The Role of Team Characteristics and Business Fundamentals in Identifying High-Growth Entrepreneurial Companies: A Machine Learning Approach

Yi Jiang[1], Bin Ke[2], Zhen Long[3], and Hongye Lyu[4]

July 24, 2025

ABSTRACT

Using data from China's New Third Board and machine learning, this study examines the value of team characteristics ("jockey") and business fundamentals ("horse") in identifying high-growth entrepreneurial firms. We define team features using management resumes and business fundamentals using financial statements. To measure team characteristics from unstructured resumes, we employ (i) manual data collection and (ii) ChatGPT-based automatic feature extraction, including text embeddings or human-defined features. Our results show that models using manual collection and ChatGPT-extracted features perform similarly, but both surpass text embedding models. While both team features and business fundamentals predict high growth, business fundamentals add no incremental value once team features are included. Additionally, our best machine learning models outperform VC investors in identifying high-growth firms. These findings contribute to the "jockey vs. horse" debate, showing that team characteristics are stronger predictors of high-growth ventures, but they appear to be still underweighted in VC investment decisions.

We thank Yibin Liu, Daniel Rabetti and workshop participants at National University of Singapore, for helpful comments. The article is written with the assistance of ChatGPT.

[1] School of Economics and Management, Harbin Institute of Technology (Shenzhen), Xili University Town, Shenzhen, 518055, Guangdong, China. Email: jiangyi@hit.edu.cn.

[2] Department of Accounting, Business School, National University of Singapore, Mochtar Riady Building, BIZ 1, #07-30, 15 Kent Ridge Drive, Singapore 119245. Tel: +65 66013133. Fax: +65 67736493. Email: bizk@nus.edu.sg.

[3] School of Accounting, Zhongnan University of Economics and Law, Wuhan, Hubei Province, China. Email: zhenlong@whu.edu.cn.

[4] NUS (Chongqing) Research Institute, Chongqing, China. Email: e0787953@u.nus.edu.

1. Introduction

Venture capital (VC) serves as a crucial financial intermediary for financing early-stage entrepreneurial firms with high growth potential (Howell, 2020; Kaplan & Lerner, 2010). Lerner & Nanda (2020) find that nearly half of the IPOs by nonfinancial firms founded post-1980 are VC-backed, and most of the top global firms by market capitalization in May 2020 receive early venture capital backing. Research points to a causal link between VC financing and firm innovation (Lerner & Kortum, 2000) and firm performance (Bernstein et al., 2016).

Despite the fact that VCs prefer entrepreneurial firms exhibiting high growth (Chemmanur et al., 2011; D. Guo & Jiang, 2013), predicting high growth firms is inherently challenging. This is because most early-stage ventures' growth trajectories are uncertain and difficult to assess (Hall & Woodward, 2010; Kerr et al., 2014). Kaplan et al. (2009) propose a theoretical framework in which VCs evaluate both team characteristics ("the jockey") and business fundamentals ("the horse") when making investment decisions. In entrepreneurial ventures, particularly at early stages where uncertainty and information asymmetry are high, VCs tend to place greater emphasis on the team. This emphasis aligns with Rajan's (2012) model, which posits that entrepreneurial human capital is a critical differentiator in the early life of a firm.

Prior survey research (Gompers et al., 2020) and experimental research (Bernstein et al., 2017) suggest that VCs tend to prioritize team characteristics over business fundamentals in their investment decisions. However, it remains unclear whether this preference leads to optimal decision-making. Recent work suggests that the effect of CEO characteristics on firm performance diminishes for publicly listed firms when evaluated using out-of-sample predictive models that incorporate both CEO and firm-level information (Shimao et al., 2025). Moreover, even when team characteristics do have predictive

power, VC investors may still fail to fully utilize them. For instance, Cohen et al. (2013) find that R&D capabilities, though relatively simple to assess, are frequently underweighted in valuation processes.

To provide direct evidence on the value of team characteristics in predicting high-growth startup firms, we propose several interrelated research questions. The first research question (RQ1) assesses the value of team characteristics ("the jockey") in predicting high-growth startup firms. The second research question (RQ2) evaluates the value of business fundamentals ("the horse") in predicting high-growth startup firms.

Team characteristics ("the jockey") and business fundamentals ("the horse") may interactively affect venture performance. Arif et al. (2025) provide evidence that the value of managerial capital is contingent on both firm fundamentals and external market conditions, suggesting that team and firm attributes do not exert their influences in isolation. Accordingly, our third research question (RQ3) explores whether integrating team characteristics and business fundamentals within predictive models further enhances the ability to identify high-growth firms.

We investigate the three research questions using entrepreneurial firms listed on China's National Equities Exchange and Quotations (NEEQ), commonly known as the New Third Board (NTB). Established in 2013, the NTB offers a platform for entrepreneurial companies to list their shares, offering lower listing thresholds than traditional IPO markets. Unlike conventional exchanges that require profitability and revenue benchmarks, the NTB admits firms based on a viable business model and at least two years of audited financials, making it a key platform for innovation-driven startups and small to medium-sized enterprises (SMEs). Our sample covers 12,567 startup companies listed on the NTB from 2013 to 2022.

It is difficult to test the three research questions because being privately held, most entrepreneurial

firms offer little information about their management teams and business fundamentals. The NTB is an ideal setting for investigating our research questions due to its extensive disclosure requirements. For team characteristics, NTB regulations mandate detailed disclosure of top management team resumes, which provide rich but unstructured data. In this study, the top management team refers specifically to directors, supervisors, and senior executives, as disclosed by NTB firms in their prospectuses. Similarly, NTB-listed firms must submit at least two years of audited financial statements before listing, offering structured and reliable business fundamentals data. These statements serve as a high-quality source for assessing the role of business fundamentals for predicting high-growth firms. Furthermore, NTB firms continue to publish audited financial statements annually after listing, allowing for longitudinal analysis of firm performance. This enhances our research design by enabling predictive modeling of how team and business fundamentals information influence subsequent business outcomes.

We measure high growth using the sales growth rate within the first three years post-listing. We choose a three-year window for two key reasons. First, the dataset spans from 2013 to 2022, making a longer timeframe impractical as it would significantly reduce available observations. Second, three years provide a meaningful forecasting horizon for venture capitalists seeking to make informed investment decisions.

To align with VC investment strategies, we transform continuous sales growth into a binary variable, distinguishing firms with exceptional growth from the rest. This approach captures the essence of how VCs evaluate high-potential startups. The classification threshold for high growth is set at 200% cumulative sales growth over three years, equivalent to an annualized rate of approximately 45%. This threshold is based on historical growth distributions and aligns with firms in the top 10-15% of the dataset, ensuring that only ventures with truly exceptional growth are classified as high-growth firms.

With this outcome measure defined, we investigate the predictive value of team characteristics, business fundamentals, and their combination in identifying high-growth firms. Specifically, we aim to identify NTB firms with high post-listing growth potential using only information available at the time of the NTB listing. We restrict our models' predictors to pre-listing data only because we aim to address the prediction challenge of venture capitalists who need to assess a startup firm's growth potential based on information available at the time of the NTB listing. While we could construct a more accurate prediction model at a later point in time (e.g., one year after listing) by incorporating more post-listing information sources such as stock trading data, the value of such a prediction model to VCs is reduced as the NTB firm's valuation will likely be much more expensive due to reduced uncertainty of its future growth.

To address the potential interdependence between team characteristics and business fundamentals, we employ XGBoost (T. Chen & Guestrin, 2016), a tree-based ensemble method known for capturing nonlinear relationships, interaction effects among predictors, and handling missing values effectively. By integrating both dimensions within a single model, XGBoost allows for flexible modeling of any latent relationships between team characteristics and business fundamentals.

To avoid any look-ahead biases, we structure our training and test data as follows:

- 2018 Test Year: Training data includes firms listed in 2013 and 2014.
- 2019 Test Year: Training data includes firms listed in 2013, 2014, and 2015.

For each test year, we randomly allocate 20% of the training sample as a validation set to optimize hyperparameter selection for our machine learning models. Once optimal hyperparameters are determined, we combine the training and validation samples to train the final models, assessing their out-of-sample predictive performance on the test set.

Following Bao et al. (2020), we evaluate our model's ability to rank high-growth firms using NDCG@K (Normalized Discounted Cumulative Gain at rank K), a metric that assesses whether the firms ranked highest by the model are indeed those that exceed the 200% sales growth threshold. To ensure meaningful comparisons, we set K to the actual number of NTB firms in each test set that achieved post-listing three-year sales growth exceeding 200%. This approach mirrors real-world conditions. Additionally, we incorporate PRECISION@K and RECALL@K to provide a comprehensive assessment. PRECISION@K measures how many of the top K firms identified by the model actually achieve high growth, while RECALL@K quantifies how many high-growth firms are successfully captured within the top K predictions.

Our first research question explores the value of team-related information from top management resumes to predicting high-growth firms. As resumes are rich yet unstructured data sources, varying in format across individuals and firms, we propose three main approaches to feature extraction. The first, human-coded features, involve the manual extraction of explicitly defined variables based on established theoretical constructs, with humans directly applying predefined frameworks to structure resume data. The second, hybrid feature extraction, integrates large language models (LLMs), such as GPT-3.5, into the extraction process. While still grounded in the same theoretical framework as human-coded features, this method employs GPT-3.5 to assist in identifying and structuring relevant information from resumes, which is then reviewed and aggregated by human coders. The third approach, machine-coded features, departs from predefined constructs and instead relies on embedding models to automatically identify patterns within resume text. To minimize potential look-ahead biases, we employ GPT-3.5 rather than the more powerful GPT-4 in the hybrid feature extraction process, as the training data cutoff date is September 2021 for GPT-3.5 compared to April 2023 for GPT-4.

As part of the first research question, we compare the effectiveness of hybrid features, human-coded features, and machine-coded embeddings in predicting high-growth firms. While manual coding is highly accurate, it is labor-intensive and costly. In contrast, GPT-generated features offer a faster and more cost-effective alternative. To assess whether GPT can serve as a viable substitute, we evaluate their predictive performance. Our results show that the hybrid features model achieves an NDCG score of 0.309, outperforming the human-coded features model (0.245). This suggests that the GPT model not only processes data more efficiently but also maintains a high level of accuracy in predicting high growth.

Additionally, we examine whether machine-coded features based on embeddings can capture team-related information that human-coded and hybrid methods might overlook. Our results indicate that the machine-coded Features model achieves an average NDCG score of 0.205. This score is lower than both the hybrid features (0.309) and the human-coded features (0.245). These findings suggest that while embeddings encode a broad range of information, they may overlook critical features necessary for high growth prediction.

Overall, our findings highlight that team characteristics hold substantial predictive value in identifying high-growth firms. To extract information from unstructured resumes, GPT emerge as a strong alternative to manual coding, offering comparable accuracy while significantly improving efficiency. Meanwhile, the inferior performance of text embeddings suggests that predictive models benefit from feature extraction informed by domain knowledge rather than relying solely on automated deep-learning representations.

Next, we examine our second research question, which focuses on the value of business fundamentals information ("the horse") in predicting high-growth firms. We use data from a firm's three

standard financial statements to capture its business fundamentals as financial statements are designed to measure a firm's financing, investing, and operating activities. We employ two primary approaches: a theory-driven Selected Financial Data Method and a data-driven Comprehensive Financial Data Method. The theory-driven approach, inspired by Hand (2005), extracts key financial metrics linked to entrepreneurial firms, such as cash holdings, non-cash assets, long-term debt, revenues, R&D expenditures, cost of sales, and SG&A expenses. In contrast, the data-driven approach utilizes all key components from the balance sheet, income statement, and cash flow statement, allowing machine learning models to identify patterns beyond theory-guided indicators.

Our findings confirm that business fundamentals contain predictive value in identifying high-growth firms. Specifically, the Selected Financial Data Method achieves an average NDCG score of 0.239, outperforming the Comprehensive Financial Data Method, which scores 0.185. This result suggests that a targeted, theory-driven selection of financial indicators is as effective, if not superior, to a purely data-driven approach. These findings highlight the importance of strategically selecting business fundamentals features, as theory-based approaches may enhance predictive performance in identifying high-growth firms. However, the business fundamentals model underperforms the team characteristics model.

For our third research question, we investigate the benefit of combining team ("jockey") and business fundamentals ("horse") information in predicting high-growth firms. To test this, we combine the most effective team information extraction methods (human-coded features and hybrid features) with the most effective business fundamentals information extraction method (Selected Financial Data). The best integrated model achieves an NDCG score of 0.305. For comparison, hybrid features alone yield an NDCG score of 0.309, human-coded features score 0.245, and Selected Financial Data alone

records 0.239. This indicates that while team information significantly enhances prediction when combined with business fundamentals information, business fundamentals information offers only marginal additional value when team information is already considered.

These results collectively highlight team information plays a more critical role than business fundamentals information in predicting high-growth startup firms. This finding aligns with prior survey and experimental research suggesting that VCs prioritize the "jockey" over the "horse" when evaluating startups (Bernstein et al., 2017; Gompers et al., 2020). However, a crucial question remains: do VCs fully appreciate the value of team characteristics in predicting high growth firms?

Persistent biases in VC decision-making such as gender (Ewens & Townsend, 2020; Howell & Nanda, 2024; Kanze et al., 2018) and race (Fairlie et al., 2022) suggest that VC investors may not always make optimal use of available information. While VCs claim to value team-related factors, their ability to extract, interpret, and effectively utilize these signals in real investment decisions is uncertain.

To investigate this, we benchmark our model's performance against real-world VC investment decisions. To ensure a fair comparison using listing information, we focus on firms that receive new VC investments within the two years preceding their listing date. Applying this criterion, we identify 3,936 such NTB firms within our sample.

For each year under review, we assess VC investors' ability to identify high-growth firms by calculating the average value of the 3-year realized high-growth dummy variable. The mean value of the outcome measure for VC-backed firms serves as a measure of VCs' selection precision. Additionally, we compute the recall rate of VC investments, defined as the proportion of all high-growth firms that received VC backing.

To compare VC performance with our predictive models, we define N as the number of NTB firms

with VC investments in each year. For each model, which features different sets of information, we compute both PRECISION@N and RECALL@N, considering the top N firms ranked by machine learning as predicted high-growth firms.

Our findings show that our machine learning models exhibit strong predictive ability, often outperforming VCs. On average, VC investment achieves a precision of 0.136, while models leveraging team information alone yield a higher precision: 0.171 for human-coded features and 0.164 for hybrid features. Similarly, VCs achieve a recall rate of 0.300, whereas team-based models reach 0.388, capturing a larger share of actual high-growth firms. Integrating business fundamentals and team information further enhances performance, achieving the highest precision (0.190) and recall (0.450).

These results suggest that our ML models, particularly those incorporating team information, outperform VC investors in identifying high-growth firms. This implies that VCs may not be fully leveraging the most predictive signals available to them. Moreover, our findings reveal a critical gap between VCs' stated emphasis on team characteristics and their ability to assess them effectively in practice. Ultimately, "valuing teams" does not necessarily equate to "evaluating teams effectively," highlighting potential inefficiencies in VC decision-making.

One potential concern regarding the relatively poor performance of VC investment compared to our ML model is that future sales growth may not be the primary factor driving VC investment decisions. However, we mitigate this concern with three key points.

First, prior research (Chemmanur et al., 2011; D. Guo & Jiang, 2013) demonstrates that VC in both the U.S. and China are more likely to invest in firms with high sales growth. Additionally, these studies highlight that VCs play an active role in enhancing a firm's sales growth post-investment. This underscores the importance of sales growth as a key consideration throughout different stages of VC

involvement.

Second, we examine new VC investments made within three years post-listing for all firms in the test years (2018 and 2019). If VCs prioritize firms with strong post-listing growth, they should be more likely to invest in firms demonstrating high performance after listing. We find results consistent with this prediction.

Third, we find that our best ML model's predicted probabilities can explain post-listing VC investments. This finding suggests that VCs invest in firms with high-growth potential only after they receive more confirmation information post the listing, even though these high growth firms could be predicted at the time of listing. This finding suggests that VCs have acted too late, missing opportunities to invest earlier in high-growth firms.

The rest of the paper is organized as follows. Section 2 reviews how this study contributes to literature. Section 3 provides the institutional background of our setting. Section 4 details the methods for extracting team information, while Section 5 details business fundamentals information extraction and other aspects of the research design. Section 6 evaluates the predictive performance of different models. Section 7 presents the SHAP analysis to interpret feature importance based on our machine learning model. Section 8 benchmarks our results against VC investment performance. Finally, Section 9 concludes the paper.

2. Prior Literature and Contributions

Our study makes three main contributions. First, our study contributes to the entrepreneurial finance literature by assessing the relative importance of team information and business fundamentals information in predicting firm success. Prior research in this domain has primarily focused on causal inference, aiming to identify the factors that explain VC investment behavior. These studies can be

broadly classified into three categories. The first category, VC surveys and interviews studies, relies on surveys or interviews to quantify how investors evaluate different factors when making investment decisions (J. H. Block et al., 2021; Gompers et al., 2020). The second category, experimental manipulation of information, involves controlled experiments where specific startup characteristics are altered to observe how VC investors react (Bernstein et al., 2017; Shepherd, 1999). The third category, archival studies, examines past startup performance and founder track records to identify success patterns, often using survival analysis or regression (J. Block et al., 2019; Kaplan et al., 2009).

Our study shifts from causal inference to prediction by evaluating the usefulness of prior causal theories on jockey vs horse in predicting high-growth firms. Traditionally, accounting and finance research emphasizes causal inference, aiming to identify underlying reasons behind corporate behaviors and outcomes. However, as Guo et al. (2025) argue, causal inference is not an end but rather a critical element within a broader analytical framework that encompasses descriptive, predictive, and prescriptive analyses.

Guo et al. (2025) highlight that causal inference, as part of descriptive analysis, provides insights into historical relationships and informs predictive models. While causal inference is essential for theory development, predictive analysis evaluates the practical utility of these theories in forecasting real-world outcomes. Given growing concerns regarding the robustness, replicability, and economic significance of causal inference findings, predictive analysis offers a complementary approach by assessing the real-world applicability and effectiveness of causal theories through out-of-sample predictions.

Consistent with recent discussions by Gow et al. (2023), who emphasize the complementary value of predictive methods in complex empirical contexts, our study evaluates how well causal theories

translate into predictive performance. Specifically, we advance the jockey vs. horse debate by demonstrating that relative importance of jockey vs. horse in predicting high-growth startup firms.

Second, our study contributes to the machine learning literature on the prediction of entrepreneurial firm performance. Although a growing body of research uses machine learning to predict entrepreneurial success (e.g., Krishna et al., 2016; Sharchilev et al., 2018; L. Zhang et al., 2022; S. Zhang et al., 2021), few works systematically distinguish the predictive roles of team characteristics versus business fundamentals. Two notable exceptions are Davenport (2022) and Lyonnet and Stern (2022).

Lyonnet and Stern (2022) investigate how VCs make investment decisions, using machine learning models as benchmarks. Like our study, they find that VC decisions are not fully optimal. However, their analysis omits the timing of investments, complicating comparisons between actual VC decisions and model predictions. Moreover, they do not directly assess the relative importance of the "jockey" versus the "horse" in forecasting firm performance. Their reliance on survey data from individual entrepreneurs limits their ability to capture team-level dynamics. In contrast, our study incorporates detailed team-level data and explicitly evaluates the role of such characteristics in predicting outcomes. We demonstrate that including team-level information beyond CEO traits enhances both predictive accuracy and robustness. Finally, their dependent variable—value added in year five—is a profit-based metric that may be ill-suited for evaluating performance in high-growth startups that often experience substantial early losses.

Similarly, Davenport (2022) analyzes VC investment decisions using Pitchbook data on startups emerging from top accelerator programs, with post-accelerator investment and exit events as outcome variables. Like Lyonnet and Stern (2022), he benchmarks VC decisions against machine learning

models and finds evidence of suboptimal decision-making. His model incorporates financial metrics and CEO-specific qualitative data at the time of funding but omits broader team-level characteristics. In contrast, our study leverages full team resumes and financial statements, employing manual coding and GPT-based tools to extract richer, theoretically grounded features. By assessing the predictive value of team attributes alongside firm fundamentals—and focusing on realized high-growth outcomes rather than intermediate financing milestones—we provide a more comprehensive and accurate framework for forecasting entrepreneurial success.

Our last contribution advances the literature on the use of Large Language Models (LLMs) in processing textual data in accounting and finance. Previous studies have primarily used LLMs for text interpretation and summarization, extracting insights from complex financial documents. Applications include analyzing conference calls (Jha et al., 2023; Kim et al., 2023), news articles (Lopez-Lira & Tang, 2023), 10-K filings (Armstrong, 2023; Kim et al., 2023), and analyst reports (Li et al., 2023). We are one of the first to explore the usefulness of LLMs in extracting information from unstructured resumes. In addition, a critical gap in the literature is the lack of direct comparisons between LLM-generated and human-coded data in structured information extraction tasks. While prior research highlights LLMs' capabilities in processing financial text, it does not assess whether LLMs can match or exceed human performance in tasks traditionally requiring manual coding. Our study fills this gap by conducting a comparison between GPT-generated hybrid features and manually human-coded features in processing team information within the context of a major managerial decision.

3. Institutional Background

This study examines entrepreneurial firms listed on China's National Equities Exchange and Quotations (NEEQ), commonly known as the New Third Board (NTB). Launched in 2013 with the

approval of the State Council[1], the NTB serves as a national over-the-counter exchange tailored for early-stage and growth-oriented enterprises. It offers significantly lower listing thresholds compared to conventional initial public offerings (IPOs) on the main stock exchanges. Traditional IPO markets impose stringent financial and operational requirements—such as sustained profitability, revenue thresholds, and minimum scale—that are often unattainable for young ventures. In contrast, the NTB's listing criteria are purpose-built for innovation-driven startups and small and medium-sized enterprises (SMEs). Applicants need only demonstrate a viable business model and provide at least two years of audited financial statements. This streamlined admission process reflects the NTB's policy mandate to foster entrepreneurial growth by filling a critical funding gap for promising young companies.

Reflecting above supportive policies, the NTB has grown rapidly into a major marketplace for entrepreneurial finance. By 2022, it hosted 6,580 listed companies and recorded a total trading volume of 79.8 billion RMB, facilitating approximately 23.1 billion RMB in SME financing[2]. This breadth and liquidity have made the NTB a focal point for venture capital (VC) investment. Many NTB-listed companies are venture-backed and tend to be more mature than the average startup, aligning them with the later-stage firms in typical VC portfolios (Jiang & Ke, 2022).

Beyond its market impact, the NTB offers a particularly rich environment for research that requires detailed information on both entrepreneurial teams and business fundamentals. In terms of team information, NTB regulations mandate extensive disclosure about each company's management and governance personnel. Listing documents and subsequent filings detail the composition of the board of directors, the supervisory board, and the senior executive team. This disclosure includes a wealth of information such as names, positions, gender, dates of birth, terms of office, and professional

[1] https://www.gov.cn/zhengce/content/2013-12/14/content_4024.htm
[2] https://www.neeq.cc/static/statisticdata.html

backgrounds[3]. Notably, the lack of a standardized format for these resumes poses a unique challenge for extracting team-level information.

Similarly, NTB-listed firms must disclose at least two years of audited financial statements prior to listing. These filings yield a wealth of structured data on each firm's business fundamentals information at the time of listing. This trove of business fundamentals information provides the foundation for assessing the role of business fundamentals ("the horse") for predicting high-growth firms.

Moreover, since NTB firms continue to publish audited financial statements annually after listing, the data enables a longitudinal analysis of firm performance. This feature strengthens the research design by allowing observation of how early team and business fundamentals correspond to subsequent outcomes, thereby permitting predictive models of high growth.

In summary, the NTB offers an empirical setting that directly supports this study's research question on the value of entrepreneurial team and business fundamentals information and facilitates the analysis of how these factors jointly contribute to high growth.

4. Team Information Extraction

To evaluate the predictive value of team characteristics ("the jockey") in identifying high-growth firms, we explore methods for extracting team-related information from top management team resumes.[4] As defined in NTB firm prospectuses, the top management team includes directors, supervisors, and senior Executives, in accordance with NTB's disclosure requirements. Resumes are rich but unstructured data sources, unlike standardized financial statements. They vary in format across

[3] Standards for the Contents and Formats of Information Disclosure by Unlisted Public Companies No. 9, No. 10, and No. 13

[4] As the CEO of startup firms tends to play a dominant role in major business decisions, one may wonder whether it is sufficient to consider the CEO's characteristics alone in predicting firm performance. As shown in Appendix 4, we find that our machine learning prediction models exhibits both better and more robust performance when using team information than using CEO-alone information.

individuals and firms, contain qualitative information that is difficult to quantify, and embed implicit signals (e.g., tone, strategic arrangement of content) that may influence firm performance but are challenging to extract systematically. These complexities necessitate structured feature extraction methods.

To address this challenge, we classify feature extraction methods into three main approaches. The first, human-coded features, consists of explicitly defined features based on established theoretical knowledge and extracted directly by humans using predefined frameworks. The second, hybrid feature extraction method, leverages large language models (LLMs), such as GPT, to assist in identifying relevant information, while still grounded in predefined theoretical frameworks similar to human-coded features. The third approach, machine-coded features, involves the fully automatic extraction of patterns from resume content using machine learning techniques, particularly through embeddings.

Each approach has distinct advantages and limitations. In the following sections, we examine these methods in greater detail and evaluate their effectiveness in predicting the high growth of entrepreneurial firms.

4.1 Human-Coded Features

The extraction of human-coded features is grounded in well-established theories of team effectiveness and entrepreneurial success. Unlike machine-coded methods, which rely on implicit pattern recognition, human-coded features require a structured causal framework to ensure that the extracted variables are theoretically meaningful and empirically relevant.

Our first step involves extracting individual attributes from resumes. Drawing on prior research in entrepreneurship and VC, we identify theoretically justified and empirically measurable proxies that can be reliably derived from resume data. To maintain rigor, we focus on attributes that are both

observable and quantifiable.

Extensive research in entrepreneurship and VC has established that certain individual attributes significantly impact firm success. Specifically, studies have highlighted the importance of demographic characteristics (e.g., gender (Sharchilev et al., 2018), age (Esen et al., 2023), ethnicity (Calder-Wang et al., 2021), nationality (Calder-Wang et al., 2021), educational background (Arroyo et al., 2019)), professional experience (e.g., entrepreneurial and industry experience (Te et al., 2022)), and network affiliations (e.g., prior company (Beckman et al., 2007), university spin-offs (Ortín-Ángel & Vendrell-Herrero, 2010), VC connections) in entrepreneurial success.

Building on these insights, we systematically extract ten key individual attributes from the resumes, selected for their theoretical relevance and empirical measurability. To capture team characteristics, we aggregate these individual attributes within an analytical framework that aligns with established models of team effectiveness.

To guide the integration of individual attributes into meaningful team-level constructs, we draw upon validated frameworks from Industrial/Organizational Psychology and Organizational Behavior. Specifically, we base our approach on the team effectiveness frameworks of Mathieu et al. (2017, 2019), which highlights how team composition, structure, and mediating mechanisms shape organizational outcomes.

The foundation of our approach builds on the evolution of team effectiveness models, particularly the transition from the Input-Process-Outcome (IPO) model (McGrath, 1964) to the Input-Mediating-Outcome (IMO) model (Ilgen et al., 2005; Marks et al., 2001). The IPO model posits that team performance is driven by inputs—such as individual, team, organizational, and contextual characteristics—that shape outcomes through team processes, such as communication and coordination.

However, subsequent research reveals that the relationship between inputs and outcomes is more complex than initially assumed. Team processes alone could not fully account for how inputs translate into results, leading to the IMO model, which introduces mediating mechanisms as essential factors bridging inputs and outcomes. These mediating mechanisms include behavioral processes, collective team affect, cognition, and emergent states, which dynamically influence firm performance (Ilgen et al., 2005; Marks et al., 2001).

Building on this, Mathieu et al. (2017) expand the IMO framework to account for the dynamic and co-evolving nature of teams. They emphasize that key inputs can be categorized into structural features and compositional features, which, in conjunction with mediating mechanisms, influence team and firm-level outcomes. These three components—inputs, mediating mechanisms, and outcomes—interact over time, forming reciprocal relationships that shape organizational performance.

Within this framework, individual attributes are systematically aggregated into three core dimensions: compositional features, structural features, and mediating mechanisms. This categorization facilitates the translation of individual attributes into team-level constructs, aligning with established models of team effectiveness. Next, we elaborate on each category, clarifying their theoretical basis and the specific aggregation methods applied.

Compositional features relate to “the combination of members” characteristics by Mathieu et al. (2017). This category includes average member attributes, diversity, and faultlines. Average member attributes are essentially the mean values or summary indexes that represent members’ characteristics collectively. These metrics provide a summary of what the "average team member" is like. Diversity, on the other hand, refers to the heterogeneity of team member characteristics. It captures how much individuals deviate from the team mean on various attributes. Group faultlines, as defined by Lau &

Murnighan (1998), are hypothetical dividing lines that split a group into subgroups that are homogeneous with respect to a particular set of diversity attributes. Faultlines reflect the potential for subgroup formation within a team based on the alignment of multiple characteristics. This concept extends beyond traditional diversity research by not only considering the presence of various attributes but also emphasizing their specific configuration. The faultline approach assumes that multiple characteristics can align simultaneously, leading to potential divisions within the team.

To translate individual attributes into team compositional features, we apply systematic aggregation methods. Average member attributes are computed using mean values (first-order measures), while diversity is assessed through variance or tailored diversity indices. Specifically, we assess demographic attributes such as gender (proportion and variance), age (mean and variance), ethnicity and nationality (proportions and diversity scores), and education (average educational attainment, elite university proportions, and educational diversity). Additionally, we measure professional experience through entrepreneurial experience (proportion and variance) and industry experience (mean and variance). We also incorporate network affiliations, such as prior company associations (mean and diversity scores), university spin-offs (proportion and variance), and VC connections (proportion and variance). A comprehensive summary of these proxies, including detailed measurement formulas, is provided in Panel A of Table 1.

To aggregate individual attributes into a team-level measure of faultlines, we adopt the Average Silhouette Width (ASW) method by Meyer et al. (2014). Faultlines refer to potential subgroup divisions within a team based on the alignment of multiple attributes. The ASW method quantifies how well team members fit within their respective subgroups by evaluating the degree of homogeneity within each subgroup and the distinctness between different subgroups.

For measuring faultlines, we adopt the Average Silhouette Width (ASW) method by Meyer et al. (2014). Faultlines refer to potential subgroup divisions within a team based on the alignment of multiple attributes. The ASW method quantifies how well team members fit within their respective subgroups by evaluating the degree of homogeneity within each subgroup and the distinctness between different subgroups. It evaluates subgroup structure through two key steps[5]:

Step 1: Clustering the team into potential subgroups.

Step 2: Calculating the silhouette width for each cluster configuration to determine the optimal subgroup structure.

For each configuration, the ASW algorithm computes a silhouette width for every team member. This metric indicates how well a member fits within their assigned subgroup compared to other possible subgroups. The silhouette width ranges from -1 to +1, where values closer to +1 indicate a strong alignment with the subgroup (high homogeneity), and values near 0 suggest ambiguity in subgroup assignment. The configuration with the highest average silhouette width is selected as the best representation of the team's subgroup structure, reflecting the optimal faultline strength. Higher ASW values indicate stronger and more distinct faultlines, suggesting clearer subgroup divisions within the team.

In our study, faultline strength is derived by clustering across 10 key individual attributes: gender, age, ethnicity, nationality, education, entrepreneurial experience, industry experience, prior company affiliation, university spin-offs, and VC connections. A comprehensive overview of these proxies, including explicit measurement definitions, is provided in Panel B of Table 1. We use the ASW value to measure the faultline strength of the management teams by incorporating the individual attributes of

[5] See appendix 1 for a detailed breakdown of these steps.

all team members.

Structural features refer to the formalized organizational aspects of a team, which influence how team members interact and perform. These include variables such as task scope and complexity, team interdependence, virtuality, and structural contingencies.

Structural features typically relate to formal organizational dynamics that may not be fully observable through resumes alone. To address this limitation, we use the individual attributes of the ultimate individual controlling shareholder as proxies for structural influence within the team. This approach acknowledges the disproportionate impact that key individuals could exert on team structure and decision-making. We identify the ultimate individual controlling shareholders using a hierarchical approach: first, we consider the Chairman of the Board; if their data is unavailable, we use the CEO's information. If neither is accessible, structural features are recorded as missing.

We then aggregate the 10 key individual attributes of these controlling figures as proxies for structural features: gender, age, ethnicity, nationality, education, entrepreneurial experience, industry experience, prior company affiliations, university spin-offs, and VC connections. Incorporating these attributes helps capture the relative significance of key individuals within the team. A comprehensive overview of these proxies, including precise measurement formulas, is provided in Panel A of Table 1.

Mediating mechanisms include the psychological and behavioral processes that bridge inputs and outcomes. These include conflict, creativity, cohesion, trust, decision-making, information sharing, team processes, among others. These mechanisms determine how team inputs translate into firm performance.

While mediating mechanisms are critical for understanding team dynamics and performance, they are inherently dynamic and interaction-based and thus cannot be directly measured from static resume

data. However, because compositional and structural features are influenced by mediating mechanisms, certain aspects of these processes may be indirectly inferred. Despite these indirect links, we recognize the limitation that our analysis focuses on inputs, and does not capture the mediating dynamics themselves.

Building upon this structured framework, we employ a specific methodology for extracting human-coded features. This traditional method involves a team of research assistants located in mainland China manually extracting key data from individual resumes and converting this unstructured information into structured formats.[6] Subsequently, data from individual resumes is aggregated to generate human-coded features. The process involves using algorithms such as regular expressions[7] to assist in data extraction. Despite its precision, this method is labor-intensive, costly, and time-consuming.

Human-coded features benefit from direct insights drawn from existing literature and domain knowledge. Integrating this knowledge ensures that the extracted variables align closely with established theories and models known to influence business outcomes. This approach facilitates a transparent, interpretable, and theory-driven extraction of team-related information. Human-coded features offer a high degree of accuracy. However, this process demands substantial resources, as it is labor-intensive, costly, and time-consuming. These constraints limit its scalability, making manual coding less practical for large-scale datasets commonly encountered in business research. In this study, we examine the predictive performance of human-coded features in forecasting high-growth firms.

[6] As the Chinese RAs performed the manual data collection before the availability of the ChatGPT Plus subscription, we are certain that the RAs have not used ChatGPT for the feature extraction.

[7] Regular expressions are pattern-matching tools used in text processing to identify and extract specific data from unstructured formats. For example, a regular expression like \d{2}/\d{2}/\d{4} can be used to identify date formats, helping to extract dates of employment or education from resumes. While regular expressions facilitate the initial identification of key information like phone numbers, email addresses, and educational credentials, they require careful implementation by research assistants to ensure data is accurately structured for analysis.

4.2 Hybrid Features

While the extraction process differs, the hybrid feature extraction method is grounded in the same theoretical framework and focuses on the same set of individual attributes as the human-coded approach. The ten key individual attributes, and the aggregation into team-level constructs (compositional features, structural features, and mediating mechanisms) all remain consistent. The critical distinction lies in the methodology used to extract these attributes from resumes: rather than relying on a team of research assistants manually converting unstructured resume data into structured formats, this approach employs LLMs to streamline and scale the extraction process.

This method integrates GPT-3.5 into the initial stages of individual resume analysis. Carefully crafted prompts guide GPT in extracting targeted information, transforming unstructured resume data into structured formats[8]. Human coders then aggregate these structured outputs at the team level, following the same analytical framework used for human-coded features. This approach leverages the speed and scalability of GPT, while maintaining theoretical rigor in variable selection and aggregation.

To minimize potential look-ahead biases, we employ GPT-3.5 rather than the more advanced GPT-4[9], as GPT-3.5's training data is capped at September 2021, whereas GPT-4 includes data up to April 2023. Our dependent variable, which indicates whether a start-up firm is identified as a high-growth firm, is assessed over the three years following its public listing. For the 2018 test year, this outcome is evaluated over 2019 to 2021, with financial statements typically disclosed around mid-April of each following year (for example, 2019 reports were released in April 2020). As a result, there is a small risk of look-ahead bias for the 2018 test year. However, this risk becomes negligible for the 2019 test year,

[8] Appendix 2 describes the methodology and specific prompts used to construct the hybrid features from individual resumes.
[9] While GPT-3.5 supports few-shot and chain-of-thought prompting when manually instructed, its performance is weaker than GPT-4 in both respects. GPT-4 demonstrates greater consistency, deeper reasoning, and more reliable inference when handling few-shot examples and multi-step reasoning tasks. We use GPT-3.5 to minimize look-ahead bias, given its earlier training data cutoff.

since GPT-3.5's training data includes almost no information from the period covering 2020 to 2022 outcomes. Additionally, in our subsequent machine-coded feature extraction, we also utilize the text-embedding-3-large model, which shares the same September 2021 training cutoff. This further ensures that our analysis minimizes the potential for look-ahead bias. Importantly, our subsequent results are robust across both test years, with particularly strong results observed for 2019, reinforcing the validity of our approach and indicating that any residual look-ahead bias is unlikely to influence our findings.

The hybrid feature extraction method offers substantial advantages, including rapid processing, scalability, and cost-efficiency. However, the use of GPT introduces specific challenges, such as potential overgeneralization to irrelevant contexts, misinterpretation of sector-specific terminology, and ambiguity arising from its probabilistic responses. We aim to assess whether hybrid features can effectively substitute for human-coded features, considering that manual coding, while highly accurate, is slow and resource-intensive, whereas GPT-assisted extraction is faster and more cost-effective. Therefore, we compare the performance of both human-coded and hybrid features in forecasting high-growth firms.

4.3 Machine-Coded Features

Machine-coded feature extraction represents a paradigm shift from the human-coded and hybrid methods previously discussed. Unlike human-coded or hybrid feature extraction, which relies on insights drawn from existing literature and domain knowledge, machine-coded extraction employs machine learning algorithms to autonomously identify and generate features. These features, often derived from complex patterns within the data, may not be immediately interpretable by humans but offer a means of uncovering insights.

To unlock these insights, we encode human language into a mathematical space. Each dimension

in this space can be seen as a feature, which is a distilled essence of the information contained within the text. Whether it's the frequency of words, the co-occurrence of terms, or the semantic proximity between concepts, each numerical value represents a distilled component of the original textual content. As highlighted in the work of Gentzkow et al. (2019), this text representation is fundamental to text analysis. The following overview traces the evolution of embedding techniques, illustrating how each advancement incrementally addresses prior limitations and enhances our ability to capture textual meaning.

The initial techniques in text representation largely revolved around quantitative assessments of textual data, primarily through word counting mechanisms. Among the most prominent of these early methods are the bag-of-words (BoW) model, n-gram analysis, and Term Frequency-Inverse Document Frequency (TF-IDF).

The BoW model represents text data by breaking down content into individual words, disregarding syntax and word order. This approach transforms text into a sparse matrix where each row represents a document and each column represents a unique word in the overall corpus, with cell values quantifying the presence of words in each document. Building on the BoW model, n-gram analysis introduces context to the representation by considering sequences of “n” words as single features. For example, while a unigram (1-gram) considers individual words, a bigram (2-gram) analyzes pairs of consecutive words, and a trigram (3-gram) focuses on triplets. TF-IDF further refines the approach by weighing the frequency of words against their commonness across documents. Term Frequency (TF) assesses the prevalence of a word within a single document, while Inverse Document Frequency (IDF) measures how common or rare a word is across the entire corpus. By multiplying these values, TF-IDF assigns higher scores to words that are frequent in a document but rare across others, thereby highlighting words

that are potentially more indicative of a document's unique content.

Moving beyond mere frequency counts, more advanced models use contextual information to produce dense vector representations, known as word embeddings. Typical models include Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014).

Word2Vec operates on the principle that words appearing in similar contexts tend to have similar meanings. It employs two distinct architectures: Continuous Bag of Words (CBOW), which predicts a word based on its surrounding context, and Skip-Gram, which reverses this approach by predicting the surrounding context of a word. This dual approach allows Word2Vec to encode semantic and syntactic similarities into the vector space, facilitating the identification of word relationships based on their usage in vast corpora. GloVe, on the other hand, extends the concept of context further by incorporating global statistical information into the embedding process. It constructs a co-occurrence matrix to quantify the frequency at which words appear together across the corpus, thereby integrating both local and global linguistic cues. This method emphasizes both local context (similar to Word2Vec) and global statistical information, resulting in word vectors that capture rich semantic relationships.

Despite the advancements Word2Vec and GloVe represent in capturing semantic richness and accuracy through contextual information, they encounter a fundamental limitation: the representations they generate are static. Once a word vector is learned, it remains unchanged regardless of the context in which the word is used. This leads to the “meaning conflation deficiency,” a phenomenon where the multiple meanings of a polysemous word are conflated into a single vector representation.

This limitation sets the stage for the introduction of contextual embeddings, which aim to overcome the static representation challenge by providing word vectors that adapt based on the sentence context. To address the limitations of static representations, groundbreaking models such as ELMo

(Peters et al., 2018), GPT-1 (Radford et al., 2018), and BERT (Devlin et al., 2018) have been developed, each leveraging unique approaches to contextual understanding and meaning in text.

ELMo advances contextual information utilization through a deep, bidirectional LSTM (Long Short-Term Memory) network. Unlike traditional models that produce a uniform embedding for each word, ELMo creates dynamic embeddings by analyzing text in both forward and backward directions. This bidirectional LSTM architecture enables ELMo to capture the full context surrounding each word, allowing for variable representations based on usage. GPT-1 shifts the paradigm to a transformer architecture, focusing on unsupervised learning to predict the next word in a sequence. GPT-1 uniquely employs a unidirectional (left-to-right) context understanding, generating embeddings based on the preceding text. This approach primarily considers the text leading up to a word, limiting its ability to integrate the full sentence context into each word's representation. When fine-tuned on specific tasks, GPT-1's pre-trained model adapts to a wide range of NLP applications. BERT represents a leap in contextual embedding technology with its bidirectional transformer architecture. Unlike GPT-1, BERT is designed to understand the context of a word based on all of its surroundings (both left and right of the word) simultaneously. This comprehensive context awareness is achieved via masked language modeling, where BERT learns to predict hidden words based on their bidirectional context. BERT's methodology allows for an unprecedented level of understanding, setting new benchmarks for a variety of NLP tasks.

Together, ELMo, GPT, and BERT represent a paradigm shift in text representation. By transitioning from LSTM to transformer architectures and expanding from unidirectional to bidirectional context analysis, these models provide dynamic, context-aware embeddings. This evolution marks significant progress in our capacity to process the complexities of human language,

moving beyond the constraints of static embeddings to a more flexible understanding of text.

The above discussion on the text representation explores the significant advancements in contextual embeddings through models like ELMO, GPT, and BERT. In light of this, our research design leverages OpenAI's publicly available API, specifically the text-embedding-3-large model.[10] This model is employed to process and analyze the resumes of the entire Top Management Team from a single firm, converting the whole team's resume into an embedding vector. An embedding vector, in this context, is a numerical array that represents machine-coded features of the text. Each vector comprises 256 elements, a dimensionality that reflects the embedding size.

Embeddings efficiently capture a wide range of linguistic and contextual patterns, enabling the detection of complex relationships that may be imperceptible to human coders. This approach offers rapid processing, scalability, and cost-efficiency, making it a powerful tool for large-scale text analysis. However, a key limitation of this method is its lack of interpretability. It remains unclear to what extent machine-coded features can extract team-level information compared to human-coded or hybrid features derived from existing literature and domain knowledge. To address this, we will investigate whether embeddings can capture additional insights that human-coded or hybrid methods. Our evaluation will compare the predictive performance of embeddings against human-coded and hybrid features in forecasting high-growth firms.

5. Research Design

5.1 Business fundamentals information Extraction

To evaluate the predictive value of business fundamentals ("the horse") in identifying high-growth firms, we examine methods for extracting business fundamentals-related information. We use a firm's

[10] Like GPT-3.5, this embedding model lacks knowledge of events after September 2021. See https://platform.openai.com/docs/guides/embeddings.

three standard financial statement data to capture its business fundamentals as financial statements are designed to measure a firm's financing, investing, and operating activities. In our setting, firms listed on the NTB are required to disclose at least two years of audited financial statements before listing. These filings contain a rich source of structured data on each firm's business at the time of listing. Extracting business fundamentals information from these financial statements is more straightforward than extracting team-related information because accounting practices have continuously evolved to enhance the reliability and clarity of financial data. In essence, we are building on a well-established foundation—financial statements are carefully curated, high-quality sources of business fundamentals information. This wealth of data allows us to develop and refine methods for extracting and quantifying firm characteristics, or what we refer to as the key attributes of the "horse," based on detailed financial disclosures.

We use two primary methods to analyze financial data: one is theory-driven, and the other is purely data-driven.

Selected Financial Data Method: Since data on entrepreneurial firms is often difficult to obtain, research in this area is relatively scarce. Inspired by Hand (2005) in the context of VC investments, we adapt this approach to extract business fundamentals information. This method identifies key financial metrics of entrepreneurial firms future success: including cash holdings, non-cash assets, long-term debt, annual revenues, R&D expenditures, cost of sales, and SG&A expenses[11]. A detailed definition of these variables is provided in Panel C of Table 1. In Hand (2005), raw numbers undergo a log(1+n) transformation in regression models to account for non-linearity and outliers. However, since we use XGBoost for classification, we apply the raw numbers directly without transformation.

[11] We exclude stock option dilution as it is irrelevant to our setting.

Comprehensive Financial Data Method: This method leverages all principal components from the three financial statements—balance sheet, income statement, and cash flow statement—without prior feature selection. By incorporating the full scope of financial data, this approach enables a holistic analysis of a firm's financial profile.

We evaluate the performance of these methods to determine whether business fundamentals (the 'horse') possess predictive value in identifying high-growth firms. By comparing their effectiveness in forecasting high-growth firms, we assess the effective approach for extracting business fundamentals information from financial disclosures.

5.2 Measuring High-growth Firms

In the context of venture capital, the primary objective is to identify firms with the potential for rapid growth, as these ventures typically yield the highest returns. Therefore, venture capital investors prioritize identifying outliers in the growth distribution. We regard VC investment as an implicit predictive model, wherein VCs are hypothesized to select firms likely to exhibit high growth based on the information available at the time of investment. We measure high firm growth using the sales growth rate within the first three years post-listing.

In the context of venture capital, the objective is to identify firms with the potential for high growth, as these ventures typically yield the highest returns. We regard VC investment as an implicit predictive model, wherein VCs are hypothesized to select firms likely to exhibit high growth based on the information available at the time of investment. We measure high growth using the sales growth rate within the first three years post-listing. We choose a three-year window for two key reasons. First, the dataset spans from 2013 to 2022, making a longer timeframe impractical as it would significantly reduce available observations. Second, three years provide a meaningful forecasting horizon for venture

capitalists seeking to make informed investment decisions.

To align with VC investment strategies, we transform continuous sales growth into a binary variable, distinguishing firms with exceptional growth from the rest. This approach captures the essence of how VCs evaluate high-potential startups. The classification threshold for high growth is set at 200% cumulative sales growth over three years, equivalent to an annualized rate of approximately 45%. This threshold is based on historical growth distributions and aligns with firms in the top 10-15% of the dataset, ensuring that only ventures with truly exceptional growth are classified as high-growth firms.

Our choice of a three-year analysis window is driven by two key considerations. First, the constraints of our dataset, covering the period from 2013 to 2022, limit the feasibility of a longer timeframe, as extending the window would significantly reduce the available data. Second, a meaningful forecasting horizon is essential for generating actionable insights for VC seeking to make informed investment decisions.

Given the venture capital industry's emphasis on identifying high-growth firms, we transform continuous sales growth metrics into a binary variable. This approach better aligns with the investment strategies of VCs, who are primarily interested in distinguishing firms with exceptional growth potential from the rest. By framing growth as a binary outcome, we capture the essence of what constitutes “high growth” among entrepreneurial firms.

In determining an appropriate cutoff for classifying high-growth firms, we initially reviewed growth percentiles within the dataset to identify thresholds aligned with venture capital investors’ investment goals. As mentioned, our approach to dividing training and test data is structured as follows. For the 2018 Test Year, the training data consists of firms listed in 2013 and 2014, while for the 2019 Test Year, the training data includes firms listed in 2013, 2014, and 2015. Cutoff criteria are selected

strictly based on these historical training datasets, thereby preventing any influence from future unseen data and preserving model validity.

The percentile calculation involves computing post-listing 3-year sales growth rates. For firms listed in 2013 and 2014, growth is measured over two overlapping periods (2014-2016 and 2015-2017), pooling all observed growth rates into a single dataset for percentile ranking. Similarly, for firms listed from 2013 to 2015, growth rates measured over the periods 2014-2016, 2015-2017, and 2016-2018 are pooled and ranked. The resultant statistics detailed in Table 2 present three-year growth rates across multiple percentiles. Upon analyzing these distributions and observing the consistency and robustness of growth rates, we choose to adopt a fixed cutoff value of 200%. Our selected threshold of 200% distinctly captures growth above the top 15% yet comfortably aligns between the top 15% and top 10% thresholds. This ensures that only firms exhibiting genuinely exceptional growth are classified as high-growth entities. This fixed cutoff of 200% allows us to focus on a subset of firms most likely to achieve exceptional growth, thereby fulfilling the strategic investment goals of venture capitalists. This threshold approximates an annual growth rate of 45%. Firms exceeding the threshold are classified as "1", indicating high growth, while those below are classified as "0", indicating lower growth levels.

5.3 Sample and Summary Statistics

Table 3 presents the detailed procedures used for sample selection. Our analysis draws from two comprehensive ten-year datasets on NTB firms, both spanning from January 16, 2013, to February 6, 2022. The datasets mark the beginning of NTB in 2013 and continue up to the last day of data collection in 2022. The first dataset, sourced from CSMAR, includes top management team information for 13,020 listed NTB firms. The second dataset, obtained from Choice, comprises initial financial records for 13,019 NTB firms and additional sales data for the same entities.

To measure high-growth firms, we also source the sales data from Choice. Our criteria require each firm to have data extending into the third year beyond their initial listing. This requirement reduces the sample to 12,787 NTB firms covering the period from 2013 to 2019. By integrating the resume, financial, and sales data using the stock ID as a unique identifier, the final analytical sample consists of 12,567 NTB firms.

Our sample consists of 12,567 NTB firms. Of these, 8,647 firms have valid sales data for both the year of listing and the third year post-listing. 3,920 firms lacking third-year post-listing data are treated as low growth[12]. For the 8,647 firms with complete data, sales growth is calculated using the formula

$$(\text{final_sales} - \text{initial_sales})/\text{initial_sales} \times 100\%$$

when initial sales are greater than zero. Sales growth exceeding 200% is classified as high growth, and growth below this threshold is classified as low growth[13].

---

[12] Among the remaining 3,920 firms, 3,823 lack sales records beyond two years post-listing, suggesting they are delisted. To accurately classify these delisted firms, we establish an interim two-year growth threshold of 108%, consistent with our three-year benchmark of 200% growth (approximately 45% annual growth). Despite applying this more lenient interim threshold, only 225 delisted firms exceed the 108% growth rate. Given this minimal number and to preserve classification consistency, we categorize all 3,823 delisted firms as low-growth. The remaining 97 firms introduce further complexity, as they lack sales data specifically for the third year post-listing but possess records for subsequent years. Possible reasons for this anomaly include missing data, or temporary delisting and relisting. After careful examination, we classify these firms as low-growth based on the following considerations. First, 12 firms from this group exceed the interim 108% growth threshold over two years. However, to maintain methodological uniformity across the sample, these firms are also classified as low-growth. Second, 7 additional firms (all listed in 2015) demonstrate sales growth surpassing our annual threshold between their fourth and sixth years post-listing. Categorizing these as high-growth would introduce forward-looking bias, as it would rely on data unavailable at the third year mark. Thus, these firms are classified as low-growth. Lastly, the remaining 78 firms, which do not achieve any predefined growth thresholds within the analyzed timeframes, are naturally categorized as low-growth.

[13] Special cases are handled as follows: 1. Negative Initial Sales: When initial sales are negative, we calculate growth using the method typically employed in firms' annual reports:

$$(\text{final_sales} - \text{initial_sales})/(-\text{initial_sales}) \times 100\%$$

The key aspect of this formula is the denominator, "$-$initial_sales", which ensures that the growth rate is properly interpreted. 2. Zero Initial Sales: Since division by zero is not possible, we manually review these three firms case by case: One firm ("宝利小贷", stock ID: 834340) has incorrect data, which we manually correct using information from its renewed annual report. Another firm reports zero sales consistently throughout the sample period and was classified as low-growth. The third firm

We present summary statistics for the high growth firms in Table 4. Panel A summarizes the annual distribution of firms classified as high growth. Across the entire sample period (2013–2019), approximately 7.1% of the 12,567 firms achieved high growth. However, the proportion of high-growth firms declined noticeably over time, starting from 13.6% in 2013 and reaching a low of 4.7% in 2017, before slightly rebounding to 9.0% in 2019. In Panel B, we compare the financial characteristics of high-growth and remaining (low-growth) firms at listing. Key metrics analyzed include firm size (TOTAL ASSETS) and sales, profitability (ROA and ROE), leverage, past sales growth, tangible assets and cash holdings. Statistical comparisons using t-tests and rank-sum tests reveal significant differences: high-growth firms generally have smaller assets, lower past sales growth, and weaker profitability metrics (ROA). There are also meaningful differences observed in leverage, cash holdings, and asset tangibility. Detailed definitions of variables are provided separately in Panel C of Table 1.

About team information, we present summary statistics for human-coded features in Panel A of Table 5 and for hybrid features in Panel B of Table 5.

For Human-Coded Features in Panel A, first, we examine key individual attributes of firm ultimate individual controllers. On average, controllers are approximately 47 years old, predominantly male (89.3%), and nearly all share the same nationality (99.2%). In terms of education, controllers typically hold around one advanced degree, with a small proportion holding degrees from elite institutions (5.2% from C9 League universities, 12% from 985 universities). Entrepreneurial experience is limited, with about 9.2% of controllers identified as serial entrepreneurs. Controllers possess considerable industry experience, averaging 28.3 years, and prior affiliations with approximately three companies. Only 4.6%

with zero initial sales records sales of approximately 0.15 million yuan in its third year. Based on this, we decide to also classify it as low-growth.

of controllers have a VC background, suggesting that venture capital experience is not a common characteristic among top executives.

We then summarize average member attributes. Team members exhibit similar demographics, with around 71.8% being male and an average age of approximately 42 years. Education profiles indicate most team members hold less than one advanced degree on average, with relatively few from elite institutions. Entrepreneurial experience is infrequent among members, and prior company experience averages around 2.4 companies per member. VC experience among team members is 5.5%, again indicating limited exposure to venture capital networks.

Regarding diversity, the variance in age among team members is high (mean: 76.69), indicating that teams consist of individuals from various age groups. Nationality diversity is very low (0.016), reflecting homogeneous national backgrounds. Education diversity in terms of the number of degrees has a mean variance of 0.449, showing some variation in educational backgrounds. The industry diversity measure is significantly high, suggesting that teams have members with a wide range of years of working experience. The diversity score for prior companies is 0.749, showing moderate variation in past company experiences.

Finally, the faultline measure, based on ASW values, averages approximately 0.553 with a standard deviation of 0.134, suggesting moderate subgroup divisions within management teams based on their attributes.

Panel B presents the summary statistics for hybrid features. While the two team information extraction methods share broad similarities [14], two key differences emerge. First, regarding

[14] Both feature sets share identical definitions for most variables with two notable exceptions.
1)University spinoff attribute: university_controller, and university_average in Human-coded Features refer to university affiliation as whether an individual was previously a student at a university. In contrast, the hybrid features broaden this definition to include whether an individual was previously a student or has ever worked at a university.

entrepreneurial experience, hybrid features identify 84.6% of controllers as serial entrepreneurs, in contrast to the 9.2% from human-coded features. Similarly, 27.1% of team members are classified as serial entrepreneurs by GPT, compared to 3.5% manually. This discrepancy suggests GPT may apply broader criteria to infer entrepreneurial backgrounds, whereas human-coded features likely adhere to stricter definitions. Second, with respect to industry diversity, hybrid features yield a lower variance in work experience years (134.79) compared to human-coded features (167.80), though average experience years remain similar (18.904 vs. 18.254). This implies GPT may smooth outliers, whereas human-coded features could capture more granular variation.

To examine the relationship between these methods, Panel C presents their pairwise correlations. Overall, the results indicate that while manual and hybrid features align closely, two key differences persist. Specifically, entrepreneurial experience and industry diversity display relatively weaker correlations. The relatively low correlations for entrepreneurial experience (0.127 for controllers, 0.259 for average team members) mirror the discrepancies in Panel B's summary statistics, reinforcing the idea that GPT applies broader criteria to classify entrepreneurial backgrounds. Similarly, industry diversity ($r = 0.209$) exhibits lower agreement, suggesting that GPT may smooth outliers more than manual coding.

To further assess the accuracy of these methods, Appendix 3 presents an additional validation using a randomly sampled set of 100 manager resumes. This exercise manually reviews resumes to establish a ground-truth benchmark for comparison, creating four distinct versions of features: (1) human-coded

2)Ethnicity attribute: ethnicity_controller, ethnicity_average, ethnicity_diversity are not in Human-coded Features because ethnicity data is not extracted in these manually coded features.
These differences arise because we refine the measurement framework after completing the manual extraction. However, incorporating these updates into the manual human coding process is impractical, as it is prohibitively slow and costly to repeatedly ask research assistants to redo the work. By contrast, the GPT-based method dynamically adapts to these refinements, ensuring consistency with the updated framework while maintaining efficiency and scalability.

features, (2) hybrid features, (3) Ground truth for human-coded features, and (4) Ground truth for hybrid features. While both methods align well with their respective ground truth, certain attributes including entrepreneurial experience exhibit weaker correlations. These discrepancies further support our previous findings. This limitation likely reflects the inherent constraints of GPT-3.5's capabilities. As previously discussed, we employ GPT-3.5 to minimize potential look-ahead biases. Given the improvements observed in later LLMs, we leave a more comprehensive evaluation of advanced models for future research.

Due to the complex nature and technical derivation of the Machine-coded Features, they are not easily interpretable by humans. Thus, they are not summarized in the summary statistics.

In addition to team information, we employ two primary methods to extract business fundamentals information: Selected Financial Data Method and Comprehensive Financial Data Method. Due to the extensive number of variables in the second method, we present only the summary statistics for Selected Financial Data Method in Table 6. It presents key financial indicators from the Selected Financial Data Method for 2013–2019, including cash holdings, non-cash assets, long-term debt, revenues, R&D expenditures, cost of sales, and SG&A expenses. These statistics highlight substantial heterogeneity in firms' financial structures, reflecting variations in liquidity, leverage, revenue generation, and operational expenses.

5.4 Machine Learning Prediction

We investigate the value of team characteristics, business fundamentals, and their combination in predicting high-growth firms. Specifically, we aim to identify NTB firms likely to achieve high post-listing growth using information available at the time of listing. By restricting our predictors to pre-listing data, we align with the decision-making context of venture capitalists, who assess a firm's growth

potential based on information available at the time of initial public disclosures. Post-listing information, such as trading data, is excluded to avoid look-ahead bias and to ensure the model reflects the real-world constraints of early-stage investment decisions. While incorporating post-listing data could improve prediction accuracy over time, such models are less valuable to venture capitalists because NTB firm valuations will likely rise as uncertainty diminishes, reducing the potential for outsized investment returns.

Our objective is to predict whether a firm will achieve three-year sales growth exceeding 200%. To address the potential interdependence between team characteristics and business fundamentals, we employ XGBoost (T. Chen & Guestrin, 2016), a tree-based ensemble method known for capturing nonlinear relationships, interaction effects among predictors, and handling missing values effectively. By integrating both dimensions within a single model, XGBoost allows for flexible modeling of any latent relationships between team characteristics and business fundamentals. We evaluate the out-of-sample performance of various models using team and business fundamentals information as inputs.

Our approach to dividing training and test data is structured[15] as follows:

- 2018 Test Year: Training data consists of firms listed in 2013 and 2014.
- 2019 Test Year: Training data consists of firms listed in 2013, 2014, and 2015.

For each test year, we randomly allocate 20% of the training sample as a validation set. The purpose is to optimize the selection of hyperparameters for our machine learning models. After determining the optimal hyperparameters, we combine the training and validation samples to train the final model configurations. We adopt cross-sectional validation instead of temporal splitting due to the varying

[15] We exclude 2017 as a test year because, in the summary statistics for the dependent variable, only 153 firms were listed in 2013, compared to 2,066 firms in 2017. Selecting 2017 as the test year could introduce potential biases and inconsistencies due to the limited scope of training data. To mitigate this issue, we focus on 2018 and 2019 for testing.

number of firms listed each year. Given the limited training period (only 2–3 years), a strict temporal split would be impractical. Table 7 presents the hyperparameters and their candidate values. The out-of-sample performance of various models, using team and business fundamentals information as inputs, is assessed on the testing sample.

Building on the methodology of Bao et al. (2020), we evaluate our model's ranking performance using NDCG@K (Normalized Discounted Cumulative Gain at rank K). To provide a more comprehensive assessment, we also incorporate PRECISION@K and RECALL@K as supplementary evaluation metrics.

NDCG@K measures how well the model prioritizes true positive cases within the top K predictions. In our setting, this means assessing whether the firms our model ranks highest are indeed those that achieve three-year post-listing sales growth exceeding 200%.

To illustrate, suppose our test set includes 100 NTB firms, and 20 of them actually surpass the 200% sales growth threshold. If our model ranks these 20 high-growth firms at the very top, NDCG@K will be closer to 1, indicating strong predictive performance. Conversely, if our model ranks these firms lower in the list, NDCG@K decreases, reflecting weaker predictive accuracy. The range of NDCG@K is between 0 and 1, where:

NDCG@K = 1 means all high-growth firms are correctly ranked at the top.

NDCG@K = 0 means none of the top-ranked firms achieved the 200% growth threshold.

To ensure meaningful comparisons, we set K to the actual number of NTB firms in each test set that achieved post-listing three-year sales growth exceeding 200%. This approach mirrors real-world conditions.

Alongside NDCG@K, we compute PRECISION@K and RECALL@K, which provide additional

insights into prediction quality: PRECISION@K measures how many of the top K firms identified by the model actually achieve high growth. RECALL@K measures how many of all high-growth firms are captured within the top K predictions. Since K is set to match the number of high-growth firms in each test sample, PRECISION@K and RECALL@K yield identical values in our case.

Using performance evaluation metrics, we investigate the value of team characteristics, business fundamentals, and their combination for predicting high-growth firms.

For the first research question, we investigate the predictive value of team characteristics ("the jockey") in identifying high-growth firms. To evaluate this, we examine multiple methods for extracting features from top management team (TMT) resumes and assess their effectiveness in forecasting high-growth firms. Our investigation focuses on two parts. First, we compare the effectiveness of GPT-generated hybrid features against manual human coding. While manual coding is highly accurate, it is labor-intensive and costly, whereas GPT-generated features are faster and more cost-effective. We evaluate whether hybrid features can serve as an alternative by assessing their predictive performance. Second, we examine whether embedding can capture team information that human-coded and hybrid methods might overlook. The predictive performance of embeddings is compared against that of human-coded and hybrid features to determine their relative strengths. Finally, we use the performance results from these methods to assess whether team characteristics (the "jockey") hold predictive value in identifying high-growth firms.

For the second research question, we assess the value of business fundamentals information in predicting high-growth firms. We use a firm's three standard financial statement data to capture its business fundamentals as financial statements are designed to measure a firm's, financing, investing, and operating activities. To assess predictive performance, we compare a theory-driven Selected

Financial Data Method with a data-driven Comprehensive Financial Data Method. The theory-driven approach focuses on key financial indicators that have been traditionally recognized as important in entrepreneurial firm literature, whereas the data-driven method incorporates a broader range of financial data points, allowing machine learning models to detect patterns that may not be explicitly identified in theory-driven approaches.

For our third research question, we investigate whether integrating team characteristics with business fundamentals can enhance the accuracy of models predicting high-growth firms. To explore this integration, we implement early and late fusion methods. Early fusion combines team and business fundamentals features at the input level, creating a unified dataset before model training to capture potential synergies between the two data sources. Late fusion, on the other hand, develops separate models for team and business fundamentals information and merges their outputs, allowing each predictor to contribute independently before final aggregation.

6. Jockey vs. Horse

Table 8 presents the out-of-sample performance metrics for predicting whether a firm will achieve three-year sales growth exceeding 200%. Panel A reports the overall average model performance across all test years. Panel B and Panel C provide model performance metrics for the years 2018 and 2019, respectively. We evaluate each model using NDCG@K, along with PRECISION@K and RECALL@K, which measure precision and recall at the top K positions in the ranking list. To provide further context, we also include the proportion of high-growth firms in each test sample (average DV) within the table.

For the first research question, we investigate the predictive value of team characteristics ("the jockey") in identifying high-growth firms. To evaluate this, we examine multiple methods for extracting features from top management team (TMT) resumes and assess their effectiveness in forecasting high-

growth firms.

First, we compare the effectiveness of hybrid features against manual human coding. In Panel A of Table 8, the "hybrid features" model demonstrates superior performance with an NDCG score of 0.309 on average, compared to the "Human-Coded Features" model at 0.245 on average. This result suggests that the GPT model not only processes data more efficiently but also maintains a high level of accuracy in predicting high-growth firms.

Second, we examine whether embedding can capture team information that human-coded and hybrid methods might overlook. In Panel A of Table 8, the "Machine-Coded Features" model, based on an embedding model, shows an average NDCG score of 0.205, which is lower than both the hybrid (0.309) and human-coded features (0.245). These results suggest that while embeddings capture a wide range of information, they may not focus effectively on the most relevant features for high growth prediction as effectively as models incorporating human expertise.

Finally, our results highlight that team characteristics hold substantial predictive value in identifying high-growth firms, particularly when structured effectively. GPT emerge as a strong alternative to manual coding, offering comparable accuracy while significantly improving efficiency. Meanwhile, the inferior performance of text embeddings suggests that predictive models benefit from feature extraction informed by domain knowledge rather than relying solely on automated deep-learning representations.

For the second research question, we examine the predictive value of business fundamentals in identifying high-growth firms. To assess this, we evaluate two different approaches: the theory-driven Selected Financial Data Method and the data-driven Comprehensive Financial Data Method.

In Panel A of Table 8, both business fundamentals feature sets, "Selected Financial Data" and

“Comprehensive Financial Data”, demonstrate predictive power, achieving average NDCG scores of 0.239 and 0.185, respectively. These findings confirm that business fundamentals hold meaningful predictive value in distinguishing high-growth firms. The higher NDCG score for 'Selected Financial Data' suggests that a targeted, theory-driven selection of financial indicators can be as effective, if not superior, to a purely data-driven approach that incorporates a broader set of financial variables. This comparison indicates that processing business fundamentals information through a theoretically guided approach may enhance predictive performance, leading to more accurate identification of high-growth firms. However, the model based on business fundamentals performs worse than the one based on team characteristics.

For our third research question, we examine whether combining team characteristics with business fundamentals can enhance the accuracy of models predicting high-growth firms.

We combine the most effective team information extraction methods (human-coded features and hybrid features) with the most effective business fundamentals information extraction method (Selected Financial Data). In Panel A of Table 8, the optimal integration of Human-Coded Features and Selected Financial Data yields an NDCG score of 0.305. For comparison, hybrid features alone achieve an average NDCG score of 0.309, human-coded features score 0.245, and Selected Financial Data alone records an average NDCG score of 0.239.

These results suggest that while team information provides substantial additional predictive value when combined with business fundamentals information, the reverse is not as significant, business fundamentals information offers only limited incremental relevance once team information is already considered. These results collectively reinforces the idea that team information plays a more significant role than business fundamentals information in predicting high-growth firms.

To ensure the robustness of our results, we conduct additional analyses. First, we recognize the potential concern of hindsight bias in choosing K based on the number of NTB firms in each test sample. To mitigate this, we also determine K by calculating the proportion of NTB firms in each training sample and multiplying it by the size of the test sample. The resulting percentages of the test samples selected for the top K are 11.2% for 2018, and 9.2% for 2019. The performance metrics, including NDCG@K, PRECISION@K, and RECALL@K, based on this adjusted K are reported in Table 9.

The revised evaluation confirms our initial findings. Team characteristics hold substantial predictive value in identifying high-growth firms. Among different extraction methods, human-coded and hybrid features remain the most effective for predicting high-growth entrepreneurial firms. Hybrid features achieve an NDCG score of 0.317, outperforming human-coded features (NDCG: 0.276) and machine-coded features (NDCG: 0.212). For business fundamentals information, it holds meaningful predictive value in identifying high-growth firms. Between the two extraction methods, the theory-driven Selected Financial Data Method (NDCG: 0.258) outperforms the data-driven Comprehensive Financial Data Method (NDCG: 0.193), highlighting the advantage of a targeted, theoretically guided approach. The results continue to demonstrate that team information plays a more significant role than business fundamentals information in predicting high-growth firms. Finally, we assess whether integrating team and business fundamentals information enhances predictive accuracy. The optimal combination of Human-Coded Features and Selected Financial Data yields an NDCG score of 0.336, compared to the individual performances of hybrid features alone (0.317), human-coded features (0.276), and Selected Financial Data alone (0.258). These results confirm that team information provides substantial additional predictive value when combined with business fundamentals information, while business fundamentals information alone offers only limited incremental relevance

once team information is already considered.

Second, a firm's geographic location and industry classification represent key background knowledge that investors typically consider when making investment decisions. As a robustness check, we evaluate whether our findings remain consistent when incorporating these factors. Specifically, we integrate geographic location and industry classification [16] with the existing team and business fundamentals information feature sets and assess predictive performance. The performance metrics, including NDCG@K, PRECISION@K, and RECALL@K, for this expanded feature set are reported in Table 10.

While adding background knowledge does not improve predictions, the findings remain consistent with our initial conclusions. Team characteristics hold substantial predictive value in identifying high-growth firms. Among different extraction methods, human-coded and hybrid features remain the most effective for predicting high-growth entrepreneurial firms. Hybrid features achieve an NDCG score of 0.346, outperforming human-coded features (NDCG: 0.299) and machine-coded features (NDCG: 0.205). For business fundamentals information, it holds meaningful predictive value in identifying high-growth firms. Between the two extraction methods, the theory-driven Selected Financial Data Method (NDCG: 0.171) outperforms the data-driven Comprehensive Financial Data Method (NDCG: 0.164), highlighting the advantage of a targeted, theoretically guided approach. The results continue to demonstrate that team information plays a more significant role than business fundamentals information in predicting high-growth firms. Finally, we assess whether integrating team and business fundamentals information enhances predictive accuracy. The optimal combination of hybrid features and Selected

[16] Geographic location is measured by the province and city where the firm is registered. Industry classification is measured based on the sub-category classification as defined in the *Guidelines for Industry Classification of Listed Companies (Revised 2012)*.

Financial Data yields an NDCG score of 0.277, compared to the individual performances of hybrid features alone (0.299), human-coded features (0.194), and Selected Financial Data alone (0.171). These results confirm that team information provides substantial additional predictive value when combined with business fundamentals information, while business fundamentals information alone offers only limited incremental relevance once team information is already considered.

In summary, our findings reveal three key insights. First, team characteristics hold substantial predictive value in identifying high-growth firms. Among different extraction methods, human-coded features outperform machine-coded features in extracting predictive team information for high-growth firms. GPT-generated features achieves predictive accuracy comparable to manual human coding while significantly improving efficiency.

Second, for business fundamentals information, it holds meaningful predictive value in identifying high-growth firms. Between the two extraction methods, a theory-driven approach (Selected Financial Data Method) focusing on targeted financial indicators outperforms data-driven methods using broader variables. This suggests that theoretical insights enhance the quality of business fundamentals feature selection for high-growth firms prediction.

Last, team ("jockey") information proves more critical than business fundamentals("horse") information in predicting high-growth firms. While combining both types of information improves predictive accuracy, team information provides significant incremental value over business fundamentals information, whereas business fundamentals information contributes minimally once team information is included.

These findings emphasize that while both the jockey (team characteristics) and the horse (business fundamentals) matter, the jockey plays a more important role in predicting high-growth firms.

7. Feature Importance Analysis

After evaluating the predictive performance of our machine learning models, we conduct a model interpretation analysis to identify the most influential features driving the predictions. To examine the importance of team characteristics in identifying high-growth firms, we utilize Shapley Additive exPlanations (SHAP) values to quantify the contribution of each predictor in our best-performing model using hybrid features. SHAP values enable an interpretable decomposition of model outputs, offering insights into the relative influence of individual predictors (Lundberg et al., 2020; Lundberg & Lee, 2017; Shapley, 1953).

Table 11 presents a comparison of the top 10 most influential independent variables across feature sets for the years 2018 and 2019 in Panel A. In Panel B, we assess the collective importance of predictors by grouping features according to the theoretical framework established in Section 4.1. The framework categorizes predictors into compositional features (which include average member attributes, diversity measures, and faultline strength) and structural features.

Our SHAP value analysis highlights the dominant role of compositional features in predicting high-growth firms (Table 11, Panel A). Compositional features, which aggregate average member attributes, diversity measures, and faultline strength, accounted for 85.40% and 90.64% of the total predictive importance in 2018 and 2019, respectively. Among the compositional components, average member attributes contributed the most, explaining 44.57% of model predictions in 2018 and 56.63% in 2019, followed by diversity measures and, to a much lesser extent, faultline strength. In contrast, structural features accounted for 14.60% and 9.36% of the predictive contribution in 2018 and 2019, respectively.

Panel B further examines the top input features individually. The rankings demonstrate substantial

stability over time: industry experience (*industry_average*) remained the most important predictor across both 2018 and 2019. Other consistently influential features include *prior_company_diversity*, *age_diversity*, and *age_average*, all of which are compositional attributes. Structural features, such as the age and industry experience of controlling individuals, ranked among the top predictors in 2018 but declined in relative importance in 2019. Overall, these results underscore the dominant predictive power of team compositional characteristics over structural attributes and highlight the particular importance of relevant industry background, prior company experience, and team age as key predictors of high-growth outcomes.

8. Benchmarking Against VC

8.1 VC Decision-Making Effectiveness

Our findings indicate that team information ("jockey") outweighs business fundamentals information ("horse") in predicting high-growth firms. This aligns with prior survey and experimental research suggesting that VCs prioritize the team over the business fundamentals information when evaluating startups (Bernstein et al., 2017; Gompers et al., 2020). However, it remains an open question whether VCs fully recognize the predictive value of team characteristics in identifying high-growth firms.

While VCs assert the importance of team characteristics, research on persistent biases in decision-making, such as those related to gender (Ewens & Townsend, 2020; Howell & Nanda, 2024; Kanze et al., 2018) and race (Fairlie et al., 2022), suggests that investors may not always extract and apply information in an optimal manner. Moreover, this issue is not unique to VCs. In broader financial markets, investors also tend to underweight important yet assessable information. For example, Cohen et al. (2013) find that R&D capabilities, despite being relatively straightforward to evaluate, are

frequently overlooked in valuation processes. This raises concerns about whether VCs similarly underutilize key signals when assessing startups. Valuing team characteristics does not necessarily mean VCs can evaluate them effectively.

To explore this, we benchmark our machine learning model's predictive performance against real-world VC investment decisions. Our models, constrained to publicly available data on team characteristics and business fundamentals, capture only a fraction of the information accessible to VCs at the time of investing. If, despite these limitations, our model outperforms VCs in identifying high-growth firms, it would suggest that investors are not fully leveraging the predictive signals in their decision-making.

We regard VC investment as an implicit predictive model, wherein VCs are hypothesized to select firms likely to exhibit high growth based on the information available at the time of investment. This assumption is grounded in the venture capital objective of identifying firms with potential for rapid sales growth, which are expected to generate the highest returns.

Thus, VC-backed firms can be seen as a reflection of VC "predictions" of high sales growth. By comparing the growth outcomes of firms predicted by our model with those of VC-backed firms, we evaluate the relative performance of our machine learning model against VCs.

We first identify VC in the shareholder list of NTB firms. Since NTB firms disclose their top 10 shareholders upon listing but do not explicitly label them as VC firms, we rely on external data sources for verification. Specifically, we obtain a list of VCs from the Asset Management Association of China (AMAC) and PEDATA, focusing on those established before 2020. Additionally, we use Qichacha (QCC), a business registration information database, to verify shareholder names and assess their

industrial and commercial registration details to determine whether they qualify as VC firms[17]. Using official data from AMAC and PEDATA, we identify 5,600 VC investors for NTB firms. Meanwhile, QCC data identifies 7,087 VC investors[18].

We identify NTB firms with VC backing at the time of listing. To ensure a fair comparison with our model using the listing information, we wish to limit NTB firms with VC backing to those with new VC investments within the two years leading up to their listing date.

The timing of VC investments is determined using two main data sources: prospectus data, and industrial and commercial registration records from RESSET. The first source, prospectus data, is manually collected from NTB firms' prospectuses, which disclose historical investors and their investment dates. These data cover NTB firms listed before 2019. The second source, industrial and commercial registration records from RESSET, compiles change registration filings that firms are required to submit under the *Regulations of the People's Republic of China on Company Registration Management* whenever there are modifications to key registration details. RESSET records include the company name, change item, change date, and shareholder information before and after the change[19].

From the prospectus data, we first identify the earliest investment date for each investor in a given NTB firm. Next, we determine whether the investor qualifies as a VC. We then select NTB firms that received new VC investments within the two years preceding their listing date.

From RESSET registration data, we identify VC investments by analyzing shareholder change

[17] Using QCC, we verify shareholder names with suffixes indicating corporate entities and partnerships. We examine their industrial and commercial registration records and classify them as VC if their industry classification and business scope include keywords such as "investment," "venture capital," "securities," "finance," "asset management," or "fund." However, we do not rely solely on these keywords in the shareholder's name to avoid mistakenly identifying market makers or brokers as VC firms.

[18] We also use the VC lists from AMAC and PEDATA to conduct additional robustness checks in our analysis. The untabulated results show similar findings.

[19] RESSET tracks 25 shareholder-related change items, including investor and shareholder modifications, equity transfers, and changes in responsible persons. Examples of key change items include "Investor (equity) change," "Shareholder change," "Equity transfer," and "Responsible person change."

records. Specifically, we examine cases where a VC appears in the shareholder list after a change but was not present before, treating this as a new VC investment in the NTB firm. The corresponding change date is recorded as the VC investment date. We then verify and extract the first recorded VC investment date for each NTB firm, applying the same two-year investment window before listing to ensure consistency.

Finally, we merge the NTB firms identified from both sources to construct the final dataset. Using this approach, we identify 3,936 NTB firms that received new VC backing within two years preceding their listing date.

Table 12 presents the annual distribution of NTB firms alongside the percentage of these firms that received VCs' investment. For each year under review, we calculate the average value of the 3-year high-growth dummy variable for NTB firms with VC backing. This dummy variable is assigned a value of 1 if a firm achieves 200% or more growth within three years post-listing. The mean value of this growth dummy for VC-backed firms quantifies the precision of the VC investment, demonstrating their ability to identify high-growth prospects. Additionally, we assess the recall rate of VC investments. This measure is defined as the ratio of VC-backed firms that achieve high growth to the total number of high-growth firms identified each year.

Focusing on test year 2018 and 2019, the number and proportion of VC-backed NTB firms decline, from 20.5% in 2018 to 18.1% in 2019. However, VC investment precision improves, rising from 7.1% to 20%, suggesting greater selectivity and better identification of high-growth firms. The recall rate also increases from 20% to 40%, indicating that a larger share of high-growth firms receive VC investment in 2019.

To facilitate comparison, we denote the number of NTB firms with VC investments as N for each

year. For each predictive model detailed in Table 8 of Section 6, which features different sets of information, we calculate both PRECISION@N and RECALL@N. This denotes that we are considering the top N firms in the ranking generated by the predictive machine learning models as the predicted high-growth firms.

As shown in Table 13, our ML models generally demonstrate strong predictive ability, with precision comparable to or exceeding that of VC investment benchmark. On average, the precision of VC investment is 0.136, while models leveraging team information alone achieve higher precision, reaching 0.171 for human-coded features and 0.164 for hybrid features. Similarly, the recall rate of VC investment is 0.300, whereas team-based models achieve 0.388, indicating that they capture a larger share of actual high-growth firms.

Integrating business fundamentals and team information further enhances predictive performance. The late fusion approach, which combines team and business fundamentals information after independent processing, achieves the highest precision of 0.190 and recall of 0.450.

These results indicate that our ML models, particularly those incorporating team information, outperform VC investors in identifying high-growth firms[20]. This suggests that VCs may not be fully utilizing the most predictive signals available to them. Furthermore, our findings expose a disconnect between VCs' stated emphasis on team characteristics and their ability to evaluate them effectively in practice. Ultimately, recognizing the importance of teams does not necessarily translate into accurately assessing their potential, underscoring inefficiencies in VC decision-making.

[20] We conduct a robustness check to assess whether our findings hold when incorporating a firm's geographic location and industry classification, similar to the approach in the previous section. As shown in Table 14, ML models, particularly those leveraging team information, continue to outperform VCs in identifying high-growth firms, confirming the robustness of our results.

8.2 Alternative Explanations for VC Underperformance

One potential concern regarding the relatively poor performance of VC investment compared to our ML model is that future high sales growth may not be a primary factor influencing VC investment decisions. However, we provide three key points to mitigate this concern.

First, prior research (Chemmanur et al., 2011; D. Guo & Jiang, 2013) demonstrates that VCs in both the U.S. and China are more likely to invest in firms with high sales growth. These studies also highlight that VCs play an active role in enhancing a firm's sales growth after investment. Thus, sales growth is a crucial factor in VC decision-making across different stages of their involvement, from initial investment to post-investment firm development.

Second, we analyze new VC investments during the three-year post-listing period for all firms in the test years (2018 and 2019). If VCs prioritize firms with strong post-listing growth, they should be more likely to invest in firms demonstrating high performance after listing. To test this hypothesis, we conduct an in-sample regression, where the dependent variable (VC_3 = 1 or 0) indicates whether a firm received new VC investment within three years post-listing, and the independent variable (High_growth = 1 or 0) captures whether the firm achieved 200% or more sales growth during the same period.

We draw on prior literature to construct control variables for our regression analysis. These constructs are consistent with the input variables used in our machine learning prediction. They fall into two main categories: team characteristics and business fundamentals. For team characteristics, we include ten key attributes commonly examined in the literature: gender (Sharchilev et al., 2018), age (Esen et al., 2023), ethnicity and nationality (Calder-Wang et al., 2021), educational background (Arroyo et al., 2019), entrepreneurial and industry experience (Te et al., 2022), prior company (Beckman et al., 2007), university spin-offs, and VC connections (Ortín-Ángel & Vendrell-Herrero,

2010). We operationalize these characteristics using established measures from earlier studies: *gender_average1* (gender), *age_average* (age), *ethnicity_diversity* (ethnicity), *nationality_diversity* (nationality), *education_average1* (education), *entrepreneurial_average* (entrepreneurial experience), *industry_average* (industry experience), *prior_company_average* (prior company affiliation), *university_average* (university spin-off), and *VC_average* (VC connections). A detailed definition of these variables is provided in Panel A of Table 1.

For business fundamentals, we adopt the approach of Hand (2005) and include a set of financial variables: cash holdings, non-cash assets, long-term debt, annual revenue, R&D expenditures, cost of sales, and SG&A expenses. Following Hand (2005), we transform these variables using the log (x + 1) function. A detailed definition of these variables is presented in Panel C of Table 1.

We include year fixed effects to control for time-specific factors that may influence investment decisions. All control variables are standardized prior to analysis. To maintain a representative sample, we exclude *ethnicity diversity*, *long-term debt*, and *R&D expenditures* due to their substantial missing values—578, 374, and 167 observations, respectively. Including these variables would reduce the sample size to just 84 observations, resulting in a significant loss of observation. Our dataset comprises 768 firm-year observations (547 from 2018 and 221 from 2019), and after excluding observations with missing values on select variables, the final sample consists of 737 firms.[21] Despite some relatively high pairwise correlations between probability and financial fundamentals[22], Variance Inflation Factors (VIFs) for all regressors remain below the conventional threshold of 10 (in fact, all are below 5). This indicates that multicollinearity is not a significant concern and does not compromise the stability or

[21] Missing data are identified in cost of sales (23 observations), SG&A expenses (23), gender average (2), age average (2), nationality diversity (2), entrepreneurial average (2), industry average (2), university average (2), and non-cash assets (1), totaling 31 dropped observations.

[22] The correlation matrix is untabulated to maintain clarity, as the inclusion of all variables would result in an overly complex and unwieldy table.

interpretability of the regression estimates. Accordingly, we retain all theoretically motivated variables in the model specification.

The regression equation is as follows:

$$VC_3 = \beta_0 + \beta_1 \cdot High_growth + \beta_2 \cdot control + \varepsilon \qquad (1)$$

The results, reported in regression (1) of Table 15, show that the estimated coefficient ($\hat{\beta}_1 = 0.2426$) is positive and statistically significant. This finding supports our hypothesis that VCs are more likely to invest in firms with high post-listing growth, consistent with our prediction.

Third, we conduct a stricter evaluation by testing whether the predicted probabilities from our ML models can explain post-listing VC investment, incremental to the control variables constructed based on prior literature. This test allows us to assess whether VCs invest in firms with high-growth potential, as identified by our models, once more firm information becomes publicly available after listing. To do so, we perform an out-of-sample regression, where the dependent variable (VC_3 = 1 or 0) again indicates whether a firm received new VC investment within three years post-listing, while the independent variable ($prob_X$) represents the probability predicted by the ML models using different information sets, as outlined in Table 13. We use the same set of control variables as in the regression (1). The regression equation is as follows:

$$VC_3 = \beta_0 + \beta_1 \cdot prob + \beta_2 \cdot control + \varepsilon \qquad (2)$$

The results are reported in Table 15. The result from regression (2) to (9) shows that $\hat{\beta}_1$ is positive and statistically significant in most models. The only exception occurs when the independent variable is the probability predicted by the ML models based on Comprehensive Financial Data, where $\hat{\beta}_1$ is negative but statistically insignificant. These results indicate that, in general, the predicted probabilities from our models are positively associated with the likelihood of post-listing VC investment.

These findings offer additional insight. While VCs frequently claim to prioritize team characteristics, their pre-listing investments do not fully capture the most predictive team-related signals, which contributes to their underperformance relative to our ML models. The positive association between our model's predictions and post-listing VC investments suggests that VCs may be more likely to invest in firms with high-growth potential once more information is available. However, this pattern also highlights a potential inefficiency: better evaluation of team characteristics at the time of listing could lead to earlier and possibly more effective investment decisions.

Taken together, our combined evidence suggests that VCs do consider future high sales growth when making investment decisions. However, they are less effective at identifying high-growth firms when information is limited, especially before listing.

These findings help mitigate concerns that the poor performance of VC investments relative to our ML models is driven by high sales growth being irrelevant to VC decision-making. On the contrary, the evidence shows that VCs care about high-growth, but face challenges in assessing it accurately at early stages. The post-listing results support this view, showing that VCs may respond to clearer signals of high-growth, but often after key opportunities for earlier investment have passed.

In summary, by comparing our model's predictive accuracy with that of VCs, we demonstrate that our ML models, particularly those incorporating team information, outperform VCs in identifying high-growth firms. This suggests that VCs may not be fully leveraging the most predictive signals available to them. Moreover, our findings reveal a critical gap between VCs' stated emphasis on team characteristics and their ability to assess them effectively in practice. Ultimately, "valuing teams" does not necessarily equate to "evaluating teams effectively," underscoring inefficiencies in VC decision-making.

9. Conclusion

This study uses entrepreneurial firm data from China New Third Board (NTB) and machine learning to evaluate the value of team characteristics ("jockey") and business fundamentals ("horse") in predicting high-growth firms. By examining different extraction methods, we assess the individual and joint roles of these factors. Our findings contribute to the "jockey vs. horse" debate and reveal inefficiencies in VC investment decision-making.

First, we demonstrate that team characteristics hold substantial predictive power in identifying high-growth firms. Among different extraction methods, human-coded features outperform machine-coded features in capturing predictive team information. Notably, GPT-generated features achieve comparable accuracy to manual human coding while significantly enhancing efficiency, offering a scalable and cost-effective alternative.

Second, business fundamentals also predict high-growth firms. Between the two extraction methods, a theory-driven approach (Selected Financial Data Method) focusing on targeted financial indicators outperforms a data-driven method using broader variables. This suggests that theoretical insights improve the quality of business fundamentals feature selection in high-growth firm prediction.

Most critically, our results show that team characteristics are stronger predictors of high growth than business fundamentals. While combining both types of information improves predictive accuracy, business fundamentals add little additional value once team characteristics are included. This underscores the dominant role of the "jockey" over the "horse" in predicting high-growth firms.

Finally, our study highlights inefficiencies in VC investment decision-making. VCs claim to prioritize team characteristics, yet our best machine learning models, relying only on publicly available data, outperform real-world VCs in predicting high-growth firms. This suggests that despite recognizing

the importance of team characteristics, VCs may not fully leverage the most predictive signals in their evaluations. Simply “valuing teams” does not equate to “evaluating teams effectively”.

References

Arif, S., Bushman, R., Donovan, J., & Gopalan, Y. (2025). *Managerial capital and soft information implications for performance and CEO turnover*.

Armstrong, D. M. (2023). Measuring Tax Enforcement with Generative AI. *Available at SSRN 4639565*.

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. https://doi.org/10.1109/ACCESS.2019.2938659

Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach. *Journal of Accounting Research*, *58*(1), 199–235. https://doi.org/10.1111/1475-679X.12292

Beckman, C. M., Burton, M. D., & O'Reilly, C. (2007). Early teams: The impact of team demography on VC financing and going public. *Journal of Business Venturing*, *22*(2), 147–173. https://doi.org/10.1016/j.jbusvent.2006.02.001

Bernstein, S., Giroud, X., & Townsend, R. R. (2016). The impact of venture capital monitoring. *The Journal of Finance*, *71*(4), 1591–1622.

Bernstein, S., Korteweg, A., & Laws, K. (2017). Attracting early-stage investors: Evidence from a randomized field experiment. *The Journal of Finance*, *72*(2), 509–538.

Block, J., Fisch, C., Vismara, S., & Andres, R. (2019). Private equity investment criteria: An experimental conjoint analysis of venture capital, business angels, and family offices. *Journal of Corporate Finance*, *58*, 329–352. https://doi.org/10.1016/j.jcorpfin.2019.05.009

Block, J. H., Hirschmann, M., & Fisch, C. (2021). Which criteria matter when impact investors screen social enterprises? *Journal of Corporate Finance*, *66*, 101813. https://doi.org/10.1016/j.jcorpfin.2020.101813

Calder-Wang, S., Gompers, P. A., & Huang, K. (2021). *Diversity and performance in entrepreneurial teams*. National Bureau of Economic Research.

Chemmanur, T. J., Krishnan, K., & Nandy, D. K. (2011). How does venture capital financing improve efficiency in private firms? A look beneath the surface. *The Review of Financial Studies*, *24*(12), 4037–4090.

Chen, C., Ke, B., & Zhao, Q. (2024). Enhancing Firm Quality Measurement Using Machine Learning: The Importance of Theory-motivated Data Engineering. *Available at SSRN 4434498*.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Y., Huang, J., Li, T., & Pittman, J. (2022). It's a Small World: The Importance of Social Connections with Auditors to Mutual Fund Managers' Portfolio Decisions. *Journal of Accounting Research*, *60*(3), 901–963. https://doi.org/10.1111/1475-679X.12395

Cohen, L., Diether, K., & Malloy, C. (2013). Misvaluing Innovation. *The Review of Financial Studies*, *26*(3), 635–666. https://doi.org/10.1093/rfs/hhs183

Davenport, D. (2022). Predictably Bad Investments: Evidence from Venture Capitalists. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4135861

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv Preprint arXiv:1810.04805*.

Esen, T., Dahl, M. S., & Sorenson, O. (2023). Jockeys, horses or teams? The selection of startups by

venture capitalists. *Journal of Business Venturing Insights*, *19*, e00383. https://doi.org/10.1016/j.jbvi.2023.e00383

Ewens, M., & Townsend, R. R. (2020). Are early stage investors biased against women? *Journal of Financial Economics*, *135*(3), 653–677. https://doi.org/10.1016/j.jfineco.2019.07.002

Fairlie, R., Robb, A., & Robinson, D. T. (2022). Black and white: Access to capital among minority-owned start-ups. *Management Science*, *68*(4), 2377–2400.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, *57*(3), 535–574. https://doi.org/10.1257/jel.20181020

Gompers, P. A., Gornall, W., Kaplan, S. N., & Strebulaev, I. A. (2020). How do venture capitalists make decisions? *Journal of Financial Economics*, *135*(1), 169–190.

Gow, I. D., Larcker, D. F., & Zakolyukina, A. A. (2023). How important is corporate governance? Evidence from machine learning. *Chicago Booth Research Paper*, *22–16*, 2022–2137.

Guo, D., & Jiang, K. (2013). Venture capital investment and the performance of entrepreneurial firms: Evidence from China. *Journal of Corporate Finance*, *22*, 375–395.

Guo, K., Ke, B., & Su, Y. (2025). *The Predictive Value of Causal Theories in Detecting Informed Insider Trading*.

Hall, R. E., & Woodward, S. E. (2010). The Burden of the Nondiversifiable Risk of Entrepreneurship. *American Economic Review*, *100*(3), 1163–1194. https://doi.org/10.1257/aer.100.3.1163

Hand, J. R. M. (2005). The Value Relevance of Financial Statements in the Venture Capital Market. *The Accounting Review*, *80*(2), 613–648. https://doi.org/10.2308/accr.2005.80.2.613

Howell, S. T. (2020). Reducing information frictions in venture capital: The role of new venture competitions. *Journal of Financial Economics*, *136*(3), 676–694.

Howell, S. T., & Nanda, R. (2024). Networking Frictions in Venture Capital, and the Gender Gap in Entrepreneurship. *Journal of Financial and Quantitative Analysis*, *59*(6), 2733–2761. https://doi.org/10.1017/S0022109023000819

Ilgen, D. R., Hollenbeck, J. R., Johnson, M., & Jundt, D. (2005). Teams in organizations: From input-process-output models to IMOI models. *Annu. Rev. Psychol.*, *56*, 517–543.

Jha, M., Qian, J., Weber, M., & Yang, B. (2023). Chatgpt and corporate policies. *Chicago Booth Research Paper*, *23–15*.

Jiang, Y., & Ke, B. (2022). Narcissistic Entrepreneurs and Venture Capital Investment and Contracting. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4134311

Kanze, D., Huang, L., Conley, M. A., & Higgins, E. T. (2018). We ask men to win and women not to lose: Closing the gender gap in startup funding. *Academy of Management Journal*, *61*(2), 586–614.

Kaplan, S. N., & Lerner, J. (2010). It Ain't Broke: The Past, Present, and Future of Venture Capital. *Journal of Applied Corporate Finance*, *22*(2), 36–47. https://doi.org/10.1111/j.1745-6622.2010.00272.x

Kaplan, S. N., Sensoy, B. A., & Strömberg, P. (2009). Should Investors Bet on the Jockey or the Horse? Evidence from the Evolution of Firms from Early Business Plans to Public Companies. *The Journal of Finance*, *64*(1), 75–115. https://doi.org/10.1111/j.1540-6261.2008.01429.x

Kerr, W. R., Lerner, J., & Schoar, A. (2014). The Consequences of Entrepreneurial Finance: Evidence from Angel Financings. *Review of Financial Studies*, *27*(1), 20–55. https://doi.org/10.1093/rfs/hhr098

Kim, A. G., Muhn, M., & Nikolaev, V. V. (2023). Bloated Disclosures: Can ChatGPT Help Investors

Process Information? *Chicago Booth Research Paper*, *23–07*.

Krishna, A., Agrawal, A., & Choudhary, A. (2016). Predicting the Outcome of Startups: Less Failure, More Success. *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 798–805. https://doi.org/10.1109/ICDMW.2016.0118

Lau, D. C., & Murnighan, J. K. (1998). Demographic diversity and faultlines: The compositional dynamics of organizational groups. *Academy of Management Review*, *23*(2), 325–340.

Lerner, J., & Kortum, S. (2000). Assessing the impact of venture capital on innovation. *Rand Journal of Economics*, *31*, 674–692.

Lerner, J., & Nanda, R. (2020). Venture Capital's Role in Financing Innovation: What We Know and How Much We Still Need to Learn. *Journal of Economic Perspectives*, *34*(3), 237–261. https://doi.org/10.1257/jep.34.3.237

Li, K., Mai, F., Shen, R., Yang, C., & Zhang, T. (2023). Dissecting Corporate Culture Using Generative AI–Insights from Analyst Reports. *Available at SSRN 4558295*.

Lopez-Lira, A., & Tang, Y. (2023). Can chatgpt forecast stock price movements? Return predictability and large language models. *arXiv Preprint arXiv:2304.07619*.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Lyonnet, V., & Stern, L. H. (2022). *Venture Capital (Mis)Allocation in the Age of AI* (SSRN Scholarly Paper 4035930). https://doi.org/10.2139/ssrn.4035930

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, *26*(3), 356–376.

Mathieu, J. E., Gallagher, P. T., Domingo, M. A., & Klock, E. A. (2019). Embracing Complexity: Reviewing the Past Decade of Team Effectiveness Research. *Annual Review of Organizational Psychology and Organizational Behavior*, *6*(1), 17–46. https://doi.org/10.1146/annurev-orgpsych-012218-015106

Mathieu, J. E., Hollenbeck, J. R., Van Knippenberg, D., & Ilgen, D. R. (2017). A century of work teams in the Journal of Applied Psychology. *Journal of Applied Psychology*, *102*(3), 452–467. https://doi.org/10.1037/apl0000128

McGrath, J. E. (1964). Social psychology: A brief introduction. *Rinehart and Winston*.

Meyer, B., Glenz, A., Antino, M., Rico, R., & González-Romá, V. (2014). Faultlines and Subgroups: A Meta-Review and Measurement Guide. *Small Group Research*, *45*(6), 633–670. https://doi.org/10.1177/1046496414552195

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*.

Ortín-Ángel, P., & Vendrell-Herrero, F. (2010). Why do university spin-offs attract more venture capitalists? *Venture Capital*, *12*(4), 285–306. https://doi.org/10.1080/13691066.2010.486166

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep*

contextualized word representations. 2227–2237.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*.

Rajan, R. G. (2012). Presidential Address: The Corporation in Finance. *The Journal of Finance*, *67*(4), 1173–1217. https://doi.org/10.1111/j.1540-6261.2012.01745.x

Shapley, L. S. (1953). *A value for n-person games*.

Sharchilev, B., Roizner, M., Rumyantsev, A., Ozornin, D., Serdyukov, P., & De Rijke, M. (2018). Web-based Startup Success Prediction. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2283–2291. https://doi.org/10.1145/3269206.3272011

Shepherd, D. A. (1999). Venture capitalists' assessment of new venture survival. *Management Science*, *45*(5), 621–632.

Shimao, H., Kim, S. J., Khern-Am-Nuai, W., & Cohen, M. C. (2025). Revisiting the CEO Effect Through a Machine Learning Lens. *Management Science*, mnsc.2023.03625. https://doi.org/10.1287/mnsc.2023.03625

Te, Y.-F., Wieland, M., Frey, M., Pyatigorskaya, A., Schiffer, P., & Grabner, H. (2022). *Predicting the Success of Startups Using Crunchbase and Linkedin Data* (SSRN Scholarly Paper 4217648). https://doi.org/10.2139/ssrn.4217648

Zhang, L., Xiang, W., Zhao, C., Zhao, H., Li, R., & Wu, R. (2022). Co-promotion Predictions of Financing Market and Sales Market: A Cooperative-Competitive Attention Approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(8), 9040–9047. https://doi.org/10.1609/aaai.v36i8.20888

Zhang, S., Zhong, H., Yuan, Z., & Xiong, H. (2021). Scalable Heterogeneous Graph Neural Networks for Predicting High-potential Early-stage Startups. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2202–2211. https://doi.org/10.1145/3447548.3467383

Appendix 1. Detailed Methodology for Measuring Faultlines Using the ASW Method

Here's a detailed breakdown of ASW steps:

Step 1: Clustering Procedures

1. **Initial Subgroup Configuration**:
   - The process starts by treating each **team member** as their own subgroup. This creates a configuration where each individual is isolated in a group of size one.
   - From this initial state, a stepwise merging process is initiated, where the algorithm attempts to merge the two subgroups that are most similar to one another.
2. **Clustering Algorithms**:
   - Two common clustering techniques are used to guide the merging of subgroups:
     - **Average Linkage Clustering**: This method calculates the average distance between all pairs of members across two different subgroups. Subgroups with the smallest average distance are merged at each step.
     - **Ward's Method**: This clustering approach seeks to minimize the increase in the total within-subgroup variance at each step. It merges the subgroups that result in the smallest increase in the overall variance.
   - These two clustering techniques are applied separately, and the results from both are combined into a **set of plausible subgroup configurations**.
3. **Intermediate Configurations**:
   - The merging process continues until all members are in a single subgroup. At each step of the process, an intermediate configuration is created where a new set of subgroup structures is formed.
   - All these intermediate subgroup configurations (from individual subgroups to a single group) are retained for further analysis.

Step 2: Silhouette Width Calculation and Faultline Strength

1. **Silhouette Width (s) Calculation**:
   - Once the possible subgroup configurations are generated, the **silhouette width (s)** is calculated for each team member in each configuration. The silhouette width measures how well each individual fits within their assigned subgroup compared to other subgroups.
   - The silhouette width for a team member i is calculated as:

$$s(i) = \frac{b(i) - a(i)}{max\ (a(i), b(i))}$$

     Where:
     - $a(i)$ is the **average dissimilarity** (often measured as Euclidean distance) between team member i and the other members of their own subgroup (Cluster A).
     - $b(i)$ is the **lowest average dissimilarity** between team member i and the members of another subgroup (Cluster B).
   - **Dissimilarity** is calculated using **Euclidean distances** between individuals based on their diversity attributes, which can include both categorical (e.g., gender, ethnicity) and numerical (e.g., age, tenure) variables. For categorical variables, dummy coding is required to convert them into numerical form.

2. **Interpreting Silhouette Widths**:
    - The silhouette width $s(i)$ ranges from **-1 to +1**:
        - **Positive values** (close to 1) indicate that the team member is well-matched to their assigned subgroup.
        - **Values close to 0** indicate that the member is on the boundary between subgroups, with an ambiguous assignment.
        - **Negative values** suggest that the member may have been assigned to the wrong subgroup.
    - The **average silhouette width (ASW)** for the whole team is calculated by averaging the individual silhouette widths for all team members in a configuration. This provides a global assessment of how well the members fit into the subgroup structure as a whole.
3. **Optimal Subgroup Configuration**:
    - The configuration with the **highest ASW value** is chosen as the optimal subgroup structure. This is the configuration where the subgroup members are the most homogeneous internally and the subgroups are the most distinct from one another.
    - The **maximum ASW value** reflects the strongest possible faultlines in the team, indicating the presence of clear and well-defined subgroups.

Appendix 2. Methodology for Extracting Hybrid Features

This appendix describes the methodology and specific prompts used to construct the Hybrid features from individual resumes. The process involves two major steps: extracting key personal information, followed by extracting detailed job records.

Prompts to extract key personal information:

Task Objective: Convert unstructured Chinese text from a resume into a structured JSON format by extracting key information. Ensure all extracted data is appropriately classified and formatted.
Steps to Follow:

1. Extract Key Information: Identify and extract the following information from the resume:
    1.gender (1 for male, 2 for female),
    2.birth_year (Four-digit year of birth),
    3.ethnicity (Ethnic group),
    4.nationality (Country of citizenship),
    5.highest_degree (1: Below Bachelor,2: Bachelor's,3: Master's, 4: PhD)
    6.bachelor_university (Name of the university for Bachelor's degree),
    7.master_university (Name of the university for master's degree),
    8.phd_university (Name of the university for PhD),
    9.entrepreneur (has the individual been an entrepreneur? 1 for yes, 0 for no),
    10.university_worker(has the individual worked in a university? 1 for yes, 0 for no),

2. Format Output as JSON: Organize the extracted information into a JSON object. Each item of information should be represented by a key-value pair in the JSON. If any information is unavailable in the resume, use -999 as the value to indicate it is not applicable.

Let me show you two examples:

Example Input and Output:

Example 1:

Input resume:"陈国东先生,董事,出生于 1979 年 10 月,中国国籍,汉族,无境外永久居留权,大专学历。1999 年 6 月至今,在成功集团任职副总经理;2004 年 10 月至 2012 年 2 月在长治县西山煤业有限公司担任销售经理;2012 年 2 月至今在成功煤炭经销担任总经理;2015 年 5 月至 8 月,在有限公司担任董事;2015 年 8 月底,股份公司成立后,担任董事。2018 年 9 月 4 日因换届辞去浙江研和新材料股份有限公司董事职务。2018 年 9 月 5 日任职于浙江研和新材料股份有限公司董事职务。"

JSON output:{{"gender": "1","birth_year": "1979","ethnicity": "汉族","nationality": "中国","highest_degree": "1","bachelor_university": "-999","master_university": "-999","phd_university": "-999","entrepreneur": "0","university_worker": "0"}}

Example 2: the resume is:"丁晓霜女士,中国国籍,无境外永久居留权,1975 年生,博士研究生学历。1992 年 9 月至 1997 年 7 月北京针灸骨伤学院针灸学学士,1997 年 8 月至 2000 年 7 月中国中医研究院望京医院内科医生,2001 年 9 月至 2004 年 7 月中国中医研究院中药学硕士,2004 年 7 月至 2009 年 7 月中国协和医科大学药物研究所新药安全评价研究中心(暨北京协和建昊医药技术开发有限公司,GLP 实验室)专题负责人,2009 年 9 月至 2013 年 7 月北京协和医学院(清华大学医学部)药理学博士,2011 年 9 月至 2013 年 5 月美国亨利福特医疗中心神经科学研究室 pre-doctor,2013 年 7 月至 2014 年 6 月上海市食品药品检验所药理毒理室检验员,2014 年 10 月至今先后任本公司项目评估部经理、药物信息部经理。"

JSON output:{{"gender": "2","birth_year": "1975","ethnicity": "-999","nationality": "中国","highest_degree": "4","bachelor_university": "北京针灸骨伤学院","master_university": "中国中医研究院","phd_university": "北京协和医学院(清华大学医学部)","entrepreneur": "0","university_worker": "1"}}

The text is as follows:

{YOUR_TEXT}

Prompts to extract detailed job records:

Task Objective: Please analyze the following unstructured Chinese text to extract job records. Perform this task in a structured manner:

Steps to Follow:

1. Identify: Identify all job record from the text.
2. Extract Key Information: For each job record, extract key information including:
    1. start_date: Use a four-digit year. If no start date is provided, use "-999".
    2. end_date: Use a four-digit year. If the text indicates the current position with "今", use "今". If no end date is provided, use "-999".
    3. company_name: Extract the full name of the company. If not specified, use "-999".
    4. position_title: Extract the full job title. If not specified, use "-999".
3. Format as JSON: Format the output as a single JSON object. Within this object, each job record should be a key-value pair, where the key is 'Job Record X' (X being the sequence number) and the value is a JSON object containing the job details.

Example Input and Output:

Example 1:

Input resume:"孙良菊,女,1979 年 7 月生,中国国籍,无境外永久居住权,高中学历。2014 之前自主创业,2014 年至 2015 年在安徽文保中心任古代建筑设计师。2015 年 7 月至今就职于安徽恒

瑞新能源股份有限公司,任销售经理。"

JSON output:{{

"Job Record 1": {{"start_date": "-999", "end_date": "2014", "company_name": "自主创业", "position_title": "-999"}},

"Job Record 2": {{"start_date": "2014", "end_date": "2015", "company_name": "安徽文保中心", "position_title": "古代建筑设计师"}},

"Job Record 3": {{"start_date": "2015", "end_date": "今", "company_name": "安徽恒瑞新能源股份有限公司", "position_title": "销售经理"}}

}}

Example 2:

Input resume:"宋伟先生,1969 年 1 月出生,中国国籍,无境外永久居留权,本科学历。1996 年-2005 年任上海冠恒工业设备有限公司总经理;2005 年-2013 年,任浙江德马科技有限公司副总经理;2014 年 1 月至今,任浙江德马科技股份有限公司董事、北京帝欧咨询有限公司副董事长。"

JSON output:{{

"Job Record 1": {{"start_date": "1996", "end_date": "2005", "company_name": "上海冠恒工业设备有限公司", "position_title": "总经理"}},

"Job Record 2": {{"start_date": "2005", "end_date": "2013", "company_name": "浙江德马科技有限公司", "position_title": "副总经理"}},

"Job Record 3": {{"start_date": "2014", "end_date": "今", "company_name": "浙江德马科技股份有限公司", "position_title": "董事"}},

"Job Record 4": {{"start_date": "2014", "end_date": "今", "company_name": "北京帝欧咨询有限公司", "position_title": "副董事长"}},

}}

The text is as follows:

{ YOUR_TEXT}

Appendix 3. Validation of Human-Coded Features and Hybrid Features Using Ground-Truth

To better illustrate the accuracy of the human-coded features and hybrid features used in this paper to extract key information from resumes, we randomly sample 100 manager resumes. Two independent reviewers manually code these resumes to identify individual features employed in the project (Key Member Attributes Proxy, as defined in Table 1). To ensure accuracy, a third reviewer resolves any discrepancies through discussion and consensus. The final manually coded dataset serves as the ground truth for comparison[23].

As a result, we have four distinct versions of the key information:

1. human-coded features,
2. hybrid features,
3. Ground truth for human-coded features,
4. Ground truth for hybrid features.

It is important to note that there is no unified ground truth due to slight differences in the variables between the human-coded features[24] and hybrid features.

As panel A to panel F shows, human-coded features show strong alignment with the ground truth, though education_controller2 (0.484), show moderate alignment. GPT-coded features also achieve alignment with the ground truth, while features like ethnicity_controller (0.419) and entrepreneurial_controller (0.329) show weaker alignment.

Table A3. Validation of Human-Coded Features and Hybrid Features Against Ground Truth

This table presents a validation exercise comparing Human-Coded Features and Hybrid Features against ground-truth resume data from a randomly sampled set of 100 managers. Panel A and Panel B summarize the statistics for human-coded features and their ground truth, respectively, while Panel C reports the correlations between them. Similarly, Panel D and Panel E provide summary statistics for hybrid features and their corresponding ground truth, with Panel F displaying their correlations. All descriptive statistics are computed using non-missing observations only. The number of non-missing observations for each variable is reported alongside the summary statistics for transparency.

Panel A: Summary Statistics for Human-Coded Features

Variable	count	N (non-missing)	mean	std	min	25%	50%	75%	max
human-coded features									
gender_controller	100	98	0.735	0.444	0	0	1	1	1
age_controller	100	96	40.781	8.871	26	33	40	46	63

[23] In some cases, resumes lack relevant feature information or the extraction methods fail to identify them. These instances are treated as missing but are retained in the dataset, as XGBoost can natively handle missing values and determine optimal split directions without the need for imputation or deletion.

[24] Both feature sets share identical definitions for most variables with two notable exceptions.
1)University spinoff attribute: university_controller in Human-coded Features refers to university affiliation as whether an individual was previously a student at a university. In contrast, the hybrid features broaden this definition to include whether an individual was previously a student or has ever worked at a university.
2)Ethnicity attribute: ethnicity_controller is not in Human-coded Features because ethnicity data is not extracted in these manually coded features.
These differences arise because we refine the measurement framework after completing the manual extraction. However, incorporating these updates into the manual human coding process is impractical, as it is prohibitively slow and costly to repeatedly ask research assistants to redo the work. By contrast, the GPT-based method dynamically adapts to these refinements, ensuring consistency with the updated framework while maintaining efficiency and scalability.

nationality_controller	100	94	0.989	0.103	0	1	1	1	1
education_controller1	100	100	0.98	0.778	0	0	1	2	2
education_controller2	100	100	0.06	0.239	0	0	0	0	1
education_controller3	100	100	0.15	0.386	0	0	0	0	2
entrepreneurial_controller	100	100	0.02	0.141	0	0	0	0	1
industry_controller	100	100	16.266	12.764	1.068	6.519	13.825	23.532	80.901
prior_company_controller	100	100	2.53	1.708	0	1	2	4	8
university_controller	100	100	0.69	0.465	0	0	1	1	1
VC_controller	100	100	0.07	0.256	0	0	0	0	1

Panel B: Summary Statistics for Ground Truth of Human-Coded Features

Variable	count	N (non-missing)	mean	std	min	25%	50%	75%	max
Ground truth for human-coded features									
gender_controller	100	98	0.735	0.444	0	0	1	1	1
age_controller	100	96	40.781	8.871	26	33	40	46	63
nationality_controller	100	96	0.99	0.102	0	1	1	1	1
education_controller1	100	100	1	0.791	0	0	1	2	3
education_controller2	100	100	0.08	0.307	0	0	0	0	2
education_controller3	100	100	0.19	0.506	0	0	0	0	3
entrepreneurial_controller	100	96	0.031	0.175	0	0	0	0	1
industry_controller	100	95	15.199	10.137	0.334	6.471	14.116	22.126	41.259
prior_company_controller	100	100	2.45	1.749	0	1	2	4	8
university_controller	100	100	0.7	0.461	0	0	1	1	1
VC_controller	100	100	0.07	0.256	0	0	0	0	1

Panel C: Correlation Between Human-Coded Features and Ground Truth

	Correlation
gender_controller	0.948
age_controller	1
nationality_controller	1
education_controller1	0.968
education_controller2	0.484
education_controller3	0.68
entrepreneurial_controller	0.812
industry_controller	0.77
prior_company_controller	0.961
university_controller	0.93
VC_controller	0.846

Panel D: Summary Statistics for hybrid features

Variable	count	N (non-missing)	mean	std	min	25%	50%	75%	max
Hybrid features									
gender_controller	100	100	0.74	0.441	0	0	1	1	1

age_controller	100	100	40.61	8.934	25	33	40	46	63
ethnicity_controller	100	16	0.875	0.342	0	1	1	1	1
nationality_controller	100	94	0.979	0.145	0	1	1	1	1
education_controller1	100	100	1.1	0.704	0	1	1	2	3
education_controller2	100	100	0.09	0.351	0	0	0	0	2
education_controller3	100	100	0.2	0.532	0	0	0	0	3
entrepreneurial_controller	100	100	0.22	0.416	0	0	0	0	1
industry_controller	100	100	19.061	19.969	1.582	8.561	14.997	23.78	173.634
prior_company_controller	100	100	2.5	1.697	0	1	2	4	8
university_controller	100	100	0.81	0.394	0	1	1	1	1
VC_controller	100	100	0.05	0.219	0	0	0	0	1

Panel E: Summary Statistics for Ground Truth of hybrid features

Variable	count	N (non-missing)	mean	std	min	25%	50%	75%	max
Ground truth forHybrid features									
gender_controller	100	98	0.735	0.444	0	0	1	1	1
age_controller	100	96	40.781	8.871	26	33	40	46	63
ethnicity_controller	100	15	0.533	0.516	0	0	1	1	1
nationality_controller	100	96	0.99	0.102	0	1	1	1	1
education_controller1	100	100	1	0.791	0	0	1	2	3
education_controller2	100	100	0.08	0.307	0	0	0	0	2
education_controller3	100	100	0.19	0.506	0	0	0	0	3
entrepreneurial_controller	100	96	0.031	0.175	0	0	0	0	1
industry_controller	100	95	15.199	10.137	0.334	6.471	14.116	22.126	41.259
prior_company_controller	100	100	2.45	1.749	0	1	2	4	8
university_controller	100	100	0.72	0.451	0	0	1	1	1
VC_controller	100	100	0.07	0.256	0	0	0	0	1

Panel F: Correlation Between hybrid features and Ground Truth

	Correlation
gender_controller	1
age_controller	1
ethnicity_controller	0.419
nationality_controller	0.703
education_controller1	0.871
education_controller2	0.962
education_controller3	0.945
entrepreneurial_controller	0.329
industry_controller	0.584
prior_company_controller	0.934
university_controller	0.663
VC_controller	0.836

Appendix 4. Predictive Power of CEO-Only Characteristics

To further investigate the role of team composition in predicting high-growth firms, we conduct an auxiliary analysis using a reduced model based solely on the characteristics of a single individual. Specifically, we prioritize the Chairman of the Board; if their data is unavailable, we substitute the CEO's information. This approach enables us to isolate the predictive contribution of individual-level leadership features, similar to the methodologies employed by Davenport (2022) and Lyonnet and Stern (2022), and contrast it with the richer team-level models presented in the main text.

For this analysis, we construct a feature set derived solely from the resume of the selected individual, applying the same feature extraction procedures (human-coded features and hybrid features) as used in our primary models. Because this model focuses on a single individual per firm, several team-level constructs from Table 1, such as the Average Member Attributes Proxy, Diversity Proxy, and Faultline Proxy, cannot be computed and are therefore excluded. Instead, the CEO-only model includes variables captured under the Key Member Attributes Proxy. These proxies reflect key individual attributes, including gender, age, ethnicity, nationality, education, entrepreneurial experience, industry experience, prior company affiliations, university spin-off background, and VC connections. A detailed overview of these variables, including their measurement formulas, is provided in the Key Member Attributes Proxy columns of Panel A in Table 1.

We evaluate the CEO-only model using the same outcome variable. Specifically, we predict whether a firm achieves three-year sales growth exceeding 200%. We apply the same machine learning pipeline and the same evaluation metrics as used in the main paper.

The results for the CEO-only model are presented in Panels A, B, and C of Table A4. Overall, the predictive performance of the CEO-only features is lower than that of the full team-level models. As shown in Panel A of Table A4, the CEO-only model achieves a average NDCG score of 0.282 for the hybrid features and 0.233 for the human-coded features. In comparison, the corresponding team-level models achieve higher average scores: 0.309 for the hybrid features and 0.245 for the human-coded features. In addition, the performance of the CEO-only models is much more volatile across the two test years than our team-based models.

These findings suggest that while CEO characteristics alone carry significant predictive value, models that incorporate the broader top management team consistently outperform those relying solely on CEO features. This underscores the importance of considering the collective attributes of the team, rather than focusing exclusively on the highest ranked individual leader, when forecasting entrepreneurial success.

Table A4. Out-of-sample Performance Metrics of CEO-Only Predictive Models with Different Information Sets

This table presents out-of-sample performance metrics for predicting whether a firm will achieve three-year sales growth exceeding 200%. Panel A reports the overall average performance across all test years. Panel B and Panel C provide model performance for the years 2018 and 2019, respectively. Performance is evaluated using NDCG@K, along with PRECISION@K and RECALL@K. The proportion of high-growth firms in each test sample (average DV) is also included for context.

Panel A: Average Performance Metrics Over (2018-2019)

Input features	NDCG	average_DV	precision_at_k	recall_at_k
Team Information				
Human-Coded Features	0.233	0.078	0.188	0.188
Hybrid Features	0.282	0.078	0.262	0.262

Panel B: Model Performance for 2018

Input features	NDCG	average_DV	precision_at_k	recall_at_k
Team Information				
Human-Coded Features	0.065	0.073	0.075	0.075
Hybrid Features	0.193	0.073	0.125	0.125

Panel C: Model Performance for 2019

Input features	NDCG	average_DV	precision_at_k	recall_at_k
Team Information				
Human-Coded Features	0.4	0.09	0.3	0.3
Hybrid Features	0.371	0.09	0.4	0.4

Table 1. Variable Definitions

This table summarizes variable definitions. Panel A presents the proxies, definitions, and measurement approaches for human-coded and hybrid features, including compositional and structural features. Panel B outlines the attributes used for faultline measurement, applying the Average Silhouette Width (ASW) method. Panel C provides definitions for financial variables, including those from Selected Financial Data and those used to compare high-growth firms with other firms at the time of listing. Specific measurement formulas for each variable are provided within each panel.

Panel A. Measurement of Human-coded and Hybrid Features

		Structural Features	**Compositional Features**		
	Key Reference	Key Member Attributes Proxy	Average member Attributes Proxy	Diversity Proxy	Faultline Proxy
gender	Sharchilev et al. (2018**)**	***gender_controller:*** whether the controller is male (1) or not (0)	***gender_average1/ gender_average2***: the proportion of male and female members	Covered by Average member Attributes Proxy (Since gender is a binary variable, the variance can be directly derived from the average value through simple mathematical derivation. The average value provides both average member attributes and insight into gender diversity)	***Faultline:*** ***ASW values***: ASW value for measuring faultline strength across all individual attributes (The attributes used in faultline analysis and asw measurement is detailed in panel B)
age	Esen et al. (2023)	***age_controller***: the age of the controller	***age_average***: The average age of the top management team	***age_diversity*****:** the variance in ages of the top management team	
race/ethnicity	Calder-Wang et al. (2021)	***ethnicity_controller***: whether the controller is Han (1) or not (0)	***ethnicity_average***: Proportion of Han members	***ethnicity_diversity***: The diversity score[25]: $Diversity\ Score_i = 1 - \frac{\#\ of\ ties\ between\ team\ members\ in\ team_i\ with\ shared\ ethn}{Total\ possible\ ties\ in\ team_i}$	
nationality	Calder-Wang et al. (2021)	***nationality_controller***: whether the controller is Chinese (1) or not	***nationality_average***: Proportion of Chinese members	***nationality_diversity***: The diversity score:	

[25] For illustration, consider a TMT comprising six individuals: three Han, two Zhuang, and one Hui. The race/ethnicity diversity score for this team is determined by 1-(3+1)/(5+4+3+2+1)=11/15. This calculation considers three intra-ethnic ties among Han members, one tie among Zhuang members, against the total of fifteen possible ties within the team. The diversity score increases with greater ethnic diversity, approaching zero when homogeneity prevails and one when each member represents a distinct ethnicity.

		(0)		$Diversity\ Score_i = 1 - \frac{\#\ of\ ties\ between\ team\ members\ in\ team_i\ with\ shared\ nati}{Total\ possible\ ties\ in\ team_i}$
education	Arroyo et al. (2019)	1) ***education_controller1:*** The total number of higher education degrees obtained by the controller. 2) ***education_controller2:*** The number of higher education degrees the controller has obtained from elite institutions (C9 League[26]) 3) ***education_controller3:*** The number of higher education degrees the controller has obtained from elite institutions (985 program[26])	1) ***education_average1***: average number of higher education degrees (Bachelor's, Master's, and PhD) 2) ***education_average2***: average proportion of higher education degrees from elite institutions (C9 League[26]) 3) ***education_average3***: average proportion of higher education degrees from elite institutions (985 program[26])	1) ***education_diversity1:*** Variance in number of higher education degrees 2) ***education_diversity2:*** The diversity score: $Diversity\ Score_i = 1 - \frac{\#\ of\ ties\ between\ team\ members\ in\ team_i\ with\ shared\ scho}{Total\ possible\ ties\ in\ team_i}$
Entrepreneurial Experience	Te et al. (2022)	***entrepreneurial_controller:*** whether the controller is serial entrepreneur (1) or not (0)	***entrepreneurial_average:*** Proportion of serial entrepreneurs	Covered by Average member Attributes Proxy (Since serial entrepreneur is a binary variable, the variance can be directly derived from the average value through simple mathematical derivation. The average value provides both average member attributes and insight into entrepreneurial experience diversity)
Industry experience	Te et al. (2022)	***Industry_controller***:	***Industry_average***:	***Industry_diversity:***

[26] In 1998, China initiated the '985 program,' encompassing 39 top universities aimed at developing the Chinese higher education system through significant national and local government funding. These universities are considered the elite of China's higher education sector. The C9 League, formed as part of the first batch of the 985 program, includes nine of the most prestigious universities in China: Peking University, Tsinghua University, Fudan University, Shanghai Jiao Tong University, Nanjing University, the University of Science and Technology of China, Zhejiang University, Xi'an Jiao Tong University, and the Harbin Institute of Technology (Y. Chen et al., 2022). This classification underscores the high caliber of education and research within these institutions, making degrees from these universities a significant indicator of educational quality in our study.

		The years of work experience of the controller	The average years of work experience among the top management team	Variance in years of work experience
Prior Company Affiliation	Beckman et al. (2007)	***Prior_company_controller:*** The number of unique prior companies of the controller.	***Prior_company_average:*** Average count of unique prior companies.	***prior_company_diversity***: The diversity score: $Diversity\ Score_i = 1 - \frac{\#\ of\ ties\ between\ team\ members\ in\ team_i\ with\ shared\ prio}{Total\ possible\ ties\ in\ team_i}$
University Spin-offs	Ortín-Ángel & Vendrell-Herrero (2010)	***university_controller***: whether the controller is university-affiliated (1) or not (0)	***university_average***: Proportion of university-affiliated team members	Covered by Average member Attributes Proxy (Since university-affiliated is a binary variable, the variance can be directly derived from the average value through simple mathematical derivation. The average value provides both average member attributes and insight into university-affiliated diversity)
VC connections		***VC_controller:*** Whether the controller has previous employment experience at VC firms (1) or not (0)	***VC_average:*** Proportion of team members with previous employment at VC firms	Covered by Average member Attributes Proxy (Since previous VC employment experience is a binary variable, the variance can be directly derived from the average value through simple mathematical derivation. The average value provides both average member attributes and insight into VC connection diversity)

Panel B. Attributes Used for Measuring Faultline Strength Using the ASW Method

Attribute Name	Individual Proxy
gender	whether the individual is male (1) or not (0)
age	the age of the individual
race/ethnicity	whether the individual is Han (1) or not (0)
nationality	whether the individual is Chinese (1) or not (0)
education	1) The total number of academic degrees obtained by the individual. 2) The number of degrees the individual has obtained from elite institutions (C9 League) 3) The number of degrees the individual has obtained from elite institutions (985 program)
Entrepreneurial Experience	whether the individual is serial entrepreneur (1) or not (0)
Industry experience	The years of work experience of the individual
Prior Company Affiliation	The number of unique prior companies of the individual.
University Spin-offs	whether the individual is university-affiliated (1) or not (0)
VC connections	Whether the individual has previous employment experience at VC firms (1) or not (0)

Panel C. Variable Definitions for Selected Financial Data and Firm Characteristics at Listing

Variable Name	Definition
Selected Financial Data	
cash_holdings	Total cash and cash equivalents
non_cash_assets	Total assets excluding cash and cash equivalents.
long_term_debt	Outstanding debt obligations with maturities longer than one year.
annual_revenues	Total revenue
R&D_expenditures	Expenses incurred for research and development activities
cost_of_sales	Direct costs associated with producing goods or services sold
SG&A_expenses	Selling, general, and administrative expenses incurred in business operations
Characteristics in High-Growth vs. Other Firms at Listing	
CASH	Cash and cash equivalents
LEVERAGE (LEV)	Total debts divided by total assets.
PAST SALES GROWTH RATE	Growth rate of sales prior to listing, calculated as (sales in listing year - sales one year before listing) / sales one year before listing (when negative, the absolute value is used) ×100%
ROA	Operating profits divided by total assets.
ROE	Operating profits divided by total equity.
TANGIBLE ASSETS %	Fixed assets divided by total assets.
TOTAL ASSETS	Total assets
SALES	Sales

Table 2. Three-Year Post-Listing Sales Growth Percentile for NTB firms

This table presents the percentile rankings of three-year post-listing sales growth rates. Growth is calculated over multiple overlapping periods, pooling all observed growth rates into a single dataset for percentile ranking. The table reports growth rates at various percentiles.

Training Set	Top 50%	Top 20%	Top 15%	Top 10%	Top 8%	Top 6%	Top 4%	Top 2%
2013, 2014	33.56	127.95	170.05	241.83	295.63	352.04	462.55	1121.89
2013, 2014, 2015	41.15	130.01	165.81	225.16	265.89	315.08	417.12	808.92

Table 3. Sample Selection

This table details the sample selection process and data sources. The study uses two ten-year datasets on NTB firms (2013–2022): one from CSMAR, containing top management team information for 13,020 firms, and one from Choice, providing financial and sales data for 13,019 firms. To measure high-growth firms, firms must have at least three years of post-listing sales data, reducing the sample to 12,787 firms (2013–2019). After integrating all data using stock ID, the final analytical sample consists of 12,567 NTB firms.

Step	Description	Data Source	Initial Count	Final Count
1	Collection of top management team information for NTB firms listed from the beginning of NTB in 2013 to the last day of data collection in 2022.	CSMAR	13,020	-
2	Collection of initial financial records and additional sales data for NTB firms listed from the beginning of NTB in 2013 to the last day of data collection in 2022.	Choice	13,019	-
3	Requirement for data extending into the third year beyond initial listing to define the dependent variable.	-	-	12,787
4	Integration of resume, financial, and sales data using stock ID as a unique identifier.	Combined Data Sources	-	12,567

Table 4. Summary Statistics of High-Growth Firms

This table presents summary statistics for high-growth firms. Panel A reports the annual distribution of firms classified as high growth from 2013 to 2019. Panel B compares key financial metrics for firms classified as high-growth versus low-growth firms at the time of listing. The reported values represent the mean (median) [standard deviation] for each group. Differences between the two groups are assessed using t-tests and rank-sum tests, with corresponding p-values reported in the last column. Statistical significance is denoted as *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Variable definitions are provided in Panel C of Table 1.

Panel A: Annual Distribution of High-Growth Firms

Listing Year	Total Firms	High Growth Firms	Proportion of High-Growth Firms
2013	154	21	0.136
2014	1172	128	0.109
2015	3488	293	0.084
2016	4919	298	0.061
2017	2066	97	0.047
2018	547	40	0.073
2019	221	20	0.09
Overall	12567	897	0.071

Panel B: Financial Characteristics of High-Growth vs. Low-growth Firms at Listing

Variable	Low growth	High growth	P value for the t-test (and ranksum test)
TOTAL ASSETS	22098.78 (8325.465) [111207.1]	15858.64 (5049.157) [95611.35]	0.063* (0.000***)
SALES	16114.3 (6709.575) [61171.38]	9521.594 (3006.965) [73022.57]	0.009*** (0.000***)
ROA	0.039 (0.054) [0.225]	-0.004 (0.034) [0.303]	0.000*** (0.000***)
ROE	-0.258 (0.104) [29.293]	-5.414 (0.062) [158.855]	0.332 (0.000***)
LEVERAGE	1.594 (0.738) [37.063]	14.739 (0.544) [413.377]	0.341 (0.000***)
CASH	3412.541 (768.382) [35726.18]	3862.239 (662.394) [50546.46]	0.794 (0.002***)
PAST SALES GROWTH RATE	21.58 (-10.306)	-30.149 (-42.49)	0.000*** (0.000***)

	[914.727]	[103.362]	
TANGIBLE ASSETS %	94.893 (97.449) [7.494]	94.143 (98.215) [11.005]	0.082* (0.117)

Table 5. Summary Statistics and Comparison of Human-Coded Features and Hybrid Features

This table presents summary statistics and comparisons between human-coded features and hybrid features. Panel A reports human-coded features. Panel B summarizes the same attributes using GPT-extracted features. Panel C examines the pairwise correlations between human-coded features and hybrid features.

Panel A: Summary Statistics for Human-coded Features (Sample Period: 2013–2019)

Variable	count	mean	std	min	25%	50%	75%	max
Key member attributes								
gender_controller	12567	0.893	0.309	0	1	1	1	1
age_controller	12567	46.651	8.006	17	41	46	52	83
nationality_controller	12567	0.992	0.091	0	1	1	1	1
education_controller1	12567	0.979	0.827	0	0	1	2	4
education_controller2	12567	0.052	0.229	0	0	0	0	2
education_controller3	12567	0.12	0.345	0	0	0	0	3
entrepreneurial_controller	12567	0.092	0.29	0	0	0	0	1
industry_controller	12567	28.3	19.069	0.329	16.909	24.726	34.702	321.279
prior_company_controller	12567	3.089	1.866	0	2	3	4	19
university_controller	12567	0.651	0.477	0	0	1	1	1
VC_controller	12567	0.046	0.208	0	0	0	0	1
Average member attributes								
gender_average1	12567	0.718	0.18	0	0.6	0.75	0.857	1
gender_average2	12567	0.282	0.18	0	0.143	0.25	0.4	1
age_average	12567	41.606	5.05	24	38	41.364	44.857	74
nationality_average	12567	0.994	0.036	0	1	1	1	1
education_average1	12567	0.763	0.416	0	0.444	0.75	1	2.5
education_average2	12567	0.032	0.086	0	0	0	0	1
education_average3	12567	0.091	0.152	0	0	0	0.154	1
entrepreneurial_average	12567	0.035	0.082	0	0	0	0	1
industry_average	12567	18.254	7.765	0.342	13.325	17.32	22.117	160.804
prior_company_average	12567	2.425	1.001	0	1.8	2.444	3.026	7.5
university_average	12567	0.573	0.268	0	0.375	0.6	0.8	1
VC_average	12567	0.055	0.102	0	0	0	0.1	1
Diversity attributes								
age_diversity	12567	76.687	55.91	0	37.148	63.333	101.544	968
nationality_diversity	12567	0.016	0.086	0	0	0	0	1
education_diversity1	12567	0.449	0.254	0	0.267	0.444	0.611	4.5
education_diversity2	12567	0.992	0.031	0	1	1	1	1
industry_diversity	12567	167.801	662.582	0	46.27	90.649	164.854	50716.97
prior_company_diversity	12567	0.749	0.291	0	0.583	0.872	0.982	1
Faultline attributes								
ASW values	12567	0.553	0.134	0	0.491	0.564	0.636	0.93

Panel B: Summary Statistics for Hybrid Features (Sample Period: 2013–2019)

Variable	count	mean	std	min	25%	50%	75%	max
Key member attributes								
gender_controller	12567	0.894	0.308	0	1	1	1	1
age_controller	12567	46.577	8.05	17	41	46	52	83
ethnicity_controller	12567	0.939	0.239	0	1	1	1	1
nationality_controller	12567	0.984	0.126	0	1	1	1	1
education_controller1	12567	1.067	0.767	0	0	1	2	3
education_controller2	12567	0.063	0.258	0	0	0	0	2
education_controller3	12567	0.15	0.387	0	0	0	0	3
entrepreneurial_controller	12567	0.846	0.361	0	1	1	1	1
industry_controller	12567	26.747	14.061	0	17.198	24.378	33.485	136.205
prior_company_controller	12567	3.01	1.762	0	2	3	4	10
university_controller	12567	0.749	0.433	0	0	1	1	1
VC_controller	12567	0.033	0.18	0	0	0	0	1
Average member attributes								
gender_average1	12567	0.721	0.186	0	0.6	0.75	0.857	1
gender_average2	12567	0.279	0.186	0	0.143	0.25	0.4	1
age_average	12567	41.547	5.186	12	37.9	41.3	44.857	74
ethnicity_average	12567	0.863	0.315	0	1	1	1	1
nationality_average	12567	0.988	0.06	0	1	1	1	1
education_average1	12567	0.92	0.408	0	0.636	1	1.2	2.333
education_average2	12567	0.034	0.081	0	0	0	0	1
education_average3	12567	0.094	0.141	0	0	0	0.167	1
entrepreneurial_average	12567	0.271	0.182	0	0.143	0.25	0.375	1
industry_average	12567	18.904	6.792	0	14.306	18.158	22.745	72.669
prior_company_average	12567	2.56	0.901	0	2	2.571	3.125	8
university_average	12567	0.74	0.263	0	0.571	0.818	1	1
VC_average	12567	0.048	0.096	0	0	0	0.083	1
Diversity attributes								
age_diversity	12567	76.298	57.293	0	35.839	62.697	100.571	612.5
ethnicity_diversity	12567	0.367	0.457	0	0	0	1	1
nationality_diversity	12567	0.031	0.134	0	0	0	0	1
education_diversity1	12567	0.359	0.246	0	0.205	0.286	0.5	4.5
education_diversity2	12567	0.986	0.045	0	1	1	1	1
industry_diversity	12567	134.785	217.809	0	45.341	86.586	153.987	5099.878
prior_company_diversity	12567	0.721	0.293	0	0.533	0.833	0.972	1
Faultline attributes								
ASW values	12567	0.549	0.132	0	0.489	0.561	0.629	0.918

Panel C: Pairwise Correlations Between Human-Coded Features And Hybrid Features (Sample Period: 2013–2019)

Variable	Correlation
Key member attributes	
gender_controller	0.999
age_controller	0.997
nationality_controller	0.726
education_controller1	0.882
education_controller2	0.85
education_controller3	0.842
entrepreneurial_controller	0.127
industry_controller	0.839
prior_company_controller	0.673
university_controller	0.757
VC_controller	0.804
Average member attributes	
gender_average1	0.936
gender_average2	0.936
age_average	0.953
nationality_average	0.631
education_average1	0.838
education_average2	0.762
education_average3	0.788
entrepreneurial_average	0.259
industry_average	0.817
prior_company_average	0.872
university_average	0.701
VC_average	0.829
Diversity attributes	
age_diversity	0.927
nationality_diversity	0.578
education_diversity1	0.686
education_diversity2	0.772
industry_diversity	0.209
prior_company_diversity	0.837
Faultline attributes	
ASW values	0.782

Table 6. Summary Statistics for Selected Financial Data Method (Sample Period: 2013–2019)

This table presents summary statistics for key financial indicators extracted using the Selected Financial Data Method for firms listed between 2013 and 2019.

Variable	**count**	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
cash_holdings	12567	3444.652	36979.39	-388.088	262.981	762.591	2012.055	1909120
non_cash_assets	12567	6873.997	38132.96	0	410.377	1699.973	5409.545	2482620
long_term_debt	12567	2573.275	19457.32	-56.071	32.41	219	981.612	860802.6
annual_revenues	12567	15643.73	62112.15	-1615.34	2835.54	6348.087	13814.05	3179490
R&D_expenditures	12567	863.437	3262.364	0	160.606	381.065	804.054	81989.97
cost_of_sales	12567	930.398	2949.957	0	132.552	327.252	798.255	96820.58
SG&A_expenses	12567	2418.532	4708.96	111.293	808.447	1385.361	2529.439	219375.4

Note: All monetary values are expressed in units of 10,000 yuan.

Table 7. Hyperparameters for XGBoost

This table outlines the hyperparameters and their candidate values used during the tuning process, inspired by the methodology of C. Chen et al. (2024). We focus on fine-tuning four principal hyperparameters while maintaining fixed values for others to strike a balance between prediction accuracy and computational efficiency. The hyperparameters are presented by their standard names as used in XGBoost tutorials, with definitions also provided in the table. The first six rows list parameters set at fixed values; the remaining entries show those with multiple candidate values. Given our binary dependent variable, we set the 'objective' hyperparameter to 'binary:logistic', indicating a binary classification task. To ensure reproducibility, we use a fixed seed for the random number generation. The number of decision trees is fixed at 200. The parameters 'subsample' and 'colsample_bytree', which help add randomness and enhance robustness to noise, are kept at 0.8. The scale_pos_weight is dynamically calculated to balance the positive and negative classes, ensuring effective handling of imbalanced datasets. Our primary focus is on tuning 'learning rate', 'max_depth', 'min_child_weight' , and 'gamma' to optimize model performance.

Hyperparameters	Definitions	Candidate values
objective	The learning task and the learning objective	binary:logistic
seed	The seed of random number	42
num_round	Number of trees	200
subsample	Subsample ratio of the training instances	0.8
colsample_bytree	Subsample ratio of features when constructing each tree	0.8
scale_pos_weight	Balances the positive and negative classes by weighting the minority class.	the ratio of negative instances to positive instances if there are positive instances, otherwise set to 1.
eta	Learning rate	0.01, 0.1, 0.3
max_depth	Maximum depth of a tree	1, 2, 3, or no limit
min_child_weight	Minimum sum of instance weight (hessian) needed in a child	4, 5, 6
gamma	Minimum loss reduction required to make a further partition on a leaf node of the tree.	0.1, 0.5, 1

Table 8. Out-of-sample Performance Metrics of Predictive Models with Different Information Sets

This table presents out-of-sample performance metrics for predicting whether a firm will achieve three-year sales growth exceeding 200%. Panel A reports the overall average performance across all test years. Panel B and Panel C provide model performance for the years 2018 and 2019, respectively. Performance is evaluated using NDCG@K, along with PRECISION@K and RECALL@K. The proportion of high-growth firms in each test sample (average DV) is also included for context.

Panel A: Average Performance Metrics Over (2018-2019)

Input features	NDCG	average_DV	precision_at_k	recall_at_k
Team Information				
Human-Coded Features	0.245	0.078	0.238	0.238
Hybrid Features	0.309	0.078	0.288	0.288
Machine-Coded Features	0.205	0.078	0.2	0.2
Business fundamentals information				
Selected Financial Data	0.239	0.078	0.175	0.175
Comprehensive Financial Data	0.185	0.078	0.162	0.162
Integrated (Team & Business Fundamentals)				
Human-Coded Features merged with Selected Financial Data (early fusion)	0.187	0.078	0.213	0.213
Hybrid Features merged with Selected Financial Data (early fusion)	0.203	0.078	0.225	0.225
Human-Coded Features merged with Selected Financial Data (late fusion)	0.305	0.078	0.262	0.262
Hybrid Features merged with Selected Financial Data (late fusion)	0.291	0.078	0.275	0.275

Panel B: Model Performance for 2018

Input features	NDCG	average_DV	precision_at_k	recall_at_k
Team Information				
Human-Coded Features	0.222	0.073	0.175	0.175
Hybrid Features	0.245	0.073	0.275	0.275
Machine-Coded Features	0.164	0.073	0.2	0.2
Business fundamentals information				
Selected Financial Data	0.141	0.073	0.1	0.1
Comprehensive Financial Data	0.15	0.073	0.175	0.175
Integrated (Team & Business Fundamentals)				
Human-Coded Features merged with Selected Financial Data (early fusion)	0.202	0.073	0.225	0.225
Hybrid Features merged with Selected Financial Data (early fusion)	0.202	0.073	0.2	0.2
Human-Coded Features merged with Selected Financial Data (late fusion)	0.274	0.073	0.225	0.225
Hybrid Features merged with Selected Financial Data (late fusion)	0.238	0.073	0.25	0.25

Panel C: Model Performance for 2019

Input features	NDCG	average_DV	precision_at_k	recall_at_k
Team Information				
Human-Coded Features	0.268	0.09	0.3	0.3
Hybrid Features	0.373	0.09	0.3	0.3
Machine-Coded Features	0.245	0.09	0.2	0.2
Business fundamentals information				
Selected Financial Data	0.338	0.09	0.25	0.25
Comprehensive Financial Data	0.22	0.09	0.15	0.15
Integrated (Team & Business Fundamentals)				
Human-Coded Features merged with Selected Financial Data (early fusion)	0.171	0.09	0.2	0.2
Hybrid Features merged with Selected Financial Data (early fusion)	0.205	0.09	0.25	0.25
Human-Coded Features merged with Selected Financial Data (late fusion)	0.335	0.09	0.3	0.3
Hybrid Features merged with Selected Financial Data (late fusion)	0.343	0.09	0.3	0.3

Table 9. Out-of-sample Performance Metrics Using Proportional K Derived from Training Set

This table presents out-of-sample performance metrics for predicting whether a firm will achieve three-year sales growth exceeding 200%. Panel A reports the overall average performance across all test years. Panel B and Panel C provide model performance for the years 2018 and 2019, respectively. Performance is evaluated using NDCG@K, along with PRECISION@K and RECALL@K after adjusting K to mitigate potential hindsight bias. K is determined by calculating the proportion of NTB firms in the training sample and applying it to the test sample size. The adjusted selection percentages are 11.2% for 2018 and 9.2% for 2019. The proportion of high-growth firms in each test sample (average DV) is also included for context.

Panel A: Average Performance Metrics Over (2018-2019)

Input features	NDCG	average_DV	precision_at_k	recall_at_k
Team Information				
Human-Coded Features	0.276	0.078	0.24	0.288
Hybrid Features	0.317	0.078	0.248	0.3
Machine-Coded Features	0.212	0.078	0.174	0.213
Business fundamentals information				
Selected Financial Data	0.258	0.078	0.182	0.212
Comprehensive Financial Data	0.193	0.078	0.141	0.175
Integrated (Team & Business Fundamentals)				
Human-Coded Features merged with Selected Financial Data (early fusion)	0.197	0.078	0.182	0.225
Hybrid Features merged with Selected Financial Data (early fusion)	0.219	0.078	0.207	0.25
Human-Coded Features merged with Selected Financial Data (late fusion)	0.336	0.078	0.257	0.312
Hybrid Features merged with Selected Financial Data (late fusion)	0.315	0.078	0.257	0.312

Panel B: Model Performance for 2018 (Top K: 11.2%)

Input features	NDCG	average_DV	precision_at_k	recall_at_k
Team Information				
Human-Coded Features	0.284	0.073	0.18	0.275
Hybrid Features	0.261	0.073	0.197	0.3
Machine-Coded Features	0.18	0.073	0.148	0.225
Business fundamentals information				
Selected Financial Data	0.178	0.073	0.115	0.175
Comprehensive Financial Data	0.166	0.073	0.131	0.2
Integrated (Team & Business Fundamentals)				
Human-Coded Features merged with Selected Financial Data (early fusion)	0.223	0.073	0.164	0.25
Hybrid Features merged with Selected Financial Data (early fusion)	0.233	0.073	0.164	0.25
Human-Coded Features merged with Selected Financial Data (late fusion)	0.337	0.073	0.213	0.325
Hybrid Features merged with Selected Financial Data (late fusion)	0.287	0.073	0.213	0.325

Panel C: Model Performance for 2019 (Top K: 9.2%)

Input features	NDCG	average_DV	precision_at_k	recall_at_k
Team Information				
Human-Coded Features	0.268	0.09	0.3	0.3
Hybrid Features	0.373	0.09	0.3	0.3
Machine-Coded Features	0.245	0.09	0.2	0.2
Business fundamentals information				
Selected Financial Data	0.338	0.09	0.25	0.25
Comprehensive Financial Data	0.22	0.09	0.15	0.15
Integrated (Team & Business Fundamentals)				
Human-Coded Features merged with Selected Financial Data (early fusion)	0.171	0.09	0.2	0.2
Hybrid Features merged with Selected Financial Data (early fusion)	0.205	0.09	0.25	0.25
Human-Coded Features merged with Selected Financial Data (late fusion)	0.335	0.09	0.3	0.3
Hybrid Features merged with Selected Financial Data (late fusion)	0.343	0.09	0.3	0.3

Table 10. Out-of-sample Performance Metrics of Predictive Models with Different Feature Sets (Incorporating Background Knowledge)

This table presents out-of-sample performance metrics for predicting whether a firm will achieve three-year sales growth exceeding 200%. Panel A reports the overall average performance across all test years. Panel B and Panel C provide model performance for the years 2018 and 2019, respectively. Performance is evaluated using NDCG@K, along with PRECISION@K and RECALL@K after incorporating geographic location and industry classification into the existing team and business fundamentals information feature sets. The proportion of high-growth firms in each test sample (average DV) is also included for context.

Panel A: Average Performance Metrics Over (2018-2019)

Input features	NDCG	average_DV	precision_at_k	recall_at_k
Team Information				
Human-Coded Features	0.194	0.078	0.15	0.15
Hybrid Features	0.299	0.078	0.25	0.25
Machine-Coded Features	0.205	0.078	0.2	0.2
Business fundamentals information				
Selected Financial Data	0.171	0.078	0.125	0.125
Comprehensive Financial Data	0.164	0.078	0.15	0.15
Integrated (Team & Business Fundamentals)				
Human-Coded Features merged with Selected Financial Data (early fusion)	0.176	0.078	0.2	0.2
Hybrid Features merged with Selected Financial Data (early fusion)	0.187	0.078	0.188	0.188
Human-Coded Features merged with Selected Financial Data (late fusion)	0.268	0.078	0.2	0.2
Hybrid Features merged with Selected Financial Data (late fusion)	0.277	0.078	0.262	0.262

Panel B: Model Performance for 2018

Input features	NDCG	average_DV	precision_at_k	recall_at_k
Team Information				
Human-Coded Features	0.178	0.073	0.1	0.1
Hybrid Features	0.252	0.073	0.25	0.25
Machine-Coded Features	0.164	0.073	0.2	0.2
Business fundamentals information				
Selected Financial Data	0.145	0.073	0.1	0.1
Comprehensive Financial Data	0.139	0.073	0.15	0.15
Integrated (Team & Business Fundamentals)				
Human-Coded Features merged with Selected Financial Data (early fusion)	0.187	0.073	0.2	0.2
Hybrid Features merged with Selected Financial Data (early fusion)	0.21	0.073	0.225	0.225
Human-Coded Features merged with Selected Financial Data (late fusion)	0.263	0.073	0.2	0.2
Hybrid Features merged with Selected Financial Data (late fusion)	0.248	0.073	0.275	0.275

Panel C: Model Performance for 2019

Input features	NDCG	average_DV	precision_at_k	recall_at_k
Team Information				
Human-Coded Features	0.21	0.09	0.2	0.2
Hybrid Features	0.346	0.09	0.25	0.25
Machine-Coded Features	0.245	0.09	0.2	0.2
Business fundamentals information				
Selected Financial Data	0.197	0.09	0.15	0.15
Comprehensive Financial Data	0.188	0.09	0.15	0.15
Integrated (Team & Business Fundamentals)				
Human-Coded Features merged with Selected Financial Data (early fusion)	0.166	0.09	0.2	0.2
Hybrid Features merged with Selected Financial Data (early fusion)	0.165	0.09	0.15	0.15
Human-Coded Features merged with Selected Financial Data (late fusion)	0.273	0.09	0.2	0.2
HybridFeatures merged with Selected Financial Data (late fusion)	0.306	0.09	0.25	0.25

Table 11. SHAP Value Analysis of Feature Importance by Theoretical Framework and Feature Rankings (2018–2019)

This table presents the SHAP value-based feature importance analysis for the hybrid feature model. The results compare the top predictors for the years 2018 and 2019 and highlight the consistency of feature importance rankings over time. Panel A presents the SHAP value contributions by theoretical category. Compositional features represent the sum of average member attributes, diversity measures, and faultline strength, consistent with the theoretical framework described in Section 4.1. Structural features reflect the attributes of key controlling individuals. Values are normalized SHAP contributions expressed as percentages. Panel B reports the top 10 most influential features, their theoretical classification, rankings across both years, and the corresponding changes in rank.

Panel A: SHAP Value Contributions by Theoretical Category for 2018 and 2019

Theoretical Category	2018 SHAP Contribution (%)	2019 SHAP Contribution (%)
Compositional Features (Aggregate of 1–3)	85.4	90.64
1. Average Member Attributes	44.57	56.63
2. Diversity	40.08	32.8
3. Faultline	0.75	1.2
Structural Features	14.6	9.36

Panel B: Top 10 Input Features Based on SHAP Values for 2018 and 2019

Input features	Top 10 Feature Name	Theoretical Category	Ranking in 2018	Ranking in 2019	Difference in Ranking
Hybrid Features	industry_average	Average Member Attributes	1	1	0
	prior_company_diversity	Diversity	2	2	0
	age_diversity	Diversity	3	3	0
	age_average	Average Member Attributes	4	4	0
	age_controller	Structural Features	5	11	6
	industry_controller	Structural Features	6	9	3
	education_diversity1	Diversity	7	13	6
	prior_company_average	Average Member Attributes	8	5	3
	education_average1	Average Member Attributes	9	6	3
	industry_diversity	Diversity	10	8	2
	university_average	Average Member Attributes	15	7	8
	entrepreneurial_average	Average Member Attributes	22	10	12

Table 12. Annual Distribution of NTB Firms and VC Investment Metrics

This table shows the distribution of NTB firms, the percentage receiving VC investment, and investment precision and recall. Precision measures the share of VC-backed firms exceeding 200% growth within three years, while recall reflects the proportion of high-growth firms receiving VC investment.

Year	VC-backed firms	Total firms	percentage	precision	recall
2013	18	154	0.117	0.222	0.19
2014	167	1172	0.142	0.138	0.18
2015	1012	3488	0.29	0.095	0.328
2016	1906	4919	0.387	0.061	0.393
2017	681	2066	0.33	0.047	0.33
2018	112	547	0.205	0.071	0.2
2019	40	221	0.181	0.2	0.4

Table 13. Predictive Performance of ML Models vs. VC Investment

This table compares the predictive performance of machine learning models to VC investment using PRECISION@N and RECALL@N, where N represents the number of NTB firms receiving VC investment each year.

Panel A: Average Performance Metrics Over (2018-2019)

Input features	precision_at_N	recall_at_N
Team Information		
Human-Coded Features	0.171	0.388
Hybrid Features	0.164	0.388
Business fundamentals information		
Selected Financial Data	0.155	0.363
Comprehensive Financial Data	0.122	0.300
Integrated (Team & Business Fundamentals)		
Human-Coded Features merged with Selected Financial Data (early fusion)	0.138	0.325
Hybrid Features merged with Selected Financial Data (early fusion)	0.155	0.363
Human-Coded Features merged with Selected Financial Data (late fusion)	0.190	0.450
Hybrid Features merged with Selected Financial Data (late fusion)	0.181	0.425
VC Investment		
VC investment	0.136	0.300

Panel B: Model Performance for 2018

Input features	precision_at_N	recall_at_N
Team Information		
Human-Coded Features	0.116	0.325
Hybrid Features	0.152	0.425
Business fundamentals information		
Selected Financial Data	0.134	0.375
Comprehensive Financial Data	0.143	0.4
Integrated (Team & Business Fundamentals)		
Human-Coded Features merged with Selected Financial Data (early fusion)	0.125	0.35
Hybrid Features merged with Selected Financial Data (early fusion)	0.134	0.375
Human-Coded Features merged with Selected Financial Data (late fusion)	0.179	0.5
Hybrid Features merged with Selected Financial Data (late fusion)	0.161	0.45
VC Investment		
VC investment	0.071	0.2

Panel C: Model Performance for 2019

Input features	precision_at_N	recall_at_N
Team Information		
Human-Coded Features	0.225	0.45
Hybrid Features	0.175	0.35
Business fundamentals information		
Selected Financial Data	0.175	0.35
Comprehensive Financial Data	0.1	0.2
Integrated (Team & Business Fundamentals)		
Human-Coded Features merged with Selected Financial Data (early fusion)	0.15	0.3
Hybrid Features merged with Selected Financial Data (early fusion)	0.175	0.35
Human-Coded Features merged with Selected Financial Data (late fusion)	0.2	0.4
Hybrid Features merged with Selected Financial Data (late fusion)	0.2	0.4
VC Investment		
VC investment	0.2	0.4

Table 14. Predictive Performance of ML Models (Incorporating Background Knowledge) vs. VC Investment

This table compares the predictive performance of machine learning models (incorporating background knowledge) to VC investment using PRECISION@N and RECALL@N, where N represents the number of NTB firms receiving VC investment each year.

Panel A: Average Performance Metrics Over (2018-2019)

Input features	precision_at_N	recall_at_N
Team Information		
Human-Coded Features	0.155	0.363
Hybrid Features	0.156	0.375
Business fundamentals information		
Selected Financial Data	0.138	0.325
Comprehensive Financial Data	0.134	0.325
Integrated (Team & Business Fundamentals)		
Human-Coded Features merged with Selected Financial Data (early fusion)	0.125	0.300
Hybrid Features merged with Selected Financial Data (early fusion)	0.150	0.350
Human-Coded Features merged with Selected Financial Data (late fusion)	0.175	0.400
Hybrid Features merged with Selected Financial Data (late fusion)	0.193	0.450
VC Investment		
VC investment	0.136	0.300

Panel B: Model Performance for 2018

Input features	precision_at_N	recall_at_N
Team Information		
Human-Coded Features	0.134	0.375
Hybrid Features	0.161	0.45
Business fundamentals information		
Selected Financial Data	0.125	0.35
Comprehensive Financial Data	0.143	0.4
Integrated (Team & Business Fundamentals)		
Human-Coded Features merged with Selected Financial Data (early fusion)	0.125	0.35
Hybrid Features merged with Selected Financial Data (early fusion)	0.125	0.35
Human-Coded Features merged with Selected Financial Data (late fusion)	0.125	0.35
Hybrid Features merged with Selected Financial Data (late fusion)	0.161	0.45
VC Investment		
VC investment	0.071	0.2

Panel C: Model Performance for 2019

Input features	precision_at_N	recall_at_N
Team Information		
Human-Coded Features	0.175	0.35
Hybrid Features	0.15	0.3
Business fundamentals information		
Selected Financial Data	0.15	0.3
Comprehensive Financial Data	0.125	0.25
Integrated (Team & Business Fundamentals)		
Human-Coded Features merged with Selected Financial Data (early fusion)	0.125	0.25
Hybrid Features merged with Selected Financial Data (early fusion)	0.175	0.35
Human-Coded Features merged with Selected Financial Data (late fusion)	0.225	0.45
Hybrid Features merged with Selected Financial Data (late fusion)	0.225	0.45
VC Investment		
VC investment	0.2	0.4

Table 15. Regression Results of Post-Listing VC Investment

This table presents regression results examining the relationship between post-listing VC investment and firm growth. Regression (1) evaluates whether high post-listing sales growth is associated with new VC investment, while Regressions (2) – (9) test whether ML model-predicted probabilities can explain post-listing VC investment. All regressions include control variables for team characteristics and business fundamentals at listing with year fixed effects. T-statistics are in parentheses. Statistical significance is indicated as follows: * p<0.1, ** p<0.05, *** p<0.01.

	(1) VC_3	(2) VC_3	(3) VC_3	(4) VC_3	(5) VC_3	(6) VC_3	(7) VC_3	(8) VC_3	(9) VC_3
High_growth	0.2426***								
	(3.794)								
Prob_ Human-Coded Features		0.1925*							
		(1.829)							
Prob_Hybrid Features			0.4660**						
			(2.229)						
Prob_Selected Financial Data				0.0341					
				(0.136)					
Prob_Comprehensive Financial Data					-0.0130				
					(-0.118)				
Prob_ Human-Coded Features merged with Selected Financial Data (early fusion)						0.4246			
						(1.629)			
Prob_Hybrid Features merged with Selected Financial Data (early fusion)							0.4683*		
							(1.938)		
Prob_ Human-Coded Features merged with Selected Financial Data (late fusion)								0.3340*	
								(1.730)	
Prob_Hybrid Features merged with Selected Financial Data (late fusion)									0.5598*
									(1.769)
gender_average1	0.0122	0.0100	0.0166	0.0119	0.0119	0.0114	0.0113	0.0096	0.0137
	(0.688)	(0.558)	(0.922)	(0.658)	(0.663)	(0.633)	(0.628)	(0.535)	(0.761)

age_average	-0.0469**	-0.0445*	-0.0381	-0.0537**	-0.0540**	-0.0489**	-0.0490**	-0.0451*	-0.0434*
	(-2.018)	(-1.862)	(-1.565)	(-2.292)	(-2.300)	(-2.078)	(-2.089)	(-1.890)	(-1.802)
nationality_diversity	-0.0168	-0.0183	-0.0193	-0.0183	-0.0182	-0.0181	-0.0187	-0.0178	-0.0182
	(-1.007)	(-1.087)	(-1.152)	(-1.086)	(-1.080)	(-1.074)	(-1.115)	(-1.061)	(-1.084)
education_average1	0.0508**	0.0483*	0.0508**	0.0519**	0.0521**	0.0464*	0.0463*	0.0479*	0.0498**
	(2.067)	(1.944)	(2.054)	(2.087)	(2.098)	(1.855)	(1.854)	(1.925)	(2.008)
entrepreneurial_average	0.0246	0.0297*	0.0289*	0.0287*	0.0288*	0.0276	0.0270	0.0290*	0.0279
	(1.444)	(1.730)	(1.686)	(1.672)	(1.674)	(1.605)	(1.573)	(1.692)	(1.629)
industry_average	0.0343	0.0424*	0.0466**	0.0346	0.0348	0.0346	0.0359	0.0406*	0.0406*
	(1.505)	(1.817)	(1.980)	(1.503)	(1.514)	(1.509)	(1.564)	(1.751)	(1.752)
prior_company_average	0.0173	0.0190	0.0240	0.0172	0.0169	0.0234	0.0225	0.0197	0.0229
	(0.967)	(1.050)	(1.313)	(0.945)	(0.934)	(1.265)	(1.230)	(1.088)	(1.246)
university_average	0.0113	0.0123	0.0062	0.0094	0.0092	0.0116	0.0111	0.0129	0.0090
	(0.479)	(0.516)	(0.261)	(0.392)	(0.383)	(0.485)	(0.466)	(0.537)	(0.378)
VC_average	0.0223	0.0226	0.0251	0.0277	0.0275	0.0255	0.0266	0.0235	0.0266
	(1.259)	(1.254)	(1.408)	(1.548)	(1.539)	(1.423)	(1.490)	(1.309)	(1.488)
cash_holdings	-0.0036	0.0003	-0.0000	-0.0029	-0.0017	-0.0058	-0.0055	-0.0038	-0.0071
	(-0.161)	(0.013)	(-0.001)	(-0.124)	(-0.074)	(-0.257)	(-0.248)	(-0.169)	(-0.318)
non_cash_assets	0.0248	0.0261	0.0288	0.0265	0.0255	0.0320	0.0318	0.0295	0.0334
	(1.149)	(1.199)	(1.324)	(1.182)	(1.156)	(1.446)	(1.448)	(1.349)	(1.507)
annual_revenues	0.0935***	0.0916***	0.0907***	0.0929**	0.0870***	0.1353***	0.1306***	0.1123***	0.1251***
	(3.597)	(3.497)	(3.470)	(2.268)	(2.943)	(3.488)	(3.847)	(3.804)	(3.756)
cost_of_sales	0.0002	0.0005	0.0014	-0.0022	-0.0031	0.0035	0.0060	0.0027	0.0042
	(0.010)	(0.018)	(0.054)	(-0.087)	(-0.124)	(0.139)	(0.239)	(0.109)	(0.168)
SG&A_expenses	-0.0286	-0.0324	-0.0317	-0.0310	-0.0314	-0.0250	-0.0249	-0.0306	-0.0288
	(-0.881)	(-0.992)	(-0.971)	(-0.945)	(-0.959)	(-0.760)	(-0.759)	(-0.936)	(-0.879)
Observations	737	737	737	737	737	737	737	737	737
Adjusted R-squared	0.055	0.040	0.042	0.036	0.036	0.039	0.041	0.040	0.040

Year Fixed Effect	YES	YES	YES	YES	YES	YES	YES	YES	YES
	In-sample								