

Quantifying Founder-Market Fit: A Machine Learning Approach to Startup Success Prediction*

Ekaterina Gonchar; Sebastian Diaz; Benjamin Schmidt; Priyanshu Yadav; Qiwei Han
Nova School of Business and Economics, Carcavelos, Portugal
59573@novasbe.pt, 59369@novasbe.pt, 59599@novasbe.pt
59061@novasbe.pt, qiwei.han@novasbe.pt

Abstract

The high failure rate of early-stage startups poses persistent challenges for venture capitalists and innovation policymakers alike. Although Founder-Market Fit (FMF), defined as the alignment between a founder's background and the domain of their startup, has rarely been systematically quantified, it is widely acknowledged in practice as a key determinant of success. In this paper, we present a novel, data-driven framework to operationalize and predict FMF using machine learning and natural language processing. We construct high-dimensional representations of founder profiles by aggregating structured data from Crunchbase, LinkedIn, and X, and apply transformer-based embeddings to quantify semantic alignment with industry verticals. FMF scores, together with features related to prestige, experience, seniority, and inferred personality traits, are incorporated into supervised models to predict startup success. Our findings show that FMF significantly improves predictive performance over baseline models and remains robust across weighting schemes and learning algorithms. By providing a scalable, interpretable, and auditable approach to founder evaluation, this study advances algorithmic entrepreneurship and offers practical insights for investors, accelerators, and policymakers seeking to improve early-stage startup assessments.

Keywords: Founder-Market Fit, Startup Evaluation, Venture Capital, Machine Learning, Natural Language Processing

This work was funded by Fundação para a Ciência e a Tecnologia (UIDB/00124/2020, UIDP/00124/2020 and Social Sciences DataLab - PINFRA/22209/2016), POR Lisboa and POR Norte (Social Sciences DataLab, PINFRA/22209/2016).

1. Introduction

Startups are central to innovation-driven economies. They catalyze innovation, disrupt industries, and generate high-skilled jobs (Choi et al., 2020). Yet despite their economic importance, failure rates remain alarmingly high: roughly 90% of ventures close within five years (Gage, 2012; Patel, 2015). This attrition introduces considerable uncertainty for stakeholders, particularly early-stage investors who must allocate capital under conditions of sparse data and high information asymmetry (Cohen & Dean, 2005).

This unpredictability has fueled interest in predictive analytics and decision support systems for venture evaluation (Arroyo et al., 2019; Du et al., 2024; Thirupathi et al., 2021). Traditional venture capital (VC) assessments rely on qualitative heuristics such as founder charisma, elite affiliations, and endorsements from influential networks (Lerner, 2009; Zider, 1998). While widespread, these subjective methods lack scalability, reduce objectivity, and often introduce systematic biases—particularly in early-stage contexts where financial and product data are limited.

To address these challenges, increasing attention has turned to *founder-market fit* (FMF), long valued by investors but seldom formalized. FMF captures the alignment between a founder's background, including education, industry experience, and psychological disposition, and the venture's market domain (Hunter et al., 2017). Intuitively, founders with deep domain understanding are better positioned to anticipate customer needs, navigate operational complexity, and adapt to market dynamics (Dencker & Gruber, 2015; Hashai & Zahra, 2022). Yet despite its prominence in practice, FMF remains anecdotal and lacks scalable, objective quantification within formal decision-making frameworks.

Building on this gap, we ask: *Can FMF be systematically quantified using machine learning and natural language processing, and does such quantification improve the prediction of startup success compared to traditional heuristics?* Framing the research question this way connects the practical problem of subjective founder evaluation with a data-driven solution. Our framework leverages modern computational methods to operationalize FMF by embedding and comparing founders' education, experience, and psychological traits against their startups' market domains. We construct high-dimensional representations of founder profiles by aggregating heterogeneous data from Crunchbase, LinkedIn, and X (formerly Twitter), and compute FMF scores based on semantic similarity between founder and industry embeddings using transformer-based NLP models (Reimers & Gurevych, 2019). The FMF score integrates educational alignment, professional experience, and job title relevance, with weights calibrated by predictive performance.

Prior research supports FMF's predictive relevance. Founders with domain-relevant experience are more likely to secure funding, gain traction, and achieve exits (Dencker & Gruber, 2015; Hashai & Zahra, 2022). These findings align with human capital theory (Becker, 1965), which emphasizes specialized knowledge, and the resource-based view of the firm (Barney, 1991), which highlights the strategic value of founder capabilities. For example, Dencker and Gruber (2015) show that prior industry experience reduces liability of newness and accelerates learning. Linguistic traits such as openness and emotional stability also correlate with resilience and team effectiveness (Freiberg & Matz, 2023).

We therefore extend FMF modeling with covariates including institutional prestige (university and employer reputation (Gompers & Lerner, 2020; Roberts, Eesley, et al., 2011)), job seniority, cumulative work experience (Arroyo et al., 2019), and inferred personality traits from social media (Freiberg & Matz, 2023). These structured and unstructured features are combined in supervised models, including gradient-boosted trees and feedforward neural networks, to predict startup outcomes: IPO or acquisition versus failure (McCarthy et al., 2023).

Our contributions are twofold. First, we develop and validate a scalable framework to quantify FMF, grounded in theory and enabled by advances in NLP and representation learning. Second, we show empirically that FMF and related features such as prestige, seniority, and personality significantly improve predictions of startup success, outperforming baselines that rely solely

on funding data or institutional affiliation.

The broader implications extend beyond technical accuracy. Current VC practices often overemphasize prestige-based heuristics, contributing to underinvestment in nontraditional founders (Gompers & Kovvali, 2018; Jeong, 2022). High-profile failures such as Theranos and WeWork illustrate the cost of insufficient rigor in founder evaluation (Carreyrou, 2018; Stein, 2023). By reconceptualizing FMF as a quantifiable, data-driven metric, this study contributes to responsible AI and equitable innovation finance, offering a predictive, interpretable, and auditable framework for early-stage investment.

2. Related Work

2.1. Startup Success Prediction and Financial Signals

Startup outcomes are often operationalized as one of three discrete exit events, including *IPO*, *acquisition*, or *failure* (Chen et al., 2009; Kaplan & Lerner, 2016). Although these results are objective and easy to codify, they are rare and laggard indicators that occur typically years after the inception of a firm. This temporal delay complicates the task of early-stage prediction. To address this limitation, researchers have introduced alternative proxy outcomes such as follow-up financing (Davila et al., 2003), survival duration (Shepherd et al., 2000), and growth milestones (Cumming & Johan, 2009). However, forecasting long-horizon exits for pre-revenue startups, particularly those in seed or early-stage rounds, remains a central methodological challenge.

Traditional research on startup evaluation has prioritized firm-level signals. These include funding stages, investor reputation, product-market characteristics, and team composition (Chen et al., 2009; Gompers & Lerner, 2020; Hellmann & Puri, 2000; Kaplan & Lerner, 2016). Round size, investor quality, and capital structure are widely used as proxies for growth potential and exit likelihood (Cumming & Johan, 2009; Davila et al., 2003). More recently, machine learning methods have been adopted to model complex patterns associated with startup outcomes (Lyonnet & Stern, 2024; McCarthy et al., 2023). However, such models are mainly focused on financial and operational indicators and often underrepresent the role of founder-specific attributes (Yang et al., 2023).

2.2. Founder Characteristics and Human Capital

A growing body of literature emphasizes the critical role of founders in determining venture outcomes. Human capital theory posits that domain-relevant knowledge, skills, and experience enhance entrepreneurial performance (Becker, 1965). Empirical studies consistently find that prior industry experience, managerial seniority, and educational attainment positively correlate with a startup's probability of success (Colombo & Grilli, 2005; Razaghzadeh Bidgoli et al., 2024; Roberts, Eesley, et al., 2011). In addition, the professional networks of founders and their career trajectories influence their access to capital, mentorship, and partnership opportunities (Gompers & Lerner, 2020; Lerner, 2009).

Despite these findings, many studies are based on relatively coarse measures of human capital (Colombo & Grilli, 2005), such as degree type, years of experience, or broad industry categories (Dalziel & Basir, 2024; Roche et al., 2020). These proxies do not fully capture the nuanced alignment between the professional history of a founder and the specific demands of the startup market. Few models explicitly formalize this alignment, leaving the concept of FMF theoretically recognized but practically underexplored.

2.3. Founder Personality and Social Media Signals

Beyond formal credentials and work history, founder personality has emerged as a significant predictor of startup performance. The availability of digital trace data, particularly from social networks, has enabled researchers to infer psychological attributes that are otherwise unobservable in traditional datasets. Freiberg and Matz (2023) used transformer-based models to infer Big Five personality traits from X data, linking traits such as openness and conscientiousness to a higher startup valuation. Similarly, McCarthy et al. (2023) demonstrated that linguistic markers of openness and emotional stability, extracted from founders' online communications, are positively associated with acquisition or IPO outcomes. Razaghzadeh Bidgoli et al. (2024) applied machine learning models to predict startup success using a mix of social media and business indicators, and found that features such as the number of LinkedIn and Twitter followers had strong predictive contribution.

These approaches illustrate the potential of using natural language processing and social signals to extract latent founder qualities at scale. Our study builds on this work by incorporating personality features inferred

from founders' social media activities on the X platform into predictive models. This allows us to account for intangible but impactful personal characteristics in a data-driven and scalable manner.

3. Data Collection and Processing

This section describes the comprehensive, multi-source data pipeline developed to construct a robust dataset of startup ventures and their founders. Drawing on structured and unstructured information from Crunchbase, LinkedIn, and X, this pipeline enables the integration of financial metrics, founder attributes, and digital behavior to support predictive modeling of startup success and FMF. Figure 1 illustrates the data architecture.

3.1. Startup Records and Financial Features

We worked with YunoAI, an Italian startup analytics firm that provides a base dataset containing 469,579 startup records in JSON format, including the founding year, country, sector tags, and web presence. This dataset was enriched by merging structured data from Crunchbase's general information, acquisitions, and funding modules. The resulting dataset included operating status, acquisition history, and detailed funding rounds (type, amount, date). Startups with at least one round of funding were retained to ensure baseline maturity, yielding a filtered set of 109,290 startups.

To characterize startup financing, we engineered features such as cumulative funding, funding frequency, and funding stage progression (e.g., seed to Series B). Company descriptions were embedded using the BERT model and mapped to 25 standard industry categories based on cosine similarity with Global Industry Classification Standard (GICS) vectors.

3.2. Stratified Sampling and Geographic Focus

A representative subset of 10,000 startups was selected to facilitate scalable model training. This stratified sample was balanced by funding stage, industry type and geography, focusing on English-speaking Western countries (US, UK, Canada, Australia). Only startups founded between 2010 and 2020 were retained to ensure comparability in both funding cycle maturity and linguistic consistency in founder communications.

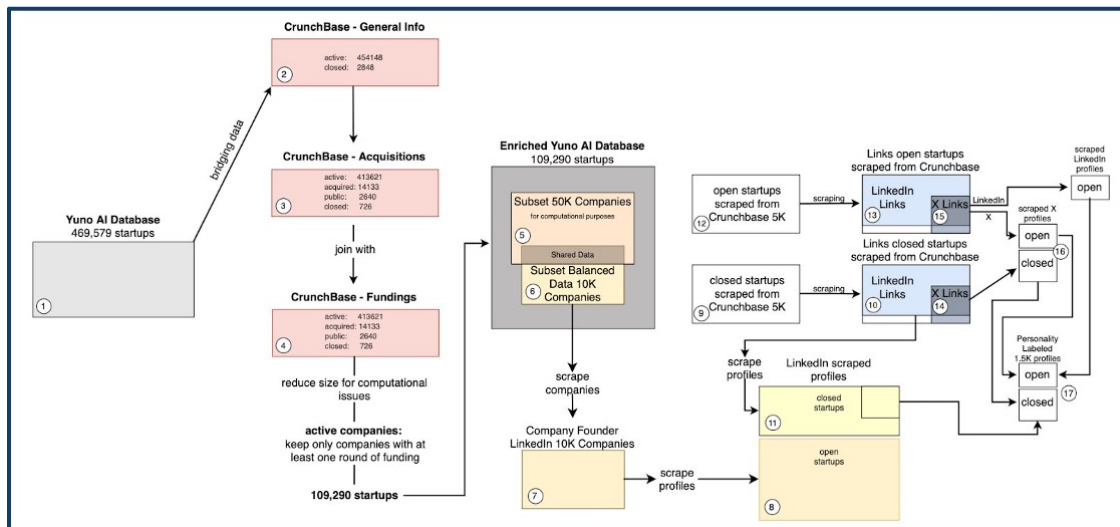


Figure 1. Data Collection and Processing Flow

3.3. Founder Identification via LinkedIn

Using LinkedIn URLs from the startup profiles, we accessed the “People” section of each company page and applied keyword heuristics (e.g., *founder*, *co-founder*, *CEO*) to identify candidate profiles. When LinkedIn links were unavailable or inactive, we implemented a fallback strategy using screenshots and Google Lens OCR with reverse image search to recover profile URLs. To ensure validity, rule-based filtering was applied to remove irrelevant or mismatched profiles, such as employees with “founder” titles at prior firms. Although not always perfectly accurate, this process substantially reduced data loss and allowed us to retain a balanced representation of founders. In total, the pipeline yielded 38,800 potential founder profiles associated with 6,811 startups. This ensured that recovered profiles were integrated with the same level of detail, such as education, professional experience, and engagement metrics, as those directly obtained from active LinkedIn links.

3.4. Founder Profile Extraction and Structuring

Each founder profile was scraped and stored in structured JSON format, including educational background (institution, degree, dates), career trajectory (positions held, durations, companies), and engagement metrics (posts, likes, comments). Engagement capture was limited to five scrolls per profile to ensure feasibility. To quantify prestige, universities were

ranked using multiple systems (QS¹, THE², ARWU³, CWUR⁴), and employers were cross-referenced with the Fortune 500⁵, Forbes Global 2000⁶ and LinkedIn Top Company List. Job seniority was inferred from title-based classification using a curated training set from LinkedIn job postings.

3.5. Industry Classification and Market Fit

The career history of each founder was assigned to industries using LinkedIn’s six-level Industry Codes V2 taxonomy⁷. These mappings enabled a semantic alignment between founders’ prior domains and the industries of their startups, which underpins the FMF score. Founders with extensive prior experience in the startup’s target industry were deemed better aligned.

3.6. Social Media and Personality Traits

Founders with public X profiles were linked via Crunchbase metadata. For those with at least 30 original posts, Big Five personality traits were inferred using a fine-tuned BERT model trained on the PANDORA dataset (Gjurković et al., 2021). The model achieved MSE = 0.0781 and MAE = 0.2308, resulting in 1,500 high-quality trait-labeled profiles.

¹<https://www.topuniversities.com/university-rankings>
²<https://www.timeshighereducation.com/world-university-rankings>
³<https://www.shanghairanking.com/rankings/arwu/2024>
⁴<https://cwur.org/>
⁵<https://fortune.com/ranking/global500/>
⁶<https://www.forbes.com/lists/global2000/>
⁷<https://learn.microsoft.com/en-us/linkedin/shared/references/reference-tables/industry-codes-v2>

This integrated data pipeline forms the empirical backbone of our predictive models, enabling joint representation of financial signals, career features, institutional prestige, and founder psychographics for startup outcome forecasting and FMF estimation.

3.7. Mathematical Formalization of FMF

We define the FMF score as a composite measure of how well a founder’s prior background aligns with the startup’s market focus. This alignment captures three core dimensions: the semantic similarity between the founder’s educational history and the market domain, the alignment of prior professional experience, and the relevance of job titles held.

Formally, the FMF score between founder i and startup j is given by:

$$\text{FMF}_{ij} = \alpha \cdot \text{sim}(f_i^{\text{edu}}, s_j) + \beta \cdot \text{sim}(f_i^{\text{exp}}, s_j) + \lambda \cdot \text{sim}(f_i^{\text{job}}, s_j) \quad (1)$$

where f_i^{edu} , f_i^{exp} , and f_i^{job} are dense vector representations of founder i ’s educational background, professional experience, and job titles, respectively, and s_j represents the startup’s industry profile, encoded using the same embedding model. α and β are tunable weights, with the job title weight implicitly defined as $\lambda = 1 - \alpha - \beta$. Our initial parameterization is grounded in prior literature: education is weighted at 30%, consistent with human capital theory and evidence linking domain-specific knowledge and institutional prestige to entrepreneurial outcomes (Dimov & Shepherd, 2005; Eesley & Wang, 2017); professional experience is weighted at 30%, reflecting resource-based and signaling perspectives that emphasize industry tenure, prestige, and network access (Hsu, 2007; Packalen & Bhattacharya, 2015); and job title relevance is emphasized (40%) based on human capital theory, which highlights the importance of leadership roles and execution capacity (Unger et al., 2011).

To account for founder credibility and behavioral traits, the extended FMF score incorporates prestige and personality:

$$\hat{\text{FMF}}_i = \gamma \cdot \text{FMF}_{ij} + \delta \cdot P_i + \theta \cdot \|\vec{p}_i\|_2 \quad (2)$$

where P_i denotes the founder’s normalized prestige score, based on university and employer rankings, and \vec{p}_i is the personality embedding vector inferred from social media posts. Unlike α and β , which have theory-informed starting values, γ , δ , and θ are tuned empirically via grid search.

3.8. NLP Feature Engineering

Unstructured founder text, including job titles, profiles, and posts from LinkedIn and X, is cleaned, tokenized, and embedded using the BERT model. To derive structured features, job roles are classified into hierarchical categories (e.g., technical, managerial, executive) with a fine-tuned BERT classifier, and past work experiences are mapped to LinkedIn’s six-level Industry Codes V2 taxonomy to compute industry alignment with startup vectors. High-dimensional embeddings are reduced via PCA to mitigate overfitting, and categorical metadata such as country and sector is represented through one-hot encoding.

3.9. Model Training and Tuning

We assess the predictive performance of FMF-based models using three supervised learning algorithms: XGBoost, Random Forest, and a two-layer feedforward neural network. Following prior literature, we operationalize startup success through discrete exit events: IPO, acquisition, or failure within five years of founding (Chen et al., 2009; Kaplan & Lerner, 2016). These events are objective, well-documented, and widely used in venture research. The feature set includes FMF scores, personality embeddings, prestige indicators, and traditional control variables such as startup age, team size, funding history, and geography.

To train and evaluate the models, we use a stratified 5-fold cross-validation scheme that preserves class balance between successful exits (IPO or acquisition) and failures. Because exits are less frequent than failures, we apply the Synthetic Minority Oversampling Technique (SMOTE) to generate additional samples of successful exits, complemented by random undersampling of the majority class. Model hyperparameters, such as learning rate, tree depth, and hidden layer size are optimized using a randomized grid search. Performance is evaluated using the F1-score and AUC, which balance precision and recall while accounting for class imbalance. Finally, to ensure interpretability, we compute feature importance values for the tree-based models to highlight the most influential predictors of startup outcomes.

4. Results

4.1. Model Performance Summary

Table 1 reports the performance of three supervised learning models trained to predict startup success: XGBoost, Random Forest, and a Multi-Layer Perceptron (MLP). Each model was trained on a

stratified, class-balanced subset of 10,000 ventures (5,000 successes, 5,000 failures) to mitigate class imbalance and facilitate fair comparison. Evaluation was performed using 5-fold cross-validation, with stratification ensuring proportional class representation across folds.

XGBoost consistently outperforms the other classifiers across all metrics, including overall accuracy, F1-score (which balances precision and recall), and the Area Under the ROC Curve (AUC), which captures the model's ability to discriminate between success and failure cases across thresholds. The model also exhibits strong balanced accuracy, indicating reliable performance across both classes in the presence of potential class distribution shifts. Additionally, its native support for feature importance make it particularly suitable for high-stakes domains like venture capital, where interpretability is essential.

While the neural network (MLP) achieves competitive AUC and generalization performance, it lacks the inherent interpretability of tree-based models. The Random Forest model, though robust and stable, trails behind both XGBoost and the MLP in predictive performance.

4.2. Influence of Prestige and Seniority

Prestige and seniority emerge as statistically relevant variables in predicting success trajectories. Founders affiliated with globally ranked universities and top-tier employers demonstrate elevated success probabilities, in line with signaling theory and network-based capital access (Gompers & Lerner, 2020; Roberts, Eesley, et al., 2011). Furthermore, job seniority is estimated using a classifier trained in LinkedIn-tagged job postings, predicting levels from Entry to Executive. This structured seniority score, along with the cumulative duration of relevant managerial experience, is positively correlated with startup performance, confirming long-standing theories of human capital accumulation (Becker, 1965).

4.3. Role of Personality Traits

Our analysis integrates psychometric features inferred from public social media text, specifically X posts. Using BERT-based transformer models fine-tuned for personality prediction, we extract Big Five trait scores for founders with sufficient social trace data. Among these, openness to experience and emotional stability (i.e., low neuroticism) show strong positive associations with startup success, likely reflecting adaptive capacity and stress resilience as critical traits for navigating early-stage

volatility (Freiberg & Matz, 2023). In contrast, traits such as agreeableness and introversion have weaker and less consistent effects.

4.4. Model Explainability via Feature Importances

To improve transparency, we compute feature importance scores from the best-performing XGBoost model, identifying the top predictors of startup success. The FMF score ranks among the most influential variables, alongside founder seniority, age, and domain-specific work experience. These global importance values highlight which attributes generally drive model predictions. While this provides a high-level view, practical deployment would also benefit from local interpretability techniques such as SHAP values, which can attribute prediction outcomes to specific founder attributes on a case-by-case basis. This would allow the framework to not only evaluate startups but also suggest actionable areas for improvement (e.g., highlighting the relative lack of industry alignment or managerial tenure). Incorporating such tools into future iterations would broaden the model's usefulness as a decision-support system for both founders and investors.

5. Model Robustness and Validation

5.1. Robustness Checks

To ensure the robustness of our findings, we conducted a series of sensitivity analyses. Specifically, we varied the hyperparameters α and β , which weight the relative importance of educational and professional components in the FMF score across a broad range from 0.1 to 0.9 in increments of 0.1. In all of these specifications, the model's F1 score fluctuated by no more than 2.4 percentage points, suggesting that the FMF embeddings retain their predictive utility in different configurations of feature weighting. This confirms that our approach to modeling the FMF is not overly sensitive to subjective parameter tuning and can generalize across different assumptions of founder-background importance.

In addition, we performed a geographic stratification analysis by segmenting the dataset into two cohorts: startups based in the United States and those based in Europe. Separate models were trained and evaluated within each regional subset. We found that both subsets achieved comparable performance (U.S.: F1 = 0.728, Europe: F1 = 0.719), indicating that our FMF-based model captures transferable founder signals that are not regionally confined. This provides empirical support for the geographic robustness of the FMF construct.

Table 1. Model Performance on Startup Success Prediction

Model	Accuracy	F1-score	AUC	Balanced Accuracy
XGBoost	0.792	0.731	0.845	0.784
Random Forest	0.765	0.703	0.814	0.763
Neural Network	0.771	0.709	0.827	0.768

5.2. External Benchmarking

To contextualize the performance of our machine learning model, we constructed a heuristic benchmark model based on a commonly used VC screening rule. This heuristic defines “high-potential” founders as those with at least one of the following: (1) educational background from an Ivy League or Top 50 global university (based on QS and THE rankings), or (2) prior work experience at a Fortune 500 company. This rule reflects conventional prestige-based filters often used in early-stage evaluation workflows (Roberts, Eesley, et al., 2011; Zider, 1998).

The heuristic model achieves an F1 score of 0.591 and an AUC of 0.682, which is significantly lower than the performance of our full model (F1 = 0.734, AUC = 0.845). This performance gap highlights two key insights. First, reliance on static, prestige-based heuristics misses meaningful information embedded in founder-career alignment and behavioral traits. Second, our FMF-based framework offers a more nuanced and data-driven approach that surpasses traditional rules of thumb, reinforcing its practical relevance for VC and accelerator selection processes.

5.3. Ablation Study

To assess the contribution of each feature group to model performance, we conduct an ablation study by systematically removing one group at a time and evaluating changes in F1 Score and AUC. Results are shown in Table 2, along with the performance drop (Δ F1) relative to the full model.

Removing FMF embeddings results in the largest performance drop, underscoring their centrality in predicting startup success. Prestige indicators also contribute meaningfully, consistent with the literature on signaling theory (Courtney et al., 2017). The personality traits inferred from the X posts offer additional predictive power, particularly when conventional metrics are not available. Founder tenure has the least impact, suggesting it adds limited value once richer alignment signals are included.

5.4. Temporal Holdout Validation

From a business perspective, one of the most critical challenges for startups lies in adapting to exogenous shocks such as technological shifts, macroeconomic crises, or behavioral changes in consumer markets (Silva et al., 2023). To test the temporal stability and generalizability of our predictive model, we implemented a holdout validation based on startup founding years. The model was trained exclusively on startups founded prior to 2018 and tested on a temporal holdout set comprising startups launched between 2018 and 2020. This setup simulates a real-world investment scenario, where decisions are made based on historical patterns applied to unseen future startups.

The model achieved a precision of 77.6% and an F1 score of 0.71 on temporal holdout, demonstrating strong generalization performance despite potential changes in market dynamics or founder behavior. This result suggests that the FMF features capture the enduring structural factors of entrepreneurial success and remain robust even in the face of recent disruptions in the startup landscape, such as the rise of remote-first startups and platform-based business models. However, we acknowledge this validation does not capture the post-COVID market fluctuations. Future research should extend temporal validation to later cohorts to assess the robustness of FMF features under pandemic-driven and subsequent structural shifts.

6. Discussion

6.1. Theoretical Implications

This study advances theoretical understanding of FMF by introducing a computational framework that operationalizes FMF at scale. While prior literature has explored FMF as a conceptual alignment between a founder’s background and venture domain (Dencker & Gruber, 2015; Hunter et al., 2017), our research offers the first integrated model that quantifies FMF using structured features such as educational prestige, industry-aligned work experience, and personality indicators inferred from public social media. This empirical formulation confirms the theoretical claim that FMF serves as a strategic

Table 2. Ablation Study: F1 Score and AUC Across Feature Groups

Feature Set	F1 Score	AUC	Δ F1 (vs. Full)
All Features (Full Model)	0.731	0.845	-
<i>w/o FMF Embeddings</i>	0.692	0.785	-0.039
<i>w/o Prestige Indicators</i>	0.708	0.801	-0.023
<i>w/o Personality Traits</i>	0.717	0.810	-0.014
<i>w/o Founder's Tenure</i>	0.724	0.816	-0.007

resource aligned with the resource-based view of entrepreneurial advantage (Barney, 1991), enabling comparative differentiation and performance gains among early-stage ventures.

6.2. Practical Implications

By operationalizing FMF and personality inference through machine learning and NLP techniques, our framework provides a scalable, modular, and interpretable augmentation to traditional due diligence. Unlike conventional heuristic-based screening, which often relies on pedigree, intuition, or informal signals, our approach enables systematic, evidence-based founder evaluation. This allows institutional investors to apply consistent criteria across diverse candidate pools, improving both efficiency and fairness in the selection process. In particular, data-driven FMF scores capture the alignment between the prior experience of founders and the domain of their startup in a quantifiable way, enabling more robust prediction even in early-stage contexts with limited financial data.

In addition, the inclusion of inferred personality traits and indicators of educational prestige expands the scope of the evaluation beyond static resumes or superficial affiliations. By combining scale with founder-level signals, the pipeline supports more inclusive investment practices and helps identify high-potential founders who might otherwise be overlooked due to non-traditional backgrounds (Gompers & Kovvali, 2018; Zider, 1998).

6.3. Limitations

While the results are promising, several limitations warrant consideration. First, our reliance on publicly available platforms such as Crunchbase, LinkedIn, and X introduces inherent biases in coverage: English-speaking founders and those active on digital professional or social networks are disproportionately represented, while founders from underrepresented groups, non-digital markets, or non-Western regions may be systematically overlooked. This sampling bias may limit external validity, and we highlight integration

of additional data sources (e.g., AngelList, regional startup databases) as a key avenue for future work to improve inclusivity and reduce dependence on a narrow set of platforms. Second, personality traits could only be inferred for founders with sufficient public activity on X (at least 30 original posts). This excludes many who lack an active social presence or restrict content access. Ablation studies indicate that models without personality features still perform robustly, suggesting these features add value but are not essential. Future research could incorporate structured assessments (e.g., MBTI) or alternative digital traces from pitch decks and accelerator applications to expand coverage.

Third, while our FMF model demonstrates strong predictive power in retrospective analysis, prospective validation in real-world VC decision settings remains an open avenue. Fourth, the present study focuses on startups that have already raised at least one round of funding, ensuring sufficient public data for modeling. This excludes ventures in ideation or incubation stages, where structured information is sparse. Extending the FMF framework to pre-investment contexts by leveraging partial founder profiles, accelerator applications, or survey-based assessments would be a valuable direction for future research. Fifth, the weighting scheme used to construct FMF remains partially heuristic despite theoretical motivation. Although robustness checks show stability across alternative specifications, future work could explore dynamic or data-driven weighting to enhance generalizability across industries, geographies, and demographics.

Finally, because our framework models founders at the individual level, it does not explicitly aggregate features across founding teams or account for complementarities among co-founders. Prior research emphasizes the role of team dynamics, skill diversity, and role complementarity in shaping outcomes (Gompers & Lerner, 2020). Future extensions could incorporate team-based representations, such as averaged embeddings, diversity indices, or interaction terms, to capture collective FMF and better reflect the multi-founder reality of most startups.

7. Conclusion

This study presents a robust, technically grounded framework for early-stage startup evaluation through the lens of FMF. Leveraging large-scale, multi-source data including Crunchbase startup metadata, LinkedIn career trajectories, and X-based psychological signals, we construct a predictive FMF score that meaningfully forecasts startup success outcomes. Our results validate key entrepreneurial theories on the alignment of human capital, institutional prestige, and psychological traits with venture performance.

By integrating natural language processing, feature engineering, and explainable machine learning, we bridge theory and practice in a scalable and interpretable manner. The FMF framework offers a promising foundation for enhancing transparency and inclusion in venture capital, especially by surfacing underrepresented founders who align well with their market domains.

Future work could explore recent advances in Large Language Models (LLMs) (Maarouf et al., 2025) multilingual data integration, real-time founder scoring, and partnership with accelerators and VC firms for deployment in live deal flow environments. Broader collaboration with responsible AI initiatives may also help ensure the ethical implementation of such tools across diverse ecosystems and geographies.

References

- Arroyo, J., Corea, F., Jimenez-Diaz, G., & Recio-Garcia, J. A. (2019). Assessment of machine learning performance for decision support in venture capital investments. *IEEE Access*, 7, 124233–124243.
- Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management*, 17(1), 99–120.
- Becker, G. S. (1965). *Human capital: A theoretical and empirical analysis, with special reference to education*. National Bureau of Economic Research.
- Carreyrou, J. (2018). *Bad blood: Secrets and lies in a silicon valley startup*. Knopf.
- Chen, X., Yao, X., & Kotha, S. (2009). Entrepreneur passion and preparedness in business plan presentations: A persuasion analysis of venture capitalists' funding decisions. *Academy of Management Journal*, 52(1), 199–214.
- Choi, D. S., Sung, C. S., & Park, J. Y. (2020). How does technology startups increase innovative performance? the study of technology startups on innovation focusing on employment change in korea. *Sustainability*, 12(2), 551.
- Cohen, B. D., & Dean, T. J. (2005). Information asymmetry and investor valuation of ipos: Top management team legitimacy as a capital market signal. *Strategic Management Journal*, 26(7), 683–690.
- Colombo, M. G., & Grilli, L. (2005). Founders' human capital and the growth of new technology-based firms: A competence-based view. *Research policy*, 34(6), 795–816.
- Courtney, C., Dutta, S., & Li, Y. (2017). Resolving information asymmetry: Signaling, endorsement, and crowdfunding success. *Entrepreneurship Theory and Practice*, 41(2), 265–290.
- Cumming, D., & Johan, S. (2009). Pre-seed government venture capital funds. *Journal of International Entrepreneurship*, 7, 26–56.
- Dalziel, M., & Basir, N. (2024). The technological imprinting of educational experiences on student startups. *Research Policy*, 53(2), 104940.
- Davila, A., Foster, G., & Gupta, M. (2003). Venture capital financing and the growth of startup firms. *Journal of Business Venturing*, 18(6), 689–708.
- Dencker, J. C., & Gruber, M. (2015). The effects of opportunities and founder experience on new firm performance. *Strategic Management Journal*, 36(7), 1035–1052.
- Dimov, D. P., & Shepherd, D. A. (2005). Human capital theory and venture capital firms: Exploring “home runs” and “strike outs”. *Journal of business venturing*, 20(1), 1–21.
- Du, D., Wang, J., & Li, J. (2024). What drives intercity venture capital investment? a comparative analysis between multiple linear regression and random forest. *Humanities and Social Sciences Communications*, 11(1), 1–13.
- Eesley, C., & Wang, Y. (2017). Social influence in career choice: Evidence from a randomized field experiment on entrepreneurial mentorship. *Research policy*, 46(3), 636–650.
- Freiberg, B., & Matz, S. C. (2023). Founder personality and entrepreneurial outcomes: A large-scale field study of technology startups. *Proceedings of the National Academy of Sciences*, 120(19), e2215829120.
- Gage, D. (2012). The venture capital secret: 3 out of 4 start-ups fail. *Wall Street Journal*, 20.
- Gjurković, M., Karan, V. M., Vukojević, I., Bošnjak, M., & Snajder, J. (2021). PANDORA talks:

- Personality and demographics on Reddit. *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, 138–152.
- Gompers, P., & Kovvali, K. (2018). The other diversity dividend. *Harvard Business Review*. <https://hbr.org/2018/07/the-other-diversity-dividend>
- Gompers, P., & Lerner, J. (2020). How do venture capitalists make decisions? *Journal of Financial Economics*, 135(1), 185–205.
- Hashai, N., & Zahra, S. (2022). Founder team prior work experience: An asset or a liability for startup growth? *Strategic Entrepreneurship Journal*, 16(1), 155–184.
- Hellmann, T., & Puri, M. (2000). The interaction between product market and financing strategy: The role of venture capital. *Review of Financial Studies*, 13(4), 959–984.
- Hsu, D. H. (2007). Experienced entrepreneurial founders, organizational capital, and venture capital funding. *Research policy*, 36(5), 722–741.
- Hunter, D. S., Saini, A., & Zaman, T. (2017). Picking winners: A data-driven approach to evaluating the quality of startup companies [arXiv:1706.04229].
- Jeong, S. (2022). VC Funding and Bias Against Minority Entrepreneurs. *Entrepreneurship Theory and Practice*, 46(2), 350–371.
- Kaplan, S. N., & Lerner, J. (2016). Venture capital data: Opportunities and challenges. *Measuring entrepreneurial businesses: Current knowledge and challenges*, 413–431.
- Lerner, J. (2009). *Boulevard of broken dreams: Why public efforts to boost entrepreneurship and venture capital have failed—and what to do about it*. Princeton University Press.
- Lyonnet, V., & Stern, L. H. (2024). *Machine learning about venture capital choices* (SSRN Working Paper: No. 4035930).
- Maarouf, A., Feuerriegel, S., & Pröllochs, N. (2025). A fused large language model for predicting startup success. *European Journal of Operational Research*, 322(1), 198–214.
- McCarthy, P. X., Gong, X., Braesemann, F., Stephany, F., Rizoïu, M.-A., & Kern, M. L. (2023). The impact of founder personalities on startup success. *Scientific Reports*, 13(1), 17200.
- Packalen, M., & Bhattacharya, J. (2015). *New ideas in invention* (tech. rep.). National Bureau of Economic Research.
- Patel, N. (2015). 90% of startups fail: Here's what you need to know about the 10%. *Forbes*. <https://www.forbes.com/sites/neilpatel/2015/01/16/90-of-startups-fail-heres-what-you-need-to-know-about-the-10-that-wont/>
- Razaghzadeh Bidgoli, M., Raeesi Vanani, I., & Goodarzi, M. (2024). Predicting the success of startups using a machine learning approach. *Journal of Innovation and Entrepreneurship*, 13(1), 80.
- Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese networks. *Proceedings of EMNLP*, 3983–3993.
- Roberts, E. B., Eesley, C. E., et al. (2011). Entrepreneurial impact: The role of mit. *Foundations and Trends in Entrepreneurship*, 7(1–2), 1–149.
- Roche, M. P., Conti, A., & Rothaermel, F. T. (2020). Different founders, different venture outcomes: A comparative analysis of academic and non-academic startups. *Research Policy*, 49(10), 104062.
- Shepherd, D. A., Douglas, E. J., & Shanley, M. (2000). New venture survival: Ignorance, external shocks, and risk reduction strategies. *Journal of Business Venturing*, 15(5-6), 393–410.
- Silva, E., Beirão, G., & Torres, A. (2023). How startups and entrepreneurs survived in times of pandemic crisis: Implications and challenges for managing uncertainty. *Journal of Small Business Strategy*, 33(1), 84–97.
- Stein, S. (2023). The wework crash: What went wrong. *CNBC*. <https://www.cnbc.com/2023/11/02/wework-crash.html>
- Thirupathi, A. N., Alhanai, T., & Ghassemi, M. M. (2021). A machine learning approach to detect early signs of startup success. *Proceedings of the second ACM international conference on AI in finance*, 1–8.
- Unger, J. M., Rauch, A., Frese, M., & Rosenbusch, N. (2011). Human capital and entrepreneurial success: A meta-analytical review. *Journal of business venturing*, 26(3), 341–358.
- Yang, Y., Fang, Y., Wang, N., & Su, X. (2023). Mitigating information asymmetry to acquire venture capital financing for digital startups in china: The role of weak and strong signals. *Information Systems Journal*, 33(6), 1312–1342.
- Zider, B. (1998). How venture capital works. *Harvard Business Review*, 76(6), 131–139. <https://hbr.org/1998/11/how-venture-capital-works>