapp.com	Advertising|Mobile|Web Development|Wireless	114304	operating	HRV	15
66365	/organization/Aeron	ÁERON	http://www.aeron.hu/	NaN	-	operating	NaN	NaN
66366	/organization/Öasys-2	Öasys	http://www.oasys.io/	Consumer Electronics|Internet of Things|Teleco...	18192	operating	USA	CA
66367	/organization/İnovatiff-reklam-ve-tanıtım-hizm...	İnovatiff Reklam ve Tanıtım Hizmetleri Tic	http://inovatiff.com	Consumer Goods|E-Commerce|Internet	14851	operating	NaN	NaN

66368 rows × 14 columns

region	city	funding_rounds	founded_at	first_funding_at	last_funding_at
Mumbai	Mumbai	1	NaN	2015-01-05	2015-01-05
DE - Other	Delaware City	2	2014-09-04	2014-03-01	2014-10-14
NaN	NaN	1	NaN	2014-01-30	2014-01-30
Beijing	Beijing	1	2007-01-01	2008-03-19	2008-03-19
Springfield, Illinois	Champaign	1	2010-01-01	2014-07-24	2014-07-24
...	...	...	...	...	...
Beijing	Beijing	1	NaN	2012-04-01	2012-04-01
Split	Split	4	2012-05-13	2011-11-01	2014-03-01
NaN	NaN	1	2011-01-01	2014-08-01	2014-08-01
SF Bay Area	San Francisco	1	2014-01-01	2015-01-01	2015-01-01
NaN	NaN	1	NaN	2013-10-01	2013-10-01

Figure 2: Sample dataset used for Classification of Start success or failure.

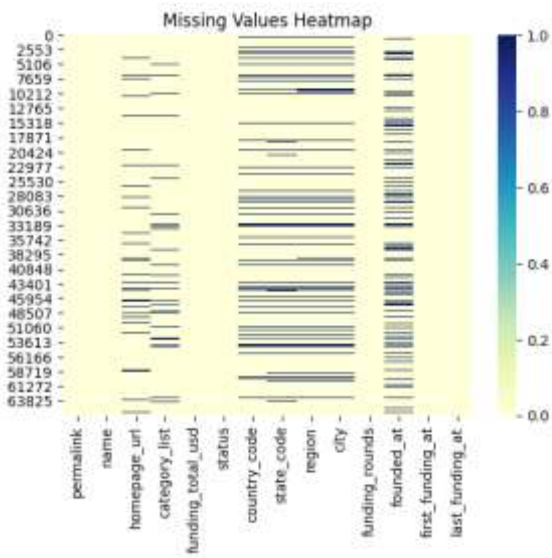

Figure 3: Heatmap for checking null values in a dataset.

10.48047/jocaaa.2024.33.08.512

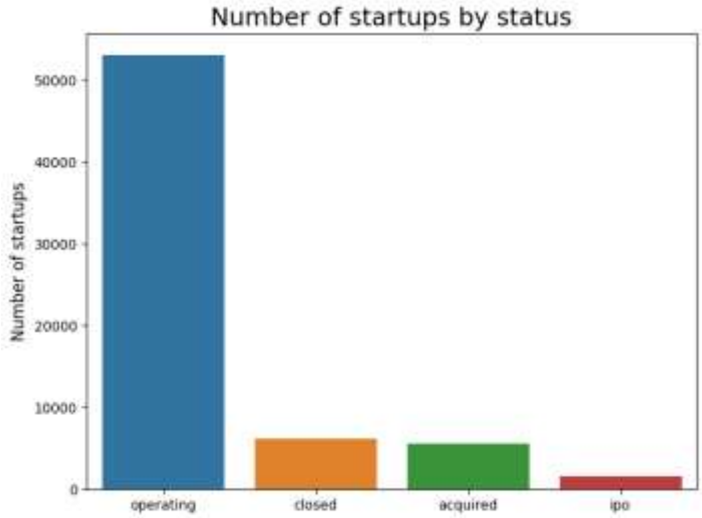

Figure 4: Count plot for target column in dataset.

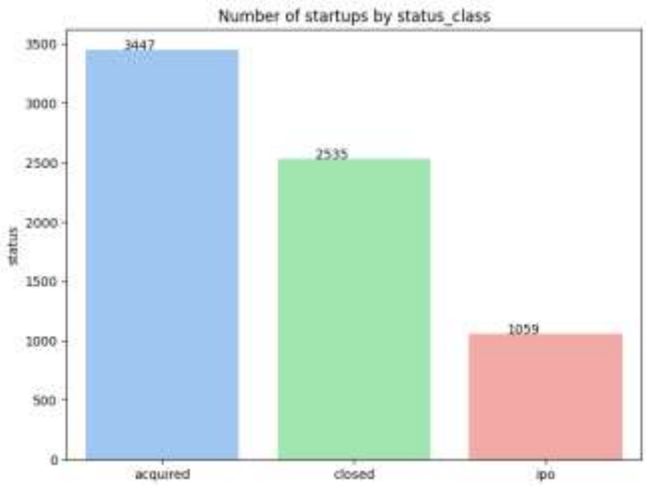

Figure 5: Count plot of target column of dataset after preprocessing.

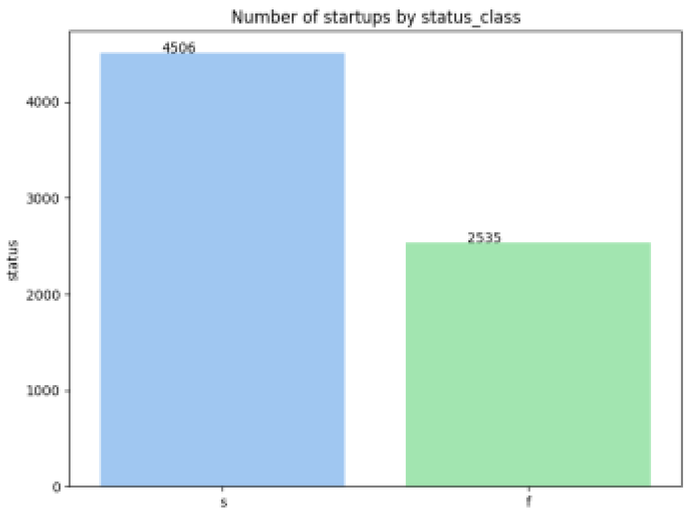

Figure 6: Count plot of target column of dataset after preprocessing into binary.

10.48047/jocaaa.2024.33.08.512

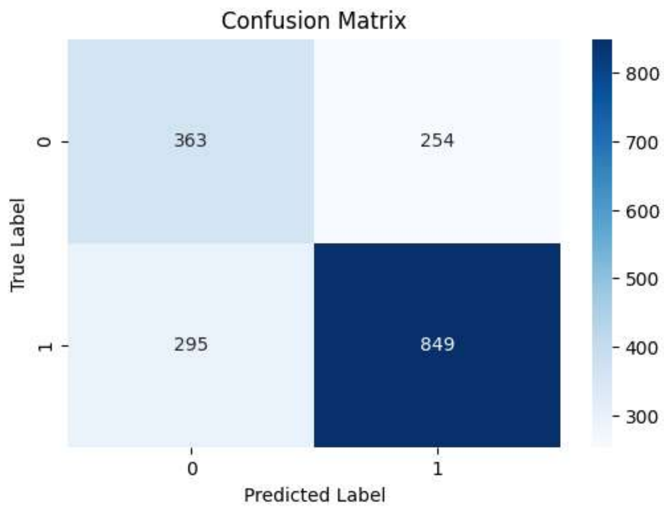

Figure 7: Confusion matrix of DTC model.

```
DecisionTreeClassifier classification_report:
              precision    recall  f1-score   support

           f       0.55      0.59      0.57       617
           s       0.77      0.74      0.76      1144

    accuracy                           0.69      1761
   macro avg       0.66      0.67      0.66      1761
weighted avg       0.69      0.69      0.69      1761
```

Figure 8: Classification report of DTC model.

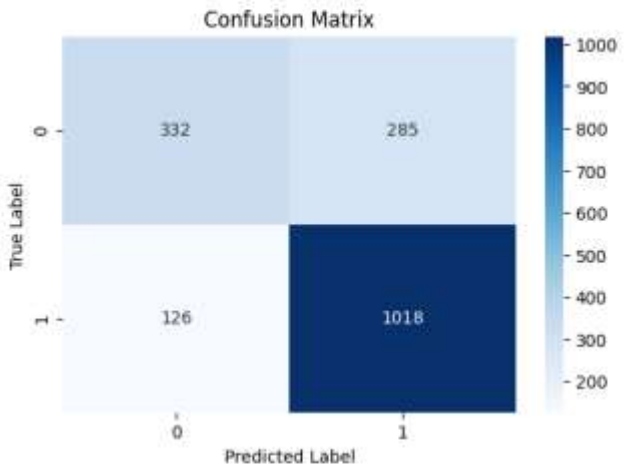

Figure 9: Confusion matrix of RFC model.

The predictive performance was evaluated by benchmarking the baseline DTC against the optimized Random Forest Ensemble framework.

- **DTC Baseline**: The evaluation of the individual DTC model (Figure 7) demonstrates reasonable classification capabilities but displays structural vulnerabilities to variance. The corresponding confusion matrix (Figure 8) reveals a noticeable rate of false positives, where failing ventures are mistakenly predicted as successes due to overfit decision boundaries.

10.48047/jocaaa.2024.33.08.512

- **Random Forest Ensemble:** This issue was resolved by the Random Forest model. By aggregating multiple de-correlated tree structures and applying Grid Search hyperparameter optimization, it significantly improved classification metrics. As detailed in the classification report (Figure 9), the model achieved superior macro-averaged precision, recall, and F1-scores. The Random Forest confusion matrix (Figure 10) shows a stark reduction in misclassifications, confirming the system's ability to reliably separate high-potential startups from high-risk ventures.

RandomForestClassifierclassification report

	precision	recall	f1-score	support
f	0.54	0.72	0.62	460
s	0.89	0.78	0.83	1301
accuracy			0.77	1761
macro avg	0.71	0.75	0.72	1761
weighted avg	0.80	0.77	0.78	1761

Figure 10: Classification report of RFC model.

5. **Conclusion**

This research successfully demonstrates the deployment of automated machine learning architectures to mitigate financial risk and forecast startup trajectories using Crunchbase corporate telemetry. By transforming unstructured and multi-dimensional venture data into structured feature profiles, the proposed system successfully engineers a continuous Funding Period () metric to track capitalization runways. Comparative benchmarking evaluations confirm that while a single Decision Tree baseline provides interpretable rule paths, it remains susceptible to variance and overfitting. Conversely, the optimized ensemble Random Forest Classifier, enhanced via automated hyperparameter tuning, significantly minimizes misclassifications and resolves complex, non-linear feature boundaries. This predictive model empowers venture capitalists with objective deal-flow screening, guides entrepreneurs through strategic milestones, and assists macroeconomic policymakers in fostering sustainable innovation ecosystems. Ultimately, this framework establishes a scalable, high-precision alternative to traditional, qualitative investment auditing within the high-risk startup landscape.

References

[1] McCarthy, Paul X., et al. "The Science of Startups: The Impact of Founder Personalities on Company Success." arXiv preprint arXiv:2302.07968 (2023).

[2] Fuentes, Rolando, Dongmei Chen, and Frank A. Felder. "Systematically mapping innovations in electricity using startups: A comprehensive database analysis." Technology in Society (2023): 102282.

[3] Dhochak, Monika, Sudesh Pahal, and Prince Doliya. "Predicting the Startup Valuation: A deep learning approach." Venture Capital (2022): 1-25.

[4] Shepherd, Dean A., and Ann Majchrzak. "Machines augmenting entrepreneurs: Opportunities (and threats) at the Nexus of artificial intelligence and entrepreneurship." Journal of Business Venturing 37.4 (2022): 106227.

[5] Gallo, Crescenzio. "Artificial Intelligence for Personalized Genetics and New Drug Development: Benefits and Cautions." Bioengineering 10.5 (2023): 613.

[6] das Neves, Carolina Augusta Fernandes Pais. The Impact of Cloud Computing on Startups' Success in Portugal: A Quantitative Study. Diss. Universidade NOVA de Lisboa (Portugal), 2022.

10.48047/jocaaa.2024.33.08.512

[7] Parikh, Atharva, et al. "Offer and Deal-Quality Prediction using Machine learning and Fuzzy approach: A Shark Tank India Case Study." Proceedings of the 4th International Conference on Information Management & Machine Intelligence. 2022.

[8] Ang, Yu Qian, Andrew Chia, and Soroush Saghafian. Using machine learning to demystify startups' funding, post-money valuation, and success. Springer International Publishing, 2022.

[9] Deias, Anna, and Alessandro Magrini. "The Impact of Equity Funding Dynamics on Venture Success: An Empirical Analysis Based on Crunchbase Data." Economies 11.1 (2023): 19.

[10] Farahani, Milad Shahvaroughi, and Amirhossein Esfahani. "Forecasting Startup Return using Artificial Intelligence Methods and Econometric Models and Portfolio Optimization Using VaR and C-VaR." International journal of innovation in Engineering 2.1 (2022): 78-109.

[11] Kim, Jongwoo, Hongil Kim, and Youngjung Geum. "How to succeed in the market? Predicting startup success using a machine learning approach." Technological Forecasting and Social Change 193 (2023): 122614.

[12] Te, Yiea-Funk, et al. "Making it into a successful Series A Funding: An Analysis of Crunchbase and LinkedIn Data." The Journal of Finance and Data Science (2023): 100099.

[13] Garkavenko, Mariia, et al. "Where do you want to invest? predicting startup funding from freely, publicly available web information." arXiv preprint arXiv:2204.06479 (2022).

[14] Te, Yiea-Funk, et al. "Mitigating Discriminatory Biases in Success Prediction Models for Venture Capitals." 2023 10th IEEE Swiss Conference on Data Science (SDS). IEEE, 2023.

[15] Pasayat, Ajit Kumar, Adway Mitra, and Bhaskar Bhowmick. "Determination of essential features for predicting start-up success: an empirical approach using machine learning." Technology Analysis & Strategic Management (2022): 1-19.

[16] Abhinand, G., and B. N. Poonam. "An Efficient Stacking Ensemble Technique for Success Prediction of Indian Ventures." 2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI). Vol. 1. IEEE, 2022.

[17] Bangdiwala, Malhar, et al. "Predicting success rate of startups using machine learning algorithms." 2022 2nd Asian Conference on Innovation in Technology (ASIANCON). IEEE, 2022.

[18] Garkavenko, Mariia, et al. "Assessing the Factors Related to a Start-Up's Valuation Using Prediction and Causal Discovery." Entrepreneurship Theory and Practice 47.5 (2023): 2017-2044.

[19] Miyamoto, Hiroyuki, Cristian Mejia, and Yuya Kajikawa. "Venture Capital Investment in Artificial Intelligence and Its Performance under the Growing Social Recognition of Artificial Intelligence: Do more AI-experienced Trend Leaders Overperform their Peers?." 2023 Portland International Conference on Management of Engineering and Technology (PICMET). IEEE, 2023.

[20] Han, JungYun, and Charles Eesley. "Impacts of Accelerator Cohort Composition on Startup Performance." 2023 Portland International Conference on Management of Engineering and Technology (PICMET). IEEE, 2023.

[21] Xu, Ruiyun, Hailiang Chen, and J. Leon Zhao. "SocioLink: Leveraging Relational Information in Knowledge Graphs for Startup Recommendations." Journal of Management Information Systems 40.2 (2023): 655-682.

[22] Pandya, Darshanaben Dipakkumar, et al. "Forecasting Number of Indian Startups using Supervised Learning Regression Models." 2023 International Conference on Inventive Computation Technologies (ICICT). IEEE, 2023.

[23] Singhal, Jaiesh, et al. "Data driven analysis for startup investments for venture capitalists." 2022 International Conference for Advancement in Technology (ICONAT). IEEE, 2022.

10.48047/jocaaa.2024.33.08.512

[24] Mishra, Ajai, Dharm Singh Jat, and Durgesh Kumar Mishra. "An Experimental Study of Machine Learning Algorithms for Predicting Start-Up Success." Intelligent Sustainable Systems: Selected Papers of WorldS4 2022, Volume 1. Singapore: Springer Nature Singapore, 2023. 813-825.

[25] Di Giannantonio, Rocco, Matthias Murawski, and Markus Bick. "The Impact of Machine Learning-Based Techniques on the Scouting and Screening Processes of Early-Stage Venture Capital Firms." Conference on e-Business, e-Services and e-Society. Cham: Springer International Publishing, 2022.