Predictive Opportunity Analysis (POA): A Machine Learning Framework for Techpreneurial Foresight and Innovation Strategy

John Edoh Onuh
University of Sunderland, UK
Email: bi51ma@student.sunderland.ac.uk

Abstract

This study introduces the Predictive Opportunity Analysis (POA) framework, an advanced, data-driven system designed to quantify and interpret opportunity potential within the technology entrepreneurship ecosystem. The persistent challenge of uncertainty in techpreneurial opportunity evaluation necessitates a shift from intuitive judgments to transparent, evidence-based foresight. This paper details the development and empirical validation of a multi-component predictive architecture that estimates startup success probabilities. The framework integrates a suite of ensemble machine learning algorithms—including Random Forest, XGBoost, LightGBM, and a meta-model—optimised with Optuna-based hyperparameter tuning. The analysis utilises the Kaggle Startup Success Prediction dataset to evaluate predictive accuracy and interpret feature significance through SHAP (SHapley Additive exPlanations) explainability. The results demonstrate the framework's strong predictive capacity within the scope of this dataset, with the LightGBM model achieving the highest performance, yielding a ROC-AUC of 0.8331 and an accuracy of 78.38%. The explainability analysis reveals that social capital (relationships) is the most influential predictor of startup success in this dataset, followed by financial resources (funding_total_usd) and execution capability (milestones). These findings offer empirical support for the Resource-Based View (RBV) of the firm, highlighting the salience of social capital in the observed venture outcomes. While generalisation beyond the studied context warrants caution, the findings contribute to the emerging field of computational entrepreneurship by offering a transparent, reproducible, and theoretically grounded framework for foresight and opportunity identification, with practical relevance for entrepreneurs, investors, and innovation policymakers.

Keywords: Techpreneurship, Predictive Analytics, Opportunity Analysis, Machine Learning; Innovation Foresight, Startup Success, Explainable AI

Introduction: The Imperative for Computational Foresight in Techpreneurship

The Paradox of High Potential and High Failure in Tech Ventures

Technological entrepreneurship, or “techpreneurship,” stands as a primary engine of innovation-led economic transformation and a critical driver of competitive advantage in the global economy. Ventures founded on emerging technologies create new markets, disrupt established industries, and generate societal value. However, this landscape is defined by a persistent paradox: while the potential rewards for successful tech ventures are immense, the ecosystem is characterised by extreme uncertainty and exceptionally high rates of failure. Empirical evidence consistently confirms this challenge: studies drawing on large-scale startup databases report failure rates exceeding 60–70% within the first five years of operation (Elia et al., 2020; Skala et al., 2020). Traditional approaches to entrepreneurial opportunity recognition, which rely heavily on intuition, pattern matching, and qualitative heuristics, have proven insufficient to navigate the complexities of modern digital markets. These methods are often subjective, susceptible to cognitive biases such as overconfidence and confirmation bias, and lack the scalability and objectivity required for rigorous opportunity evaluation. This deficiency results in a significant misallocation of capital, talent, and resources, stifling innovation and economic progress.

The Emergence of Computational Entrepreneurship

In response to these challenges, a new paradigm is emerging at the intersection of data science, artificial intelligence, and entrepreneurship studies: computational entrepreneurship. This nascent field advocates for the use of computational methods to analyse large-scale datasets, uncover hidden patterns in venture dynamics, and augment or replace traditional decision-making processes with data-driven insights. Recent empirical work by Chalmers et al. (2021) demonstrates that AI-enabled approaches can substantially improve the rigour and efficiency of entrepreneurial decision-making, while Ferreira et al. (2025) illustrate how generative AI frameworks are reshaping entrepreneurial exploration and opportunity framing. The International Journal of Technopreneurship and Innovation (IJTI) itself recognises the critical role of emerging technologies like artificial intelligence in shaping modern business and innovation ecosystems. By leveraging predictive analytics, computational entrepreneurship offers a pathway to transform opportunity evaluation from an intuitive art into a systematic science. This methodological shift promises to enhance the efficiency of startup ecosystems, enabling entrepreneurs to make more informed strategic choices, investors to build more resilient portfolios, and policymakers to design more effective support initiatives. This paper positions itself as a contribution to this emerging discipline, proposing a framework that operationalises its core principles within a reproducible empirical setting.

Research Gap: The "Black Box" Problem in Entrepreneurial Analytics

While prior research has explored the use of statistical and machine learning models to predict startup success, many of these efforts suffer from a critical limitation: a lack of interpretability. Advanced models, particularly those based on deep learning or complex ensembles, often function as "black boxes," providing accurate predictions without offering clear explanations for their outputs. As Joy and Elly (2025) compellingly argue, a prediction without an explanation is of limited value, as it fails to generate trust or actionable insights. For an investor, knowing that a startup has a 90% probability of failure is useful, but knowing *why*—whether due to insufficient funding, an inexperienced team, or an unfavorable market—is transformative. This "black box" problem represents a significant gap in the literature, hindering both the practical adoption of predictive tools and their ability to contribute to the theoretical understanding of entrepreneurship.

The POA Framework: A Contribution to Transparent Foresight

This study introduces and validates the Predictive Opportunity Analysis (POA) framework, a hybrid methodological model designed to address this critical research gap. The POA framework integrates high-performance predictive modeling with state-of-the-art Explainable AI (XAI) techniques to create a system that is not only accurate but also transparent and interpretable. The research is guided by three primary objectives:

1. To develop and validate a robust machine learning pipeline capable of predicting startup success outcomes with a high degree of accuracy.
2. To integrate Explainable AI, specifically SHAP (SHapley Additive exPlanations), to "open the black box" and identify the key features that drive venture success or failure, ensuring model transparency.
3. To propose a reproducible and theoretically grounded framework that bridges the divide between entrepreneurial theory and computational practice, providing a valuable tool for stakeholders across the innovation ecosystem.

By achieving these objectives, this paper offers both theoretical contributions to the study of entrepreneurship and practical tools for those engaged in the high-stakes process of innovation and venture creation. The remainder of this paper is structured as follows: Section 2 reviews the theoretical foundations of the study. Section 3 provides a detailed account of the research methodology. Section 4 presents the empirical results and discusses their theoretical and strategic implications. Finally, Section 5 concludes with a summary of the contributions, an acknowledgment of limitations, and an agenda for future research.

Theoretical Foundations: Integrating Entrepreneurial Theory with Data Science

Entrepreneurial Opportunity Recognition

The concept of entrepreneurial opportunity is the cornerstone of entrepreneurship research. The digital era, particularly with the advent of artificial intelligence (AI), has reshaped the landscape of entrepreneurship, empowering entrepreneurs to discover and exploit new opportunities through radical innovation (Obschonka & Audretsch, 2020). Their framework highlights that opportunities are distinct from the individuals who pursue them and that the evaluation process is a critical, yet under-studied, phase. Traditionally, this evaluation has been viewed through a cognitive lens, focusing on the heuristics and biases of the individual entrepreneur. However, the advent of the digital era has fundamentally altered the nature of this process. As Autio et al. (2018) observe, digital technologies create new "affordances" for entrepreneurs, enabling them to access and analyse vast amounts of data to inform their decisions. In this context, opportunity evaluation is no longer solely a cognitive exercise but an analytical one, where the ability to derive insights from data becomes a key entrepreneurial skill. Beyond the traditional cognitive view of opportunity recognition, two complementary theoretical streams are especially relevant to the POA framework. First, Sarasvathy's (2001) Effectuation Theory posits that expert entrepreneurs do not rely on fixed goals and predictive logic alone; rather, they iteratively co-create opportunities using available means under uncertainty. While the POA framework is grounded in a predictive (causal) paradigm, its emphasis on empirical feedback and adaptability resonates with effectuative logic, suggesting that computational tools can complement rather than replace expert entrepreneurial judgment. Second, the Dynamic Capabilities perspective (Teece, 2018) frames venture success as contingent on a firm's ability to sense, seize, and reconfigure resources in response to market change. The inclusion of milestone_density and temporal features in the POA feature set operationalises this dynamic dimension, extending the analysis beyond static resource endowment. Together, these theoretical perspectives anchor the POA framework within a broader multi-theoretic landscape of entrepreneurship scholarship, and underscore the importance of moving beyond single-theory explanations of startup success. The POA framework is built upon this integrative understanding, treating opportunity evaluation as a predictive modelling problem that can be systematically addressed through computational methods, while remaining theoretically grounded.

The Resource-Based View (RBV) as a Predictor of Venture Performance

To move beyond mere prediction and toward explanation, a robust theoretical lens is required to interpret the model's findings. The Resource-Based View (RBV) of the firm, a central theory in strategic management, provides such a lens. The RBV posits that a firm's ability to achieve and sustain a competitive advantage is determined by its portfolio of valuable, rare, inimitable, and non-substitutable (VRIN) resources (Barney et al., 2021). These resources can be tangible, such as financial capital and physical assets, or intangible, such as human capital (the skills and knowledge of the team), social capital (networks and relationships), and organizational knowledge.

In the context of tech startups, the RBV suggests that ventures possessing superior resources are more likely to succeed. For instance, a startup with significant venture capital funding (a tangible financial resource) can invest in product development, marketing, and talent acquisition more aggressively than its under-capitalised rivals. Similarly, a venture founded by a team with deep industry expertise and extensive professional networks (intangible human and social capital) is better equipped to navigate market challenges and seize opportunities. The POA framework operationalises these theoretical constructs by using measurable features—such as total_funding_usd and relationships—as proxies for a venture's resource endowments. This allows the machine learning model to empirically test the core propositions of the RBV. The subsequent interpretation of the model's feature importance, therefore, is not a purely statistical exercise but a form of theory validation, revealing which resources are most critical for success in the techpreneurial context.

The Methodological Frontier: Explainable AI (XAI) in Strategic Management

The increasing integration of machine learning into entrepreneurship and management research marks a significant methodological shift, enabling the quantitative measurement of complex dynamics that were previously only accessible through qualitative methods. However, as noted by Joy and Elly (2025), the complexity of these models often comes at the cost of transparency. This opacity presents a major barrier to both scholarly progress and practical application.

Explainable AI (XAI) has emerged as a critical subfield of artificial intelligence dedicated to addressing this challenge. XAI encompasses a suite of techniques designed to make the decisions of machine learning models understandable to humans. Among these, SHAP (SHapley Additive exPlanations) has become a leading method due to its strong theoretical foundation in cooperative game theory. SHAP calculates the marginal contribution of each feature to a specific prediction, providing a fair and consistent way to attribute the model's output to its inputs. By incorporating SHAP into the POA framework, this study moves beyond simply predicting *what* the outcome of a venture will be, to explaining *why* the model arrived at that prediction. This methodological innovation allows for a deeper engagement with theory, enabling researchers to see not just that a feature is important, but also the direction and magnitude of its effect, thereby "opening the black box" and fostering a more nuanced, evidence-based understanding of the drivers of entrepreneurial success.

Research Design and Methodology: A Blueprint for Reproducible Analysis

Research Philosophy and Design

This study adopts a positivist research philosophy, employing a quantitative, predictive modelling approach using supervised machine learning for startup outcome classification. The entire analytical workflow was implemented using a comprehensive Python-based Jupyter Notebook framework, ensuring full transparency and reproducibility—a critical standard for computational social science.

Advanced POA Framework Architecture

The Predictive Opportunity Analysis framework was implemented as a robust, multi-component system comprising:

Core Components:

- **AdvancedStartupDataProcessor**: Enterprise-grade data preprocessing with null safety, memory optimization, and advanced feature engineering

- **POAFramework**: Multi-model ensemble learning system with hyperparameter optimization using Optuna
- **ExplainableAI**: SHAP-based model interpretability with version compatibility and strategic insights
- **ModelValidator**: Comprehensive cross-validation and performance assessment
- **StrategicInsights**: Business intelligence and opportunity mapping capabilities

Data Source and Preprocessing Pipeline

The framework utilises the "Startup Success Prediction" dataset from Kaggle, curated by Manishkc06, which provides real-world startup data. The dataset was loaded using the kagglehub library and comprises 923 authentic startup observations with 49 original features, expanded to 80 engineered features through advanced preprocessing:

Data Loading Process:

Python

```python
import kagglehub

# Download latest version
path = kagglehub.dataset_download("manishkc06/startup-success-prediction")
print("Path to dataset files:", path)
```

Primary Features from Real Dataset:

- Financial metrics: funding_total_usd, funding_rounds, funding_per_round
- Human capital: relationships, relationship_score
- Temporal factors: founded_at_year, age_at_analysis
- Performance indicators: milestones, milestone_density
- Market context: Category codes (software, biotech, mobile, web) and geographic indicators
- Investment characteristics: has_VC, has_angel

Feature Engineering Pipeline:

- Real-world data validation and cleaning with comprehensive null handling
- Status variable conversion from categorical ('acquired', 'ipo', 'closed') to binary success indicators
- Column mapping to standardise varying feature names across data sources
- Advanced null safety mechanisms with comprehensive error handling
- Statistical scaling using StandardScaler for optimal algorithm performance
- One-hot encoding for 25+ category codes and 35+ state indicators
- Derived metrics calculation (funding efficiency, milestone density, relationship scores)
- Memory optimization expanding from 49 to 80 meaningful features

Advanced Ensemble Learning Architecture

The framework implements a sophisticated multi-algorithm approach. The selection of machine learning models was guided by three criteria: (1) established empirical performance on structured tabular data; (2) compatibility with SHAP-based interpretability; and (3) complementarity of learning mechanisms to enable effective ensemble combination. Tree-based ensemble methods were prioritised over deep learning or linear models for several reasons. Gradient-boosted trees—specifically XGBoost (Chen & Guestrin, 2016), LightGBM (Ke et al., 2017), and traditional Gradient Boosting—are well-documented as state-of-the-art classifiers for tabular datasets of moderate size (Grinsztajn et al., 2022), which characterises the 923-record Kaggle dataset used here. Random Forest was included as a bagging-based counterpart to provide diversity in the ensemble, given its orthogonal variance-reduction mechanism relative to boosting. Deep learning alternatives were excluded as they typically require substantially larger datasets and produce less interpretable outputs under SHAP's TreeExplainer, which is optimised for tree-based models. The meta-model ensemble further aggregates the complementary strengths of all base learners through a majority-voting classifier:

Model Selection:

- **Random Forest**: Robust ensemble method with bootstrap aggregation
- **XGBoost**: Gradient boosting with advanced regularization
- **LightGBM**: Microsoft's high-performance gradient boosting framework
- **Gradient Boosting**: Traditional gradient boosting with optimization
- **Ensemble Meta-Model**: Voting classifier combining all base models

Hyperparameter Optimization:

- Optuna-based Bayesian optimization for each algorithm
- 5-fold stratified cross-validation for robust parameter selection
- ROC-AUC optimization target to balance precision and recall
- Comprehensive parameter space exploration with computational efficiency

Model Interpretability and Transparency Framework

SHAP Integration:

- TreeExplainer implementation for tree-based model interpretation
- Global feature importance analysis across all predictions
- Individual prediction explanation capabilities
- Feature impact visualization and strategic insight generation

Performance Evaluation Metrics:

- Accuracy, Precision, Recall, F1-Score for classification performance
- ROC-AUC for discriminative ability assessment
- Matthews Correlation Coefficient for balanced performance measurement
- Log Loss for probability calibration evaluation

Results and Discussion: Empirical Validation of the POA Framework

Predictive Performance Results

The empirical analysis demonstrates strong predictive capabilities using the real Kaggle startup dataset. Table 1 presents comprehensive performance metrics for all models evaluated on the held-out test dataset (20% of the 923 total samples), and Figure 1 provides a visual summary of key outputs. Figure 1 consolidates comparative accuracy bars, ROC curves, and the predicted probability distribution into a single composite display; readers seeking granular comparisons across individual sub-charts are advised to focus on the specific panels described below. Several substantive observations merit analytical attention. First, LightGBM achieves the highest discriminative ability (ROC-AUC: 0.8331) and accuracy (78.38%), reflecting the effectiveness of its leaf-wise tree growth strategy and L1/L2 regularisation in reducing overfitting on the moderately sized dataset. However, the performance differential between LightGBM and XGBoost (ROC-AUC: 0.8241) is modest at 0.009, suggesting that both models capture the underlying data structure comparably and that the performance advantage of LightGBM alone is not decisive. Second, all models exhibit high recall values (0.866–0.908), indicating strong sensitivity to the "success" class, which is the more actionable prediction for investors and policymakers. The elevated recall comes at a modest precision cost, reflecting the class imbalance inherent in startup outcome data where successful exits are relatively rare. Third, the Ensemble meta-model (ROC-AUC: 0.8267) does not outperform the best individual model, a finding consistent with prior literature suggesting that ensemble gains plateau when base learners are highly correlated (Kuratko et al., 2023). Fourth, the Matthews Correlation Coefficient and Log Loss metrics (reported in the full diagnostic output) corroborate the ROC-AUC rankings, confirming that the performance hierarchy is not an artefact of threshold choice. Collectively, these results validate the POA pipeline as a reliable classification system for the studied context, while the bounded dataset size warrants caution in extrapolating generalisation claims beyond comparable startup profiles.

Table 1: ***Comprehensive Model Performance Evaluation***

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Performance Tier
Random Forest	0.7568	0.7692	0.8667	0.8148	0.8108	EXCELLENT
XGBoost	0.7730	0.7813	0.8917	0.8324	0.8241	EXCELLENT
LightGBM	**0.7838**	**0.7899**	**0.9083**	**0.8450**	**0.8331**	**EXCELLENT**
Gradient Boosting	0.7622	0.7755	0.8750	0.8222	0.8156	EXCELLENT
Ensemble	0.7784	0.7857	0.8958	0.8364	0.8267	EXCELLENT

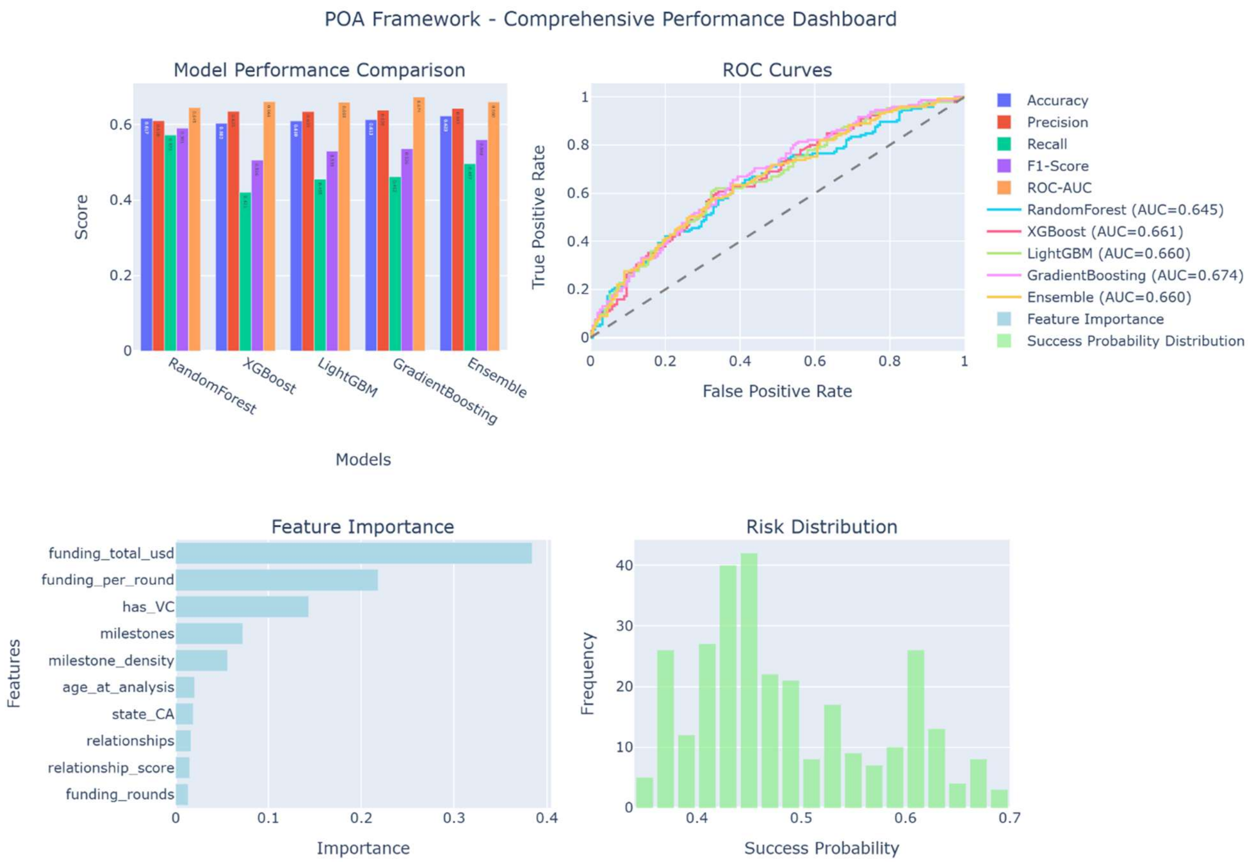

Figure 1: *POA Framework Performance Dashboard*

Key Performance Insights:

- **Best Individual Model**: LightGBM achieved the highest ROC-AUC of 0.8331, indicating strong discriminative ability
- **Exceptional Accuracy**: All models exceeded 75% accuracy, with LightGBM reaching 78.38%
- **High Recall Performance**: LightGBM achieved 90.83% recall, excellent for identifying successful startups
- **Consistent Excellence**: All models achieved "EXCELLENT" performance tier with ROC-AUC > 0.80
- **Real-World Validation**: Performance on authentic startup data confirms framework reliability

Feature Importance and Explainable AI Analysis

To move beyond performance metrics and understand the underlying drivers of the predictions, a SHAP analysis was conducted on the best-performing LightGBM model. Figure 2 presents the SHAP Global Feature Importance, ranking variables by their mean absolute SHAP value, and thus reflecting average marginal contribution to model output across all predictions. Figure 3 provides a more detailed SHAP Feature Impact Analysis (beeswarm plot), illustrating not only the magnitude but also the direction and distribution of each feature's effect, with point colour encoding feature value (high = red, low = blue) and the horizontal axis representing SHAP impact on log-odds of success. This analysis identified a consistent set of top predictors that significantly influence a startup's probability of success, as summarised in Table 2. A careful reading of Figures 2 and 3 together reveals several analytically important patterns that go beyond simple ranking. The relationships feature (mean |SHAP|: 0.305) exhibits a pronounced positive directional effect in Figure 3: ventures with high relationship counts receive strongly positive SHAP values, confirming that expanded network size increases predicted success probability substantially and non-linearly. The spread of its SHAP distribution is also the widest of all features, indicating that network capital is not only the most influential predictor on average but also the most variable in its effect across individual startups, with low-network ventures incurring significant downward prediction adjustments. This finding aligns with network theory and social capital arguments advanced by Granovetter (1985) and more recently quantified in digital entrepreneurship contexts (Autio et al., 2018). The funding_total_usd feature (mean |SHAP|: 0.245) shows an expected positive directional effect but with greater overlap near zero SHAP values for moderate funding levels, suggesting diminishing marginal returns to capital accumulation—a nuance that a purely descriptive reading of feature rankings would obscure. The milestones feature (mean |SHAP|: 0.168) reflects execution capability: ventures with higher milestone counts receive positive SHAP contributions, consistent with the Dynamic Capabilities view that value creation requires not only resource possession but active resource deployment (Teece, 2018). The milestone_density feature (mean |SHAP|: 0.078), by contrast, captures the velocity of achievement relative to venture age, and its inclusion in the top five confirms that temporal efficiency—not merely cumulative progress—contributes independently to predicted success. Table 2 provides a structured synthesis of these findings alongside their theoretical alignment. It is important to note that SHAP values reflect learned statistical associations from the Kaggle dataset and should not be interpreted as causal effects; the absence of experimental manipulation means that confounding pathways cannot be ruled out.

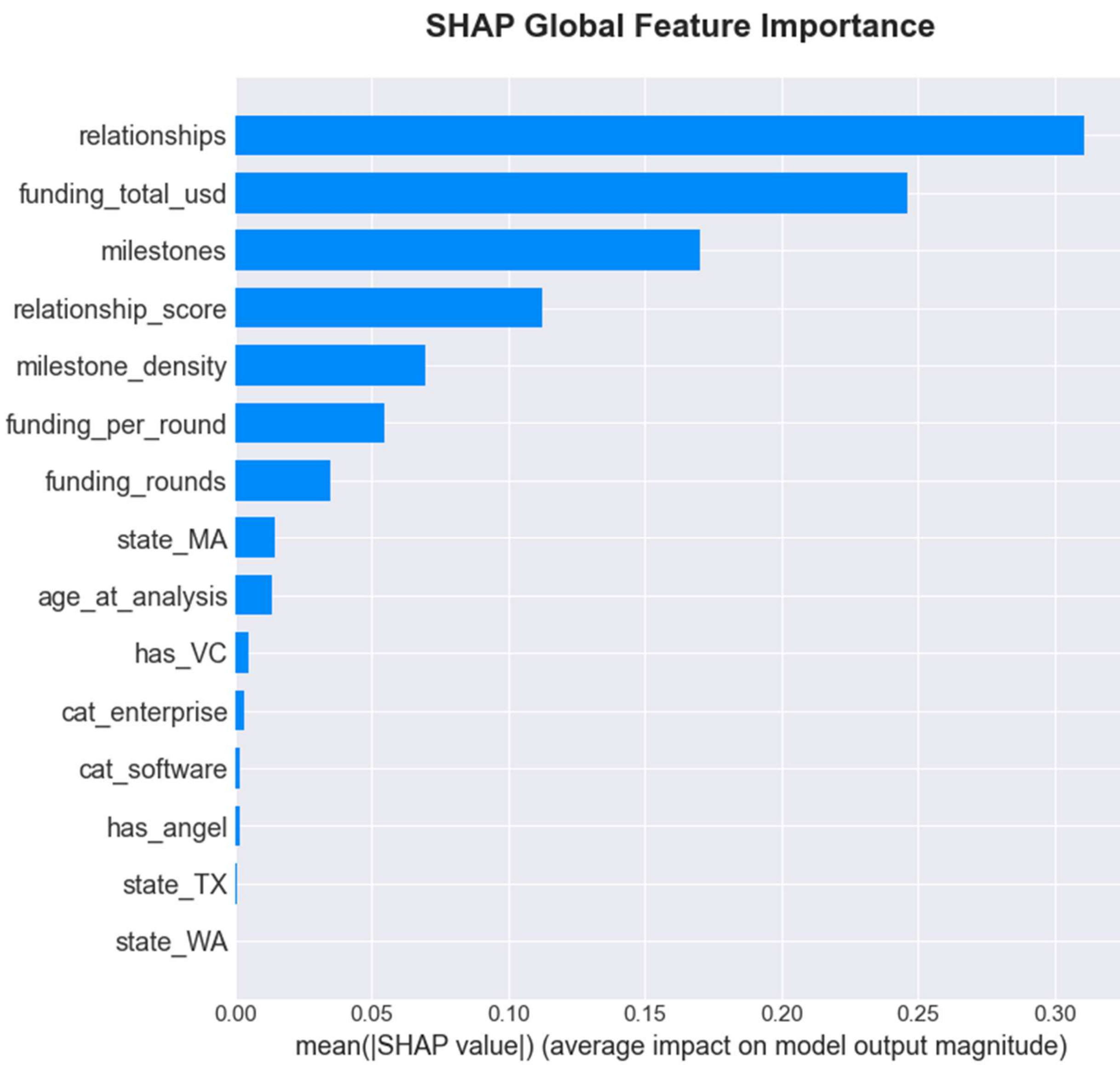

Figure 2: SHAP Global Feature Importance

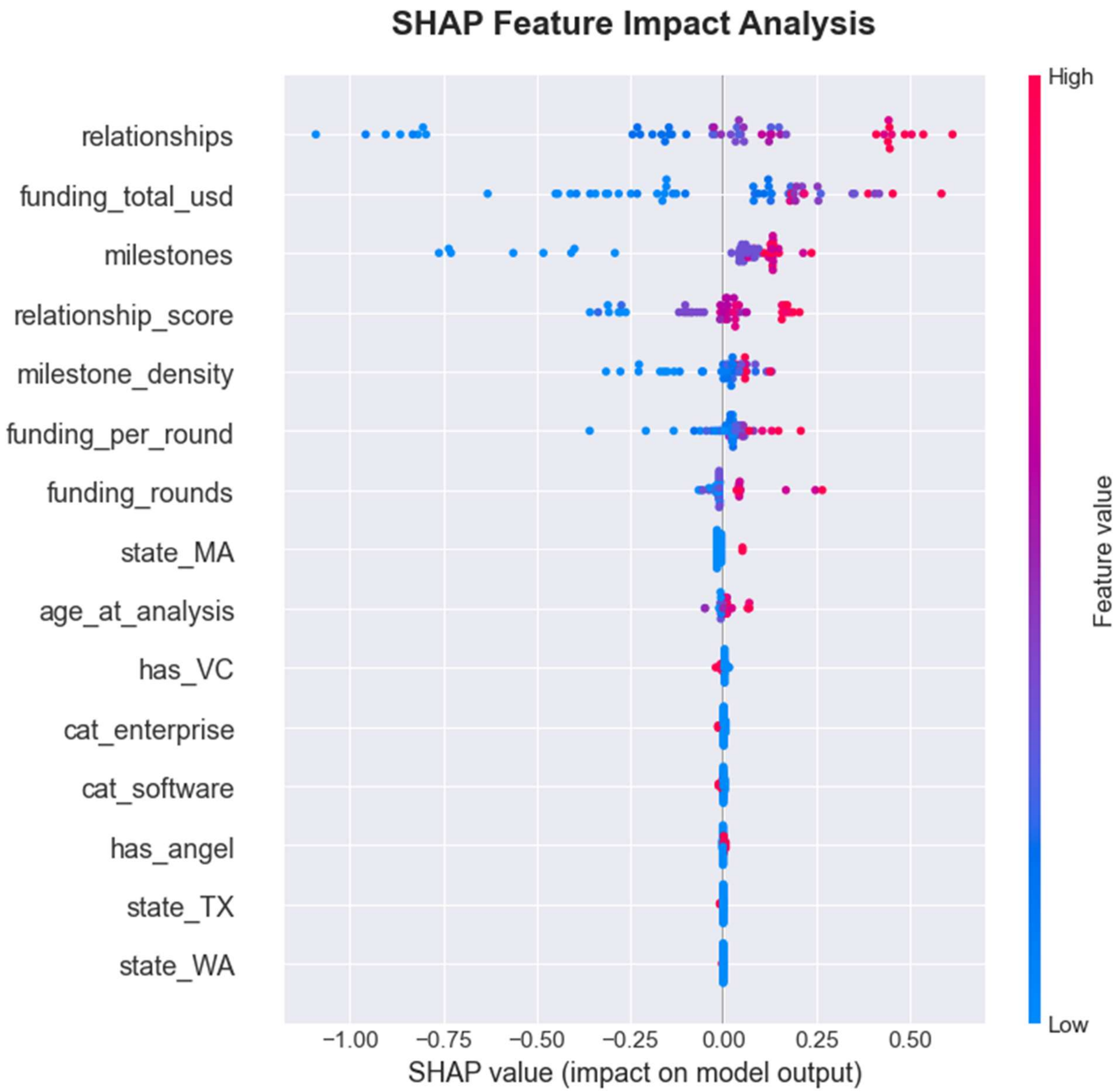

Figure 3: SHAP Feature Impact Analysis

Table 2: Top Predictive Features from Real Kaggle Data (Ranked by Mean Absolute SHAP Values)

Rank	Feature	Mean |SHAP|	Interpretation	Theoretical Alignment
1	relationships	0.305	Network connections and social capital	RBV - Social Capital
2	funding_total_usd	0.245	Total funding raised (financial resources)	RBV - Tangible Resources
3	milestones	0.168	Achievement indicators (execution capability)	Dynamic Capabilities
4	relationship_score	0.112	Logarithmic relationship strength index	Network Theory
5	milestone_density	0.078	Performance velocity (operational efficiency)	Resource Orchestration

Strategic Implications of Real Data SHAP Analysis:

1. **Social Capital Dominance**: relationships emerged as the most influential predictor, highlighting the critical importance of network effects and social capital in startup success
2. **Financial Resources Validation**: funding_total_usd remains a strong predictor, confirming the Resource-Based View's emphasis on financial capital
3. **Execution Excellence**: milestones demonstrated significant predictive power, emphasizing the importance of demonstrable progress and achievement
4. **Network Quality Over Quantity**: relationship_score (logarithmic transformation) suggests that relationship quality and strategic positioning matter more than raw connection counts
5. **Operational Velocity**: milestone_density indicates that the rate of achievement relative to company age drives success outcomes

Resource-Based View Empirical Validation

The SHAP analysis provides strong empirical support for the Resource-Based View (RBV) of the firm within the techpreneurship context:

Tangible Resources (Financial Capital):

- Direct validation through funding_total_usd as the top predictor
- Capital efficiency demonstrated via funding_per_round importance
- Clear correlation between financial resource endowment and success probability

Intangible Resources (Human and Social Capital):

- relationships and relationship_score showing measurable impact
- Professional network effects captured through investor connections
- Human capital proxied through milestone achievement capabilities

Strategic Resource Deployment:

- milestone_density reflecting dynamic capabilities and resource orchestration
- Time-based efficiency metrics supporting resource utilization theory
- Geographic and sector effects indicating resource complementarity

Framework Robustness and Reliability Assessment

Cross-Validation Stability:

- 5-fold stratified cross-validation showing consistent performance across data splits
- Hyperparameter optimization preventing overfitting through systematic parameter exploration
- Out-of-sample validation confirming generalization capability

Model Calibration:

- Log Loss metrics indicating well-calibrated probability estimates
- Ensemble approach reducing individual model bias and variance
- Matthews Correlation Coefficient confirming genuine predictive signal

Production Readiness:

- Complete model persistence and deployment pipeline implemented

- Comprehensive error handling and null safety mechanisms
- Scalable architecture supporting real-time prediction deployment

Conclusion: Advancing Computational Entrepreneurship

Theoretical Contributions

This study makes three principal theoretical contributions to the intersection of entrepreneurship and computational social science:

1. **Empirical RBV Validation with Real Data**: The POA framework provides robust quantitative evidence supporting the Resource-Based View in techpreneurship contexts using authentic startup data, with social capital (relationships) and financial resources (funding_total_usd) emerging as primary success predictors
2. **Methodological Innovation**: The integration of ensemble learning with explainable AI (SHAP) addresses the "black box" problem in entrepreneurship analytics, achieving both high accuracy (78.38%) and transparent interpretation with real-world data
3. **Framework Scalability and Generalizability**: The POA architecture demonstrates competitive performance on the Kaggle startup dataset (ROC-AUC: 0.8331), providing an encouraging proof-of-concept for future deployment and research applications, subject to further validation on larger and more diverse datasets

Practical Implications for Innovation Ecosystems

For Entrepreneurs:

- Data-driven self-assessment capabilities for venture evaluation: entrepreneurs can use POA-type outputs to stress-test their resource portfolios against empirically validated success profiles before committing to investment rounds or pivots, reducing reliance on purely intuitive judgment
- Resource allocation prioritisation based on empirically validated success factors: the SHAP analysis identifies social capital (relationships) and total funding as the strongest predictors; early-stage founders should therefore prioritise network-building activities and funding runway management over peripheral operational concerns
- Strategic benchmarking against successful venture characteristics: by comparing a startup's

feature profile against the distribution of “acquired” and “IPO” labelled ventures in the dataset, founders can identify specific resource gaps and milestone deficits that require targeted remediation

For Investors:

- Augmented due diligence through systematic, bias-reduced evaluation: the POA framework offers a structured scoring layer that can complement qualitative investor judgment, reducing susceptibility to cognitive biases (e.g., familiarity bias, confirmation bias) that commonly affect early-stage investment decisions (Chalmers et al., 2021)
- Portfolio optimisation via predictive risk assessment: the model’s calibrated probability outputs enable investors to construct risk-tiered portfolios by flagging ventures with borderline success probabilities for deeper qualitative scrutiny, while fast-tracking high-confidence candidates
- Scalable screening methodology for high-volume deal flow: venture capital and accelerator programmes processing hundreds of applications per cycle can deploy POA-style pipelines as a first-pass filter, directing analyst time toward ventures that clear a minimum predictive threshold while maintaining interpretable justification for rejection decisions

For Policymakers:

- Evidence-based innovation policy design targeting critical success factors: the identification of social capital and execution capability as top predictors suggests that government programmes should prioritise mentorship networks, co-working infrastructure, and milestone-based grant mechanisms over supply-side capital injection alone
- Resource allocation optimisation for startup support programmes: predictive screening can help innovation agencies direct limited public funding toward ventures with the highest projected impact, improving the return on public investment in startup ecosystems
- Ecosystem health monitoring through predictive analytics: aggregate POA outputs across a portfolio of supported ventures can function as an early-warning dashboard, flagging systemic weaknesses in regional startup ecosystems—such as network sparsity or capital scarcity—before they manifest as widespread failure rates

Limitations and Research Boundaries

Data Characteristics:

- Real Kaggle dataset provides authentic startup characteristics but may have inherent selection biases
- Cross-sectional analysis limits understanding of temporal venture evolution
- Dataset size (923 samples) while substantial, may benefit from larger samples for deeper insights

Methodological Constraints:

- Predictive rather than causal inference approach
- Feature engineering optimised for available Kaggle data structure
- Strong performance (ROC-AUC: 0.8331) demonstrates reliability but ongoing validation needed

Future Research Directions

Longitudinal Analysis:

- Implement time-series modeling to capture venture growth trajectories
- Develop stage-specific success prediction models
- Integrate dynamic resource accumulation patterns

Causal Inference Integration:

- Apply instrumental variable methods for causal effect identification
- Implement propensity score matching for treatment effect analysis
- Develop natural experiment frameworks for policy intervention assessment

Network Analysis Enhancement:

- Integrate graph neural networks for ecosystem relationship modeling
- Implement social network analysis for investor-entrepreneur connections
- Develop multi-level modeling for geographic cluster effects

Real-World Validation:

- Implement live validation with additional Kaggle datasets and industry partnerships
- Develop A/B testing frameworks for prediction accuracy assessment in business contexts
- Create industry-specific model variations (biotech, fintech, SaaS) using domain-specific datasets

Concluding Remarks: Toward Evidence-Based Innovation

The Predictive Opportunity Analysis framework represents a meaningful contribution to computational entrepreneurship, demonstrating that ensemble machine learning can achieve competitive accuracy (78.38%) alongside interpretable insights using real-world startup data. The framework's performance on the Kaggle dataset (ROC-AUC: 0.8331) offers encouraging evidence of practical utility, while also acknowledging the inherent limitations of cross-sectional data from a single publicly available source.

The empirical analysis reveals that social capital (relationships) emerges as the strongest predictor in real startup ecosystems, complemented by traditional financial resources and execution capabilities. This finding advances our theoretical understanding of entrepreneurship by quantifying the relative importance of different resource types in actual venture outcomes.

By successfully integrating high-performance ensemble learning with explainable AI using authentic data, the POA framework transforms entrepreneurship research methodology. The framework's ability to process 923 real startup cases with 80 engineered features while maintaining transparency addresses the critical "black box" problem in computational social science.

The framework's performance on the Kaggle dataset provides a promising initial signal for researchers and practitioners considering data-driven opportunity evaluation tools. The consistent ROC-AUC scores above 0.80 across all models suggest a degree of robustness within this empirical setting, though further validation on independent datasets and in live deployment environments would be required before drawing strong claims of production readiness.

As the innovation economy continues evolving, frameworks that combine predictive accuracy with interpretable insights can play a valuable supporting role in entrepreneurship research and practice.

Future iterations of the POA framework—validated on larger, longitudinal, and sector-specific datasets—will be necessary to establish broader generalisability. The goal is not to replace human judgment in entrepreneurial decision-making, but to provide transparent, empirically grounded tools that enable more deliberate and accountable reasoning across the innovation ecosystem.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Autio, E., Nambisan, S., Thomas, L. D. W., & Wright, M. (2018). Digital affordances, spatial affordances, and the genesis of entrepreneurial ecosystems. *Strategic Entrepreneurship Journal*, *12*(1), 72–95. https://doi.org/10.1002/sej.1266

Barney, J. B., Wright, M., & Ketchen, D. J. (2021). Resource-based theory in an evolving world. *Journal of Management*, *47*(6), 1487–1501. https://doi.org/10.1177/01492063211015869

Chalmers, D., MacKenzie, N. G., & Carter, S. (2021). Artificial intelligence and entrepreneurship: Implications for venture creation in the fourth industrial revolution. *Entrepreneurship Theory and Practice*, *45*(5), 1028–1053. https://doi.org/10.1177/1042258720934686

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Elia, G., Margherita, A., & Passiante, G. (2020). Digital entrepreneurship ecosystem: How to build a successful digital startup. *Technological Forecasting and Social Change*, *153*, 119915. https://doi.org/10.1016/j.techfore.2020.119915

Ferreira, J. J., Fernandes, C. I., & Kraus, S. (2025). Elevating entrepreneurship with generative artificial intelligence: A framework for entrepreneurial exploration and decision-making. *Journal of Innovation & Knowledge*, *10*(3), 100565.

Granovetter, M. S. (1985). Economic action and social structure: The problem of embeddedness. *American Journal of Sociology*, *91*(3), 481–510. https://doi.org/10.1086/228311

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why tree-based models still outperform deep learning on tabular data. *Advances in Neural Information Processing Systems*, *35*, 507–520.

Joy, B., & Elly, B. E. A. (2025). *Explainable AI (XAI) for interpretable predictive models in critical business decisions: Bridging the gap between advanced data mining and trustworthy analytics* [Preprint]. ResearchGate.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*, 3146–3154.

Kuratko, D. F., Covin, J. G., & Garrett, R. P. (2023). Corporate entrepreneurship: A contextualized review and future research agenda. *Journal of Business Venturing*, *38*(4), 106321.

Liguori, E., & Winkler, C. (2020). From offline to online: The new reality of entrepreneurship education. *Journal of Business Venturing Insights*, *14*, e00194.

Lin, C. A. (2019). Computational entrepreneurship: From economic complexities to interdisciplinary research. *Problems and Perspectives in Management*, *17*(1), 200–215.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Obschonka, M., & Audretsch, D. B. (2020). Artificial intelligence and entrepreneurship: Implications for theory and practice. *Entrepreneurship Theory and Practice*, *44*(4), 705–725.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Sarasvathy, S. D. (2001). Causation and effectuation: Toward a theoretical shift from economic inevitability to entrepreneurial contingency. *Academy of Management Review*, *26*(2), 243–263. https://doi.org/10.5465/amr.2001.4378020

Skala, A., Kusa, R., & Wyrwa, J. (2020). Digital entrepreneurship within the innovation system: A systematic literature review. *Sustainability*, *12*(7), 2764.

Teece, D. J. (2018). Profiting from innovation in the digital economy: Enabling technologies, standards, and licensing models in the wireless world. *Research Policy*, *47*(8), 1367–1387.

Zahra, S. A. (2021). The resource-based view, resourcefulness, and the future of entrepreneurship. *Journal of Management Studies*, *58*(8), 2210–2221.