Data Science in Finance and Economics

DSFE, 6(1): 1–30.
DOI: 10.3934/DSFE.2026001
Received: 03 February 2025
Revised: 27 September 2025
Accepted: 25 December 2025
Published: 13 January 2026

http://www.aimspress.com/journal/dsfe

Research article

Classifying high-potential startups for strategic partnerships using machine learning – The case of german digital startups

Dung Hai Dinh[1,*], Van Thanh Tran[1] and Nicole Ondrusch[2]

[1] Business Information Systems, Faculty of Economics and Management, Vietnamese-German University, Bến Cát, Vietnam
[2] Digital Transformation, Department of Software Engineering and Management, Heilbronn University of Applied Sciences, Heilbronn, Germany

* **Correspondence:** Email: dung.dh@vgu.edu.vn.

Abstract: Traditional methods for identifying strategic partners often rely on human judgment and informal networks, which would likely cause poor, strongly-biased choices, and missed opportunities. In this study, we aimed to support partner selection through a machine learning approach using publicly available startup data. First, we addressed class imbalance using SMOTE and generated negative labels through heuristic clustering. A Logistic Regression model was then trained to classify high-potential startups based on features such as technological capabilities, funding phase, innovation hub association, founders' structure, and geographical proximity. As for the results, the model achieved solid performance with 88.2% accuracy, 80% precision, and an F1-score of 0.727. Key patterns showed that technology prowess had the most prominent impact on defining a 'high-potential' startup partner, closely followed by their funding stage and close geographical distance to the focal company. This study provides a replicable, data-driven framework to implement into the early partner screening stage for companies, especially SMEs, along with insights on selection criteria for startups.

Keywords: classification; strategic partnerships; startups; digitalization; machine learning

JEL Codes: L26, M13, C38

1. Introduction

Strategic partnerships between companies have earned great importance over the last decade. Strategic partnerships, or strategic alliances, are formal agreements between two or more institutions to pool resources, capacity, and technology to reach a beneficial shared goal (PWC, 2009). In other words, companies are cooperating in the form of joint ventures, equity collaborations, and co-development for mutual benefits (Mockler, 1999). The right strategic alliances can help enterprises ease the friction of market entry and boost organizational effectiveness, along with significant reductions of costs (Cumming et al., 2013; Kang et al., 2014). Partnerships between companies lead to process innovations with significant relationships, especially for startups (Bettinelli et al., 2016). Moreover, collaborations between startups and corporations bring competitive advantages for both sides if done correctly, as they complement each other in different ways (Giglio et al., 2025; Rothaermel, 2001; World Economic Forum, 2018). Larger enterprises have extensive resources but adapt more slowly to environmental changes (Weiblen and Chesbrough, 2015), while startups are more flexible and innovative but struggle with resource constraints, hindering growth and development (Riepe and Uhl, 2020).

However, identifying suitable partners is challenging. The difficulty occurs from the large amount of information and research that decision-makers have to face to determine the best partner (Baiman et al., 2000). Because of the limited, imperfect, complex, and overwhelming data that entrepreneurs have to analyze, companies often turn to existing networks (e.g., previous suppliers), expert opinions, or well-known alternatives nearby for partnerships. This leads to a reliance on established networks, where information about new partners is mostly gathered through word-of-mouth and subjective judgments rather than objective, quantitative criteria (Geum et al., 2013; Lee et al., 2016). As a result, SMEs, in particular, often rely on socially and geographically biased choices when identifying potential partners due to their limited resource constraints, resulting in many opportunities being overlooked (Broekel and Binder, 2007; World Economic Forum, 2018).

Researchers have proposed several systematic approaches that can boost and optimize the partner selection process (Chang et al., 2008; Jeon et al., 2011). However, the approaches are considered inadequate (Guertler and Sick, 2021). Most researchers focus on R&D, which features mainly university-industry or research-industry collaborations (Cardamone et al., 2015; Chang et al., 2019; Eom and Lee, 2010; Giunta et al., 2016; Kang et al., 2019; Mohnen et al., 2018; Rybnicek and Königsgruber, 2019; Trela et al., 2021). Even though there are also studies on firm-firm alliances, startups can hardly fall under this category as they have highly distinct traits such as financial dependency, young and immature products, but exceptionally innovative potential (Allmendinger and Berger, 2020).

For corporate-startup relationships, there is a lack of research using quantitative models. A dominating 75% of studies on firm-startup alliances were interviews and case studies (Giglio et al., 2025). The findings were useful but prone to subjective opinions (e.g., from interviewees) and may not be applicable to broader use due to the lack of consistency and objective evaluation metrics. The determinants and parameters derived were found to influence partnerships also varied from case to case, which calls for a separate empirical study on each specific context.

Our study contributes to this research gap by proposing a quantitative and data-driven method, not to analyze collaboration outcomes, but for the partner identification and matching process in the startup ecosystem (the selection stage). Additionally, it goes beyond past research by providing a clear, repeatable method using publicly available data that can be retrieved easily at hand (e.g., Germany Digital Hub Startup Database). This is different from other approaches when researchers focus on expert opinions or stakeholder interviews, which might be more sophisticated and therefore less scalable.

In particular, in this study, we leverage real-world company data, provided by a medium-sized tech enterprise as the collaboration initiator, and startup data, provided by German federal authorities regarding their digital ecosystem. This enables the research to detect potential partners and their characteristics, or the factors that help set high-potential partners apart from their counterparts in the technology startup domain. Aside from that, it can also provide an overview of the German tech startup market, which is well-known for its active, flourishing, and innovative entrepreneurial culture. We propose a process to gather and process public data, use data manipulation methods such as SMOTE and PCA, then build a logistic regression (LR) model to classify high-potential startups. In other words, we would like to explore whether a typical machine learning method, such as LR, could support the partner selection by making it faster and more efficient. This is especially useful for SMEs, companies with stricter resource constraints that aim for a more straightforward, data-driven approach to identifying potential allies. The procedure proposed in this paper includes: Prescreening a large pool of startups, searching for potential collaboration partners with machine learning predictions based on historical data, and then providing a ranking for decision-makers to assess the most prominent partners.

The rest of this paper is structured as follows: In Section 2, we present a literature review related to the topic, elaborating on studies that revolve around factors affecting the partner selection process. In Section 3, we describe the data used and data manipulation methods. In Section 4, we present the methodology and model implementation of LR. In Section 5, we discuss the analytical results and implications. In Section 6, we provide limitations and conclude the paper.

2. Literature review

2.1. Partner selection process and heuristics problems

From the perspective of research institutions, partner selection is defined as identifying and choosing the most potential candidate to transform research results into an actual, practical product or innovation. This process can typically be summarized into four phases (Solesvik and Gulbrandsen, 2013; Wu and Barnes, 2014):

- Criteria formulation: Analyzing context and background information to establish criteria for selecting the right ally. This stage is often characterized by uncertainty due to limited information about potential partners and their strategic fit.
- Qualification: Searching for and narrowing down the list of potential by ranking them based on the previously formed criteria and matching profiles.
- Final selection: Deciding on the most suitable partner(s).
- Feedback: Evaluating the selection process and making necessary adjustments for future applications.

Two key challenges arise during this process: First, obtaining relevant and qualified data about different candidates is difficult due to its limited availability and scattered, vague, or biased nature, which introduces uncertainty (Baiman et al., 2000). Second, analyzing such large amounts of information requires significant time and effort, often overwhelming decision-makers' cognitive capacities (Trela et al., 2021). To address these challenges, collaboration initiators and their teams generally rely on heuristics, a human cognitive process that ignores part of the problem to save time and effort (Gigerenzer and Gaissmaier, 2011). Also referred to as 'local search bias' in this context, this phenomenon occurs when individuals typically avoid exhaustive searches, often favoring popular and familiar options within their nearby distances or previous partners (Gigerenzer and Gaissmaier, 2011; Solesvik and Gulbrandsen, 2013). This process can result in poor, biased choices and missed opportunities (Geum et al., 2013; Tello et al., 2010). It can also lead to path dependence (when past decisions heavily influence future actions) and social lock-in effects (over-reliance on existing networks), restricting organizations from exploring truly innovative or diverse partners. This is further complicated by labor-intensive manual processes, such as exhaustive database and web searches, which are time-consuming and inefficient (Fritsch and Kauffeld-Monz, 2010; Jeon et al., 2011). This problem is especially heightened for SMEs due to their limited resources.

These bias and effects created by heuristics are also especially stronger in startup–company partnerships compared to traditional relationships like university–industry or corporate–corporate alliances. Startups, being young and not well-known, often lack the reputations or collaboration history that companies usually based on for partner evaluation. As a result, decision-makers may miss promising but unfamiliar startups to be in favor of safer or more popular choices, narrowing their partner pool and missing out on truly innovative candidates. On the other hand, partnerships regarding research organizations and institutions tend to have more structured partner-matching procedures, specifically involving patent or license analysis (Geum et al., 2013; Jeon et al., 2011; Lee et al., 2016), and is therefore less dependent on heuristics.

Thus, we argue that Machine Learning approaches can help execute this process more effectively and efficiently. Criteria to determine partnerships can be automatically identified in phase 1 with models and notably less human effort. This data-driven method can also reduce complexity in phase 2 – searching and narrowing down the list of potential partners. With the availability of accessible data sources, the focal company can utilize methods like web-scraping to gain relevant data on other firms, build their comprehensive database, and rank candidates rapidly. Furthermore, it can also contribute to phase 4 – providing feedback to improve future suggestions (Trela et al., 2021). Even though the studies on applications of machine learning to the mentioned process are not new, they mostly focus on the fields of supply chain, R&D, or open innovation (Wu and Barnes, 2014, 2011), featuring mainly university-company or research institutions-company collaborations. Few researchers have addressed company-startup partnerships in the technology industry, despite its fast-paced innovation rhythm and great reliance on collaboration for growth (Hostettler and Bolay, 2014; Attah et al., 2024). The tech sector has always been considered a rapid-growth industry and is projected to grow at an average annual rate of 29% between 2024 and 2028 (Steve Fineberg et al., 2025). As a result, this research gap leads us to the first research objective:

Objective 1. To assess whether machine learning models, trained on partially public data, can effectively support the partner selection process in company-startup collaborations within the technology industry.

2.2. Determinants of potential partners

Effective partner selection is crucial in facilitating successful collaborations and strategic alliances. Existing literature highlights several key factors influencing this process. Technological and innovation capabilities are of utmost importance. Several studies directly emphasize the importance of R&D capacity and patents, from innovation-focused industries like biopharmaceuticals, research, and open innovation (Chang et al., 2019; Eom and Lee, 2010; Kang et al., 2019; Mohnen et al., 2018; Wu et al., 2022) to less technology-driven sectors like renewable energy (Raghuvanshi et al., 2022). Intense knowledge sharing is considered crucial specifically for firms that seek collaboration for a technology transfer (Günsel et al., 2019), and the expertise of the prospective partner should be regarded highly (Guertler and Sick, 2021). In this aspect, Mori et al. (2012) and Meulman et al. (2018) emphasized unique technology expertise and innovativeness level, once again highlighting the significance of technical innovation. Additionally, partners' innovation prowess should complement each other's needs and capabilities, with each partner bringing unique yet synergistic technological expertise to the collaboration (Meulman et al., 2018; Niederhauser et al., 2022; Trela et al., 2021).

Distance to the target company is another prominent determinant facilitating collaborations, as seen in the biopharmaceutical and technology sectors, where geographical proximity supports innovation, as it can facilitate interaction and reduce the cost of logistics (Giunta et al., 2016; Guertler and Sick, 2021; Mori et al., 2012). The availability of other key resources, such as skilled personnel, funding, and access to infrastructure, also plays a key role in determining the success of the partnership (Rybnicek and Königsgruber, 2019).

Other determinants that are not resource-related were also discussed in other research. Firm-level factors, such as size and age, can influence partnership outcomes. Larger firms with significant portfolio investments are more likely to attract collaborations due to their resource capabilities (Giunta et al., 2016). In addition, environmentally conscious practices and stakeholder engagement are becoming increasingly relevant, with studies emphasizing green resource integration and the importance of sustainable development in strategic alliances (Lo et al., 2022). Other factors like trust (Günsel et al., 2019; Meulman et al., 2018; Mori et al., 2012), marketing ability, and access to indigenous processes (Raghuvanshi et al., 2022) have also contributed to forming successful partnerships.

Based on the literature, we argue that a set of factors can help companies identify valuable and suitable tech startups in the technology industry. This forms our second research objective:

Objective 2. To determine a set of observable factors that distinguish high-potential startups from others in the German tech sector.

3. Data and data manipulation

3.1. Data description

To effectively assess partnership opportunities, it is compulsory to understand their context. For web scraping, we used Beautiful Soup which is a powerful library in Python that makes it easy for business users to scrape information from web pages. The dataset in this study was collected from *Germany's Digital Hub Initiative*, a project initiated by the German Federal Ministry for Economic

Affairs and Climate Action. This governmental program aims to promote cooperation between companies, startups, scientists, and investors to drive technological innovation by establishing Digital Hubs across Germany for potential cooperation. The hubs act as a shared ecosystem that provides conditions for startups to thrive, facilitating networking opportunities, resources, funding support, and new collaboration projects. Each hub has its own specialized field/competency with matching experts and resources, with details listed as follows:

- Hub Berlin: Internet of Things (IoT) & FinTech
- Hub Cologne: InsurTech (Insurance Technology)
- Hub Dresden/Leipzig: Smart Systems & Smart Infrastructure
- Hub Dortmund: Logistics
- Hub Frankfurt/Darmstadt: FinTech & Cybersecurity
- Hub Hamburg: Logistics
- Hub Karlsruhe: Artificial Intelligence
- Hub Mannheim/Ludwigshafen: Digital Chemistry & Digital Health
- Hub Munich: Mobility & InsurTech
- Hub Potsdam: MediaTech
- Hub Nuremberg/Erlangen: Digital Health
- Hub Stuttgart: Future Industries

Considering each hub's core technological competencies, each startup is able, but not limited, to join one or various hubs that they want in order to be exposed to the most acute trends, innovations, experts, and investors of the industry. In 2022, there were only 12 Digital Hubs with the participation of around 700 startups. In 2024, the number of Hubs increased to 25, and the number of ventures registered rose to more than 1000, showing how successful the Hubs were in boosting the entrepreneurial and innovation spirit among companies (Federal Ministry for Economic Affairs and Climate Action, n.d.).

Data regarding the startups were collected via web-scraping, which extracts information from the Web for later use, from the Digital Hub Initiative's search portal (Federal Ministry for Economic Affairs and Climate Action, n.d.). Data was gathered in December 2024, returning a dataset of 755 tech startups with information about their industry, technology expertise, customer type, target market, the associated hub, funding phase, founders, and location. The dataset features startups from various industries and technology expertise, but they fall under the 'tech' umbrella. The industries vary from Fintech to Mediatech, Education, Smart Systems, E-commerce, etc. Moreover, technology prowess can range from the Internet of Things (IoT) to Data Analytics, Robotics, Virtual Reality, etc., as described in Table 1.

The target variable (*high_potential*) was collected from an internal proprietary partner screening list provided by a medium-sized IT outsourcing company based in Frankfurt am Main, Germany (referred to as Company A throughout this study). In 2022, Company A conducted a selection process to identify promising startup partners for upcoming collaboration projects. These partnerships aimed to help the firm expand into new digital markets and diversify its service offerings by adopting new technologies. Although further details were not disclosed, the list shared with this study included startups that the company had labelled as high-potential candidates.

Table 1. Variable description.

Columns	Type	Description
Company	Nominal	The startup's name.
Industry	Nominal	The startup's industry. 16 unique values: AdTech, Cybersecurity, Digital Health, Digital Chemistry, E-Commerce, Education, FinTech, InsurTech, LegalTech, Logistics, MediaTech, Mobility, SaaS, Smart Infrastructure, Smart Systems, Cross-industry.
Technology	Nominal	The startup's technology expertise. 8 unique values: Artificial intelligence, Blockchain, Robotics, Virtual Reality, Hardware, Software Development, Data Analytics, Internet of Things.
Looking for	Nominal	What the startup is looking for. 4 unique values: Partner, Financing, Talents, Mentoring.
Funding phase	Ordinal	The startup's funding stage in Venture capital, showing the development of investment that startups typically go through. 5 unique values (in order) and their characteristics: + Pre-seed: Startups typically only have the initial concept prepared with little proof of success. Most of the funds typically come from the close network or platforms like crowdfunding. + Seed: Startups are completing their products. First stage of formal equity financing. Funds usually come from angel investors. + Early Stage: Startups most likely have established a functioning product. Phases include Series A and B. + Growth Stage: Startups are generally performing well. Phases include Series C funding rounds or further. + Later Stage: Startups likely may have functioned for five to seven years with a strong customer base and path to profitability (Ellison, 2023).
Hub Affiliation	Nominal	The digital hub associated with the startup 13 unique values: Hub Berlin, Hub Cologne, Hub Dresden/Leipzig, Hub Dortmund, Hub Frankfurt/Darmstadt, Hub Hamburg, Hub Karlsruhe, Hub Mannheim/Ludwigshafen, Hub Munich, Hub Potsdam, Hub Nuremberg/Erlangen, Hub Stuttgart, not part of the network.
Location	Nominal	The startup's headquarters locations.
Size	Ordinal	The startup's number of employees. 4 unique values: 1-10, 10-50, 50-100, >100.
Market	Ordinal	The startup's target market. 4 unique values: Germany, DACH, Europe, International.
B2B or B2C	Nominal	The startup's customer type. 2 unique values: B2B, B2C
high_potential	Nominal	The partner selection decision from Company A. 2 unique values: Yes, No.
Startup founders	Nominal	The startup's startup founders.
Website	Nominal	The startup's contact information.
Email	Nominal	

With the list in hand, we searched for these ventures in the Digital Hub Initiative dataset and got 53 matched results, representing the 'positive' samples, while the remaining 702 were treated as 'negative' instances. However, since they were not explicitly rejected by Company A, it may include false positives; startups that can potentially be chosen but were not as there was no exposure to Company A. Such problems will be addressed in the later parts (see Section 4.3).

3.2. Data preparation

First, since most of the features were categorical, we encoded them into dummy variables. A notable detail here is that five features are multi-labeled, meaning a data record (company) can have multiple values for the same attribute. For instance, under Technology, instead of having only one technology (e.g., AI) for each startup, a firm usually has two or more at the same time (e.g., DA, IoT, and AI). These variables include industry, technology, looking for, hub affiliation, and business type (B2B or B2C). Therefore, one-hot encoding was utilized here, where binary columns for each of the values were created. This scenario, however, can also increase the dimensionality of the dataset significantly. More columns mean there are more parameters in the model, which can increase complexity and make the model more prone to overfitting (Zhang and Zhou, 2014). It also makes the model harder to interpret, since we have to look at each tag separately, even though they belong to the same combined feature (Sinha, 2024). Therefore, we implemented dimensionality reduction to address this problem (see Section 4.1)

Moreover, label encoding was utilized for the remaining ordinal variables, including funding phase, market, and size, where each unique value was assigned a distinct integer value according to its natural order (Poslavskaya and Korolev, 2023).

We tackled the missing values in the dataset as follows: In general, the variables were all satisfactory, with the highest percentage of missing values reaching 13.64% and 11.52% for *funding phase* and *looking for*, respectively. After encoding, we addressed the missing values through iterative imputation with Bayesian Ridge regression. This method models the target feature as a function of other variables, creating estimations for the missing data based on statistical relationships between features within the dataset. This same process iterates through every attribute until all the nulls are imputed under the assumption that the data records are missing at random (Raghunathan et al., 2001; Rubin, 1976). Being computationally complicated, this multivariate approach ensures that complex relationships between variables are better preserved than simple imputations, such as using the most repeated values (Liu and Brown, 2013).

To further enhance the analysis process, we created four more variables. One feature was the distance to the target company called *dist_bin_no,* an ordinal variable categorizing distance from Company A into certain thresholds such as very close (0–20 kilometers), close (20–100 kilometers), commutable (100–200 kilometers), and far (more than 200 kilometers). To achieve this, the geographic coordinates of each company (city-wise) were retrieved from OpenStreetMap, then relative distances were computed based on geodesic calculations. Another variable derived is the number of headquarters. Moreover, the other two features were about the founders' team and their female proportion, including *num_of_founder* and *percent_female*. To get this, a list of 1622 founders was first extracted, and then we utilized a simple model predicting the most likely gender from a particular name using a gender predictor package. With the error rate assumed to be around 2% for Western names (Sebo, 2022), 1373 male and 143 female instances were detected, representing an extreme disproportion in female tech leaders –9.6 males to 1 female. The remaining 106 unknown gender predictions were imputed randomly with the same ratio, with which we calculated to get the *percent_female* of each startup's founder team.

3.3. Exploratory data analysis (EDA)

Figure 1 displays an overview of Germany's tech startup dataset. First, it can be observed that the SaaS (Software as a Service) industry has an overwhelming number of startups (over 300 – nearly half on the market), which might also explain the exceedingly high number of ventures specializing in software development.

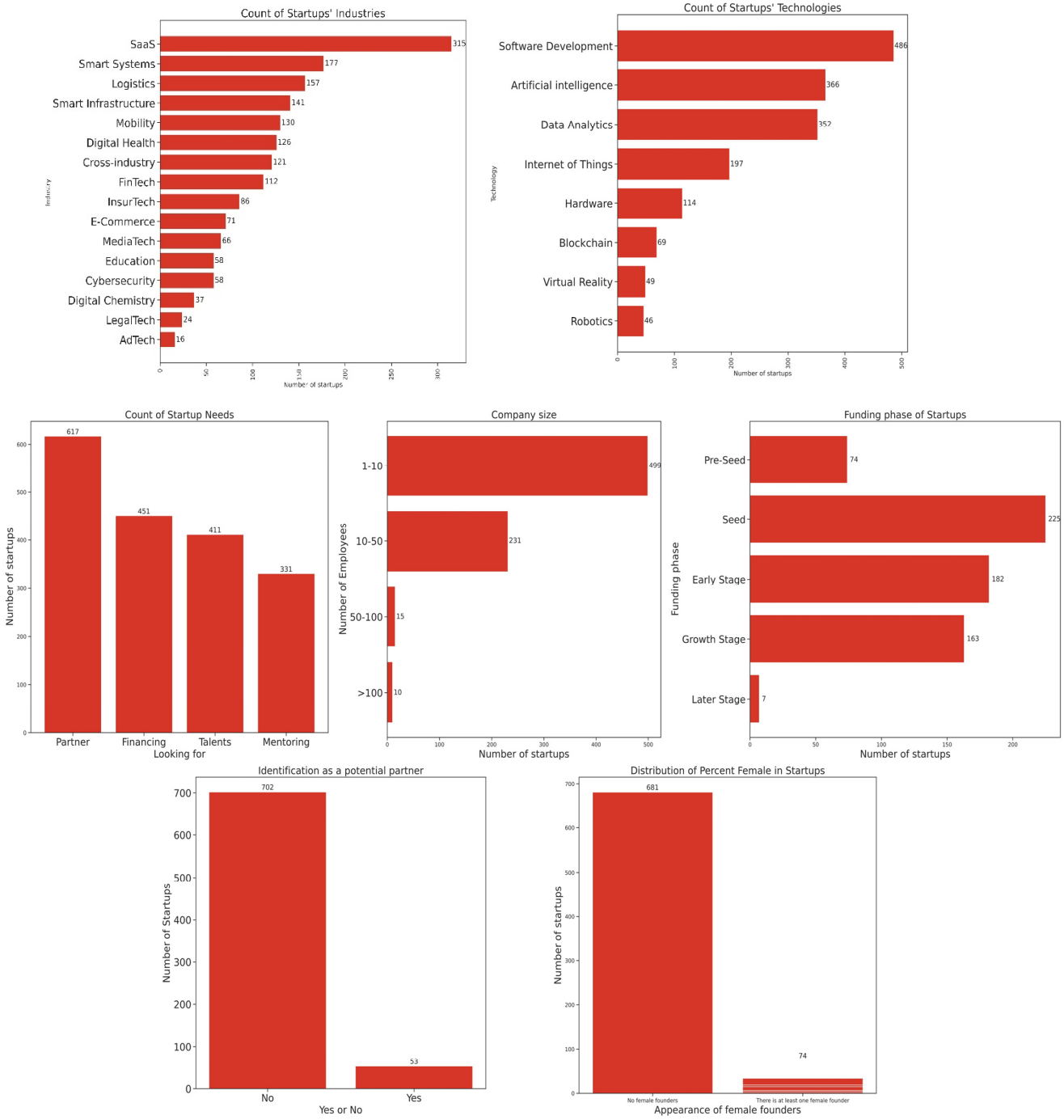

Figure 1. Descriptive characteristics.

Other popular technologies include artificial intelligence (AI) and data analytics (DA). Additionally, looking at the staggering number of small to extremely small firms (96.7% contain less than 50 employees), interestingly, financing is not their first priority since it is well-known that most

startups tend to fail due to insufficient venture capital, especially for high-tech firms (Öndas, 2021). Instead of funding, their top need is to search for potential collaborations (taking up 82%). This greatly stresses the importance of strategic alliance research for startups, supporting their vulnerabilities at the early stages of a business. This need is emphasized once more due to the large number of young firms on the market, with 40% only in their pre-seed and seed funding stages. Additionally, the dataset also showed the extreme disproportion of female leaders in the high-tech domain, with only 74 among 755 firms (9.8%) having at least one female in their founder team. Finally, the number of high-potential startups identified by the company until this stage was 53, showing how heavily unbalanced the dataset was. Appendix A1 provides further demographic characteristics.

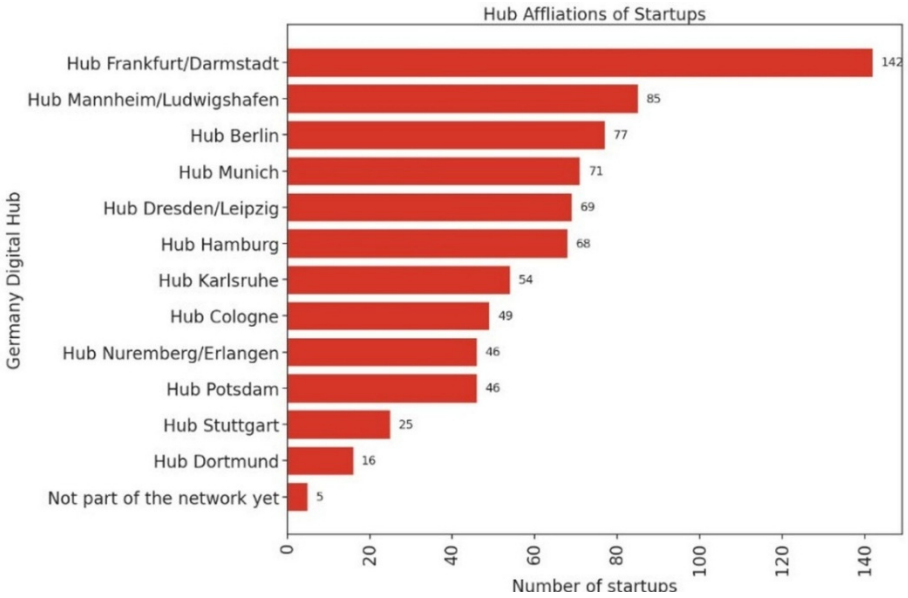

Figure 2. Hub affiliations of startups.

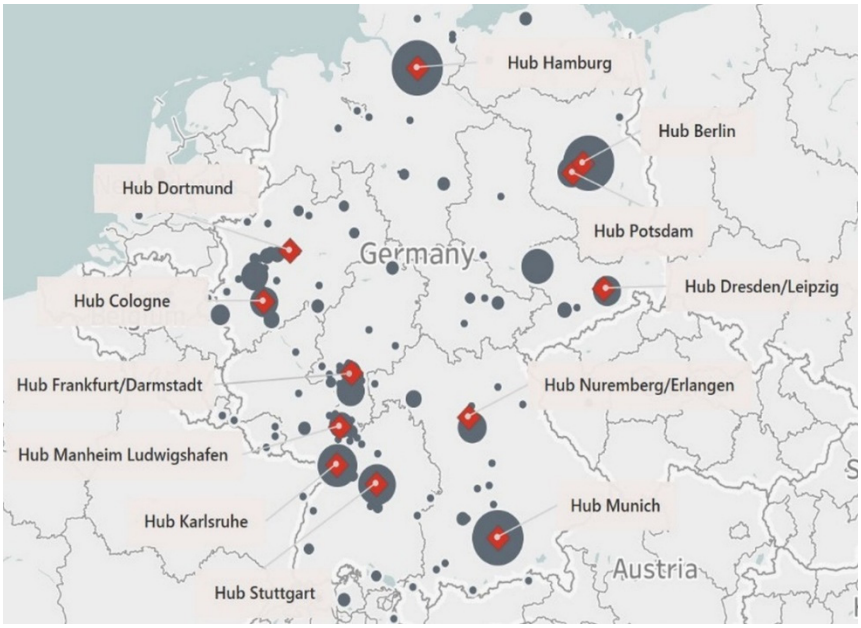

Figure 3. Count of Startups, their german digital hubs and geographic distribution. (Map source: © 2025 Mapbox, OpenStreetMap contributors. Visualization created in Tableau.)

4. Methodology and models

In this section, we describe the machine learning models and techniques used in this study, including the rationale for choosing them, their theoretical principals, parameter settings, and evaluation strategy. The steps involved: Feature selection using Chi-Squared Test, dimensionality reduction using PCA, negative sampling using heuristic-based clustering, addressing imbalance using SMOTE, model training, classification with LR, and evaluation metrics. These steps align with standard machine learning procedures in many studies, such as in Li & Yan (2025). Additionally, SMOTE, model tuning, and interpretability-focused techniques are used to discover patterns in imbalanced datasets.

4.1. Feature selection (chi-squared test)

Table 2. Independence test results (Chi-square) of features with the target variable.

Feature	Chi2	P-value	Feature	Chi2	P-value
percent_female	72.18	<0.001	Hub Munich	0.85	0.357
Hub Karlsruhe	29.57	<0.001	Hub Hamburg	0.75	0.385
Internet of Things	23.78	<0.001	InsurTech	0.69	0.407
funding_no	18.93	<0.001	MediaTech	0.62	0.431
Artificial intelligence	12.43	<0.001	Mobility	0.53	0.466
Smart Systems	11.60	0.001	FinTech	0.47	0.491
b2c	6.92	0.009	Hub Berlin	0.39	0.531
Data Analytics	5.90	0.015	Not part of the network yet	0.38	0.539
dist_bin_no	3.90	0.048	Software Development	0.34	0.558
Hub Potsdam	3.47	0.062	b2b	0.34	0.561
Financing	3.40	0.065	Education	0.30	0.582
Logistics	3.28	0.070	E-Commerce	0.22	0.637
Hub Frankfurt/Darmstadt	2.72	0.099	num_of_founder	0.16	0.688
Digital Chemistry	2.39	0.122	Blockchain	0.16	0.691
Hub Nuremberg/Erlangen	2.38	0.123	Partner	0.14	0.703
SaaS	2.15	0.143	Smart Infrastructure	0.13	0.716
num_of_female	2.11	0.146	Virtual Reality	0.10	0.754
Hub Cologne	2.05	0.152	Cross-industry	0.03	0.857
LegalTech	1.81	0.178	Hub Stuttgart	0.02	0.893
Mentoring	1.76	0.184	AdTech	0.01	0.904
size_no	1.64	0.201	num_headquart	0.00	0.955
market_no	1.28	0.259	Hub Mannheim/Ludwigshafen	0.00	0.988
Hub Dortmund	1.21	0.272	Talents	0.00	0.993
Cybersecurity	1.13	0.287	Digital Health	0.00	0.996
Robotics	1.04	0.307	Hardware	0.00	0.999
Hub Dresden/Leipzig	0.95	0.329	num_of_hub	0.00	1.000

We first explored the features using a Chi-squared test to assess their independence from each other and also with the target variable. The goal was to retain only the variables contributing significantly to the predictions, eliminating irrelevant variables. This helped in enhancing the overall performance, eliminating noise and keeping it a simple, interpretable model. Table 2 shows the independence evaluations of all variables in the dataset with the target variable, measured by the chi-square test. The chi-square test is a statistical method used to determine if there is a significant association between two categorical features. A lower p-value (under 0.05) suggests that the two variables are highly associated with one another. This means that the variables examined have a statistically significant relationship and not by mere chance.

From the table, among 53 variables, only nine display a stronger association with whether or not a startup is considered high-potential for partnerships, with a p-value under 0.05. These variables include distance to the company (in categories), percentage of females in leadership, startups from Hub Karlsruhe, funding stage, B2C type of business, operating in smart systems industries, and technology expertise of IoT, AI, or DA. In other words, the relationship between these variables and the target feature is supported statistically and not by random chance. Therefore, only these nine factors were chosen for modeling and analysis since we wanted to prioritize interpretability and simplicity, focusing more on explaining the variables behind cooperation.

4.2. Dimensionality reduction - PCA

After selecting significant features, we examined the strength of the association between the variables using the Cramer's V measure (Table 3). Its value ran from 0 (no association) to 1 (perfect association). A number closer to 1 expresses a stronger relationship. The target feature "high_potential" was most strongly associated with "funding_no" at 0.27 and "dist_bin_no" at 0.26, indicating that these two factors are particularly relevant in defining whether a startup will have a high potential for future collaborations. Another notable observation was that "IoT" and "Smart Systems" were highly correlated by 0.50. Similarly, "AI" and "Data Analytics" were moderately correlated by 0.35. Such numbers reflect how these two pairs may display an overlap for the first pair (as IoT is the representative technology for the Smart Systems industry), or complementarity for the second pair (as DA serves as the fuel for AI models to function well) regarding the technology expertise of a startup.

These findings raise multicollinearity concerns, which can greatly affect regression estimates. To address this, Principal Component Analysis (PCA) was applied to two feature pairs: (1) Internet of Things and Smart Systems and (2) Artificial Intelligence and Data Analytics, since both showed abnormal levels of association (0.50 and 0.35) compared to the remaining pairs fluctuating steadily around 0.05–0.2. PCA reduces data redundancy by combining correlated features together into uncorrelated components, avoiding overlap (Richardson, 2009). Even though this step sacrifices interpretability to some extent since these direct, raw features with meanings are dropped, it was necessary. Without PCA, our earlier version of the model produced conflicting results as one variable's coefficient flipped signs, resulting in a negative coefficient even though it showed strong positive association with the target attribute in the independence test. Therefore, applying PCA resulted in two principal components: "IoT_SmartSystems_PC1" and "AI_DataAnalytics_PC2," replacing the original overlapping features.

Table 3. Association levels between variables (Cramer's V).

	percent_female	Hub Karlsruhe	IoT	Funding no	AI	Smart Systems	b2c	DA	dist_ bin	High_ potential
percent_female	1.00									
Hub Karlsruhe	0.05	0.99								
Internet of Things	0.07	0.06	1.00							
funding_no	0.09	0.08	0.14	1.00						
Artificial intelligence	0.07	0.13	0.07	0.07	1.00					
Smart Systems	0.08	0.16	0.50	0.13	0.09	1.00				
b2c	0.11	0.06	0.10	0.17	0.10	0.10	1.00			
Data Analytics	0.06	0.03	0.12	0.15	0.35	0.06	0.07	1.00		
dist_bin_no	0.09	0.29	0.13	0.14	0.13	0.13	0.02	0.14	1.00	
high_potential	0.06	0.20	0.20	0.27	0.18	0.14	0.11	0.12	0.26	1.00

Despite the relatively strong association between "dist_bin_no" and "Hub Karlsruhe" (0.29), these two variables were kept as individual features in the model. The reason was that "dist_bin_no" measured the distance to company A, categorized into ordinal ranks, while "Hub Karlsruhe" was merely one of the 12 hubs. While both variables had a strong association with the target variable (0.2 for Hub Karlsruhe and 0.26 for dist_bin_no), they represented different aspects of its impact and were not directly correlated to each other. Therefore, they were not combined to keep the interpretability straightforward. As a result, there were only seven features, percent_female, funding_no, b2c, IoT_SmartSystems_PC1, AI_DataAnalytics_PC2, Hub Karlsruhe, and dist_bin_no, that made up the model.

4.3. Negative sampling and deduplication – heuristic-based clustering

As only positive samples were available, we adopted a weak supervision approach to produce negative labeled instances using heuristic-based clustering. Weak supervision is a technique that helps annotate data based on the available signals and correlations between variables, eliminating the need for manual labeling by assuming that these signals somewhat reveal the data outcome. This saves time and effort, and can sometimes increase predictive power compared to hours of hand labeling (Ratner et al., 2017). In our case, we first cleaned and simplified the dataset using association analysis, feature selection, and PCA, with the expectation of reducing redundant information and creating a sample space where variables revealed their true relationships. Then, we used the K-Means model to group startups into clusters based on their positions in the feature space. Clusters that contained no positive cases, or very few (maximum 5% to ensure sufficient samples to feed the models), were treated as negative samples. This helped reduce noise from potential false negatives based on the assumption that startups with little similarity or correlation to the partner list were more likely to be negative. This assumption was similar to the concept behind Multiple Instance Learning (MIL), where a label was assigned to a group (or "bag") based on whether it contained at least one positive sample (Poyiadzi et al., 2022).

After negative sampling, we obtained a total of 495 instances, including 53 positives and 442 negatives. However, these numbers were brought down to 403 records, including 53 positives and 350

negatives, after we removed duplicates to prevent data leakage. This step was necessary because, after feature selection, the dependent variables were reduced to seven, significantly increasing the likelihood of samples repeating each other. Keeping these duplicates could lead to model overfitting, as it could learn to memorize records rather than deriving patterns, resulting in overly optimistic results.

4.4. Addressing imbalance - SMOTE

Synthetic Minority Oversampling Technique (SMOTE) was applied to tackle the class imbalance in the dataset. SMOTE works by creating synthetic examples for the minority class rather than simply duplicating existing ones. It achieves this by identifying the nearest neighbors of each minority instance in the feature space and generating new data points along the line segments between them (Chawla et al., 2002). This approach helps balance the class distribution while preserving the diversity of the minority class, leading to more robust and unbiased model performance.

Although LR is widely used for its interpretability, it tends to struggle under imbalanced conditions compared to more robust models such as random forests or boosting methods (Dube and Verster, 2023). In particular, class imbalance raises a significant challenge in constructing classification models, as they tend to become biased toward the majority class, leading to poor performance in detecting minority classes, or high-potential startups in this specific case. As the class distribution is heavily skewed, models tend to simply predict the majority class for all instances, making them highly-conservative (He and Garcia, 2009). Therefore, applying SMOTE is highly necessary. This method was applied only to the training set to prevent data leakage. It oversampled the minority class to reach a ratio of 1:1 (280:280) between positives and negatives while keeping the original distribution in the test set to reflect real-world scenarios.

4.5. Training process

First, a ratio of 75:25 train-test split was applied to build the training and test set. This test set was kept out of all training steps and used to report final performance metrics.

Then, the training set applied a 5-fold cross-validation approach (due to the small sample size), where the dataset was divided into 5 equal parts. In each iteration, 80% of the data was used for training while the remaining 20% was used as the validation set. This method helped assess the models' reliability to see whether the predictions' accuracy was stable or not.

4.6. Modeling-LR

Non-linear models like neural networks have gained considerable attention due to their robust performance in capturing relationships. However, while these models can handle large, high-dimensional datasets efficiently, they often require a substantial amount of data. In our case, it would be a misfit to use neural networks because of the limited sample size (280 records entering the model) when these powerful models require at least thousands of records with many features (high-dimensional data) to fully take advantage of their benefits (Frei et al., 2022).

On the other hand, traditional linear models such as LR remain a reliable choice in business applications, especially when interpretability and small sample sizes are key problems. LR offered three key benefits in this study's context. First, it provided a straightforward and interpretable table of coefficients explaining the meanings behind classification attributes, aligning with our key objective to understand decision-making patterns. Second, it could generalize well even under a limited number of samples. Finally, because LR is a probabilistic classification technique, it could rank startups for easy comparison, sorting out the most to least potential ventures (Joseph M. Hilbe, 2015). The LR model was trained using the scikit-learn library in Python with default hyperparameters, except for max_iter, which was set to 1,000 to ensure convergence.

To have a more comprehensive assessment, we also included the performance of three other widely-used models: Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost). These models were selected due to their robust performance in various classification tasks. Comparing LR predictability against more complex models, as such, could enable us to evaluate the predictive power of LR further while considering the trade-offs between a model's interpretability and complexity.

4.7. Evaluation metrics

Model performance was assessed using these metrics:

- Accuracy: The overall correctness of predictions.
- Precision: The proportion of positive predictions that were actually correct.
- Recall: The proportion of actual positives that were correctly identified.
- F1-score: The harmonic mean of precision and recall.
- ROC-AUC: Measures the ability to distinguish between classes across thresholds.

We computed standard performance metrics, including precision, recall, accuracy, and f1-score. Precision indicated the 'quality' of our classification, while recall showed the 'quantity' of all potential startups we could detect. This meant that, on average, when the model predicted whether a specific startup had potential, precision told us how accurate this prediction was in percentage. On the other hand, a high recall showed that our model could identify a large number of potential partners. Since the business was in the pre-screening startup stage, where the goal appeared to be filtering out the most promising startups from a large pool of candidates, precision was prioritized over recall. This assumption was based on the understanding that the resulting list of startups had to go through more detailed manual investigation from Company A in the next stage, and an unnecessary amount of non-potential partners could increase the time and effort required to do so. Under such context, the accuracy of each prediction (precision) may have been more crucial than identifying all potential candidates (recall). However, a minimum acceptable level of recall was still important to avoid missing too many candidates. Therefore, in this study, a recall threshold of 50% was assumed to be a reasonable threshold for us to assess the model, or at least half of the promising startups should have been detected, while precision remained the primary metric.

5. Results and discussion

5.1. Models' performance

Table 4 displays the relative performance of four models: LR (the main model), RF, SVM, and XGBoost at the baseline decision threshold (0.5). In general, on the test set, LR exhibits a solid balance between precision (0.800) and recall (0.667), resulting in an F1-score of 0.727, and ROC AUC and accuracy of 0.882. In other words, achieving an accuracy of 0.882 means that it correctly classifies nearly 90% of all startups combined in the dataset. The precision attained is 0.80, which indicates that 80% of the startups predicted as high-potential by the model are high-potential, showing the 'quality' of each prediction. The recall score is 0.667, showing that the model successfully identifies 66.7% of actual high-potential startups, representing the 'quantity' detected by the model. A more balanced measure between precision and recall would be an F1-score of 0.727, indicating a solid overall performance. Finally, the ROC AUC value of 0.882 shows that the model effectively distinguishes the high-potential and non-high-potential classes.

Table 4. Performance of four models.

Metric / Model	RF (test set)	XGBoost (test set)	SVM (test set)	LR (test set)	LR (validation mean)	LR (validation st.dev)
Precision	1.000	1.000	0.719	0.800	0.760	0.109
Recall	0.467	0.533	0.767	0.667	0.764	0.122
F1-score	0.636	0.696	0.742	0.727	0.756	0.089
ROC AUC	0.901	0.901	0.881	0.882	0.922	0.041
Accuracy	0.840	0.860	0.840	0.882	0.945	0.020

On the other hand, RF and XGBoost achieve perfect precision (1.000), but come with lower recall scores of 0.467 and 0.533, suggesting that these models might be overly conservative, identifying only obvious positive cases to achieve the perfect precision metric, and therefore resulting in such low recall. Moreover, SVM achieves the highest recall (0.767) but has the lowest precision (0.719), which is undesirable for our partner-matching case scenario. Therefore, although XGBoost is the model with the best metrics here, LR remains a competitive and more well-rounded alternative due to its (1) overall second-highest F1-score, showing a better overall trade-off between precision and recall, (2) other metrics being equally comparable (highest accuracy, second-highest recall, or ROC AUC showing merely a 1.9 percentage point difference compared to the highest), (3) models not being overly conservative, and especially (4) when its precision can be improved through further threshold tuning (see Section 5.2). Therefore, even though it is best known for its interpretability, LR's performance has proven to be solid and can be implemented in real life.

Additionally, validation results from five-fold cross-validation also support LR's consistency and robustness. On average, LR achieves a precision of 0.760 and recall of 0.764 across folds, with an F1-score of 0.756 and ROC AUC of 0.922. The standard deviations (e.g., ±0.109 for precision, ±0.122 for recall) indicate moderate variability across folds, which is expected considering the small sample size. To improve further, threshold tuning steps can be added.

Figure 4 shows the performance of LR by classification thresholds. From the figure, recall remains high in lower thresholds, while precision starts low but gradually increases. The two curves intersect at around 0.40–0.45, which is a point where the model's forecasts begin to lean toward precision over recall. Beyond this threshold, the model becomes increasingly conservative, generating fewer positive predictions and therefore missing more actual positives.

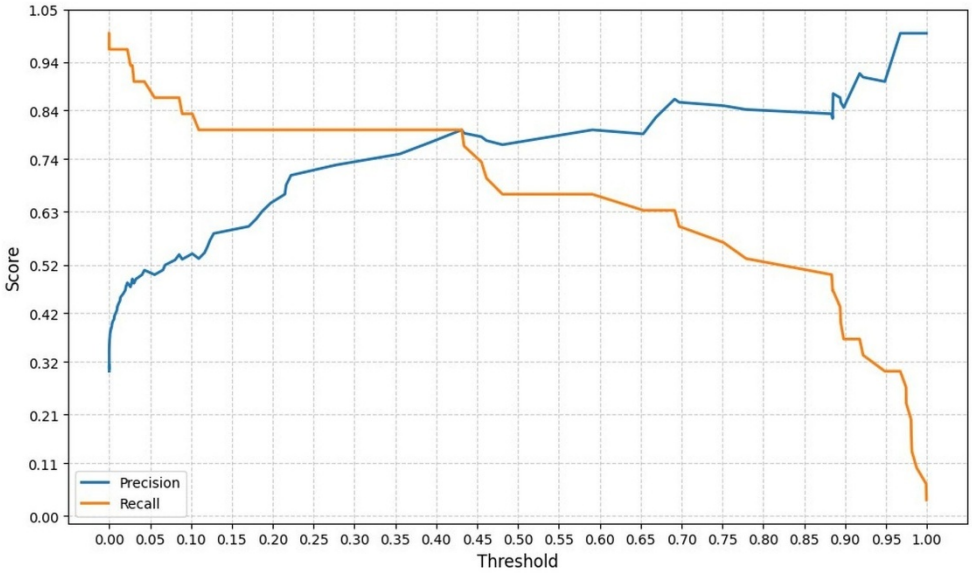

Figure 4. Precision and recall (LR) across different thresholds.

Based on this pattern, businesses can adjust the threshold to tailor their model to their operational goals. For example, in cases where precision is prioritized, such as when the cost of false positives becomes significant (e.g., wasted investigation and due diligence efforts on misfit partners), higher thresholds (e.g., ≥ 70) could yield better results. In contrast, if the enterprise's goal is to flag as many candidates as possible, lower thresholds (e.g., 0.30) would be more beneficial. In the case of Company A, for instance, where precision is prioritized with a minimum recall threshold of 50%, the company can go as high as around threshold 0.69, with precision reaching 85% and recall at 63%. With these settings, Company A can ensure 85% accuracy for each startup prediction and 63% potential startups being detected.

5.2. Impact of SMOTE

Additionally, these performance results also highlight the effectiveness of SMOTE in handling imbalance and improving predictions (Table 5). Without SMOTE, the LR model shows perfect precision (1.00) but only 53.3% recall, meaning nearly half the actual positive cases are missed. This is rather common of models trained on imbalanced data, which tend to be conservative and overly avoid false positives, leading them to skip many true positives. After applying SMOTE to the training set, recall improves moderately to 66.7%, while precision remains high at 88.0%, making the F1-score

rise from 0.696 to 0.727 and an improvement in ROC AUC from 0.797 to 0.882, showing overall better model performance.

Table 5. Performance of LR, before and after SMOTE (Threshold 0.5).

	Prediction					
Actual	Logistic regression			Logistic Regression + SMOTE		
	0	1	Total	0	1	Total
0	35	1	36	64	6	70
1	7	0	7	8	22	30
Total	42	1	227	72	28	100
Precision			1.000	Precision		0.800
Recall			0.533	Recall		0.667
F1-score			0.696	F1-score		0.727
ROC AUC			0.797	ROC AUC		0.882
Accuracy			0.860	Accuracy		0.882

The results strengthen the importance of addressing class imbalance. In this context, high precision is crucial for Company A to avoid wasting valuable resources on unsuitable candidates. Furthermore, the recall improvement enhances the model's ability to detect a larger portion of promising startups. Such findings reinforce the literature, although to a smaller extent, as SMOTE increases the specificity level of a model by 16 percentage points (Islahulhaq and Iis, 2021), 24.6 points for precision, and up to 70 points for recall of the other models (Wongvorachan et al., 2023).

5.3. Coefficient analysis

Table 6 displays the coefficients derived from the LR model for the most significant variables and their analysis. From the coefficients, five major aspects appear to significantly drive the partnership decision from company A, including technical expertise, funding stage, location, and business type of the startup.

Table 6. Logistic regression analysis of classification variables.

Feature	Variable	Coefficient	Std. Error	Odds Ratio	P-Value
Funding	funding_no	2.5808	3.372E-01	13.2078	0.0000
Technology	IoT_SmartSystems_PC1	2.9878	6.746E-01	19.8424	0.0000
Technology	AI_DataAnalytics_PC2	3.641	5.496E-01	38.1307	0.0000
Business Type	b2c	−1.5406	5.256E-01	0.2143	0.0034
Location	dist_bin_no	−0.7589	3.532E-01	0.4682	0.0317
Digital hub	Hub Karlsruhe	24.7601	3.606E+05	5.66E+10	0.9999
Founder	percent_female	−3.0365	3.217E+12	0.048	1.0000
	const	−6.5283	1.486E+00	0.0015	0.0000

First, technology know-how appears to be one of the most significant factors. The features *IoT_SmartSystems_PC1* and *AI_DataAnalytics_PC2* show remarkably strong odds ratios 19.842 and 38.131, respectively, for their influence on the target variable. This indicates that startups that focus on IoT and Smart Systems are nearly 20 times more likely, while those specializing in AI-data analytics witnessed a staggering 38 times higher chance of being considered as a potential partner. Notably, all

these technologies represent general-purpose technologies (GPTs), pervasive technologies that possess inherent potential for technical improvements (Aithal and Aithal, 2018). The finding reveals not just how important these technologies are to the company in question but also how strongly it prioritizes technology resources over other factors. Even though the types of technologies (e.g., IoT, AI, and DA) cannot be generalized to the majority, the overwhelming effects of technical aspects highly emphasize the importance of technologies in forming strategic partnerships.

This finding is in consensus with findings in the literature, which confirmed the crucial role of technological innovation (Guertler and Sick, 2021; Li et al., 2019; Mori et al., 2012; Yan et al., 2024) and the level of innovativeness of partners (Meulman et al., 2018) in business alliances. Not only is this factor crucial in the tech domain, it is also significant in the renewable energy sector (Raghuvanshi et al., 2022) and the field of open innovation (Wu et al., 2022). However, this sharply contrasts with Kuan et al.'s (Daiy et al., 2021) finding that technology is considered the least important in partner selection in the open banking industry, ranked below factors such as financial regulations, market scale, and rapid service delivery, showing how various contexts can influence the partner selection criteria. Other than that, the firm's prioritization of GPTs further confirms literature on their spillover impacts on various functionalities of a firm (e.g., productivity gains, quality of services, innovation, and customer satisfaction) (Desalegn et al., 2024; Gonzales, 2023), which applies to latecomer SMEs in this digitalizing frontier (Kopka and Fornahl, 2024). For partner firms, selecting such GPT-focused startups represents not only harnessing technological innovation but also targeting significant and robust spillover effects.

Second, a startup's financial metrics have appeared to be significantly important. The variable *funding_no*, representing the startup's funding stage, has a positive coefficient of 2.5808 and an odds ratio of 13.208. This indicates that startups in more mature funding phases are over 13 times more likely to be classified as high-potential. As shown in Figure 5, which shows the proportions of high-potential startups by funding phase after SMOTE resampling, within each extremely early phase, like pre-seed or seed, only 22.8%–25.9% are considered high-potential. Conversely, these numbers are at least twice as high for the three later stages. Economically speaking, this suggests that Company A values financial stability, aligning with other studies that ventures after being financed tend to perform better (Partal and Gönel, 2025), have more resources to implement changes through collaboration (Cardamone et al., 2015), and are solid proofs of operational stability as they have gone through various rounds of due diligence stages associated with venture capital investing. In this case study, which involves more high-risk components such as startups and the volatile technology domain, financial factors are especially important even in the first partner screening round due to how prone to failure these firms are in the early stages of their life cycle (Öndas, 2021; Patel, 2015). Such findings align with Raghuvanshi et al. and Trela et al.'s research, revealing that affordability (Raghuvanshi et al., 2022) and economic metrics (Trela et al., 2021) matter when screening for partners. However, it this is not mentioned in most studies. This is likely because it is considered in much later rounds after companies have sorted out most of the candidates.

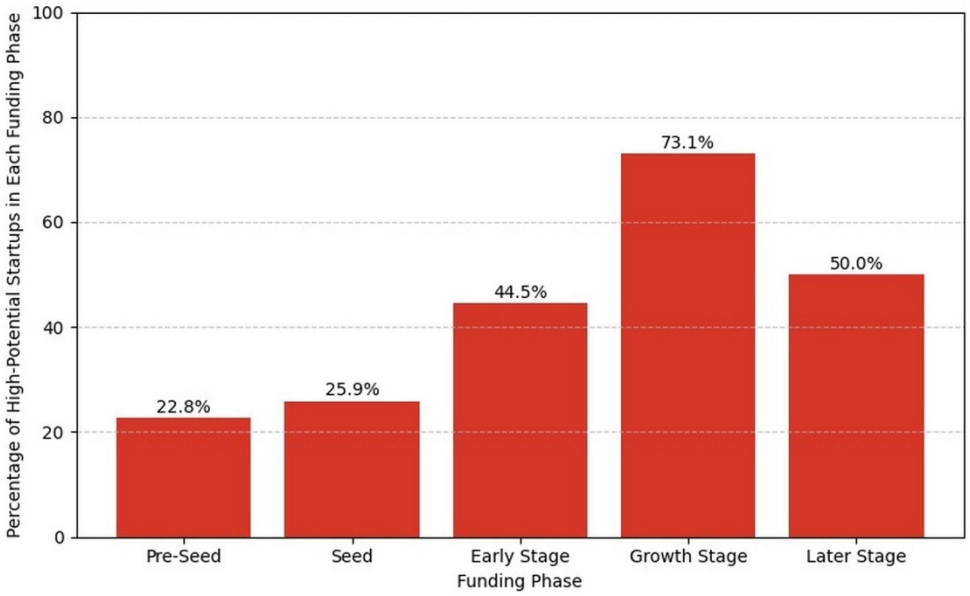

Figure 5. Proportion of high-potential startups by funding phase.

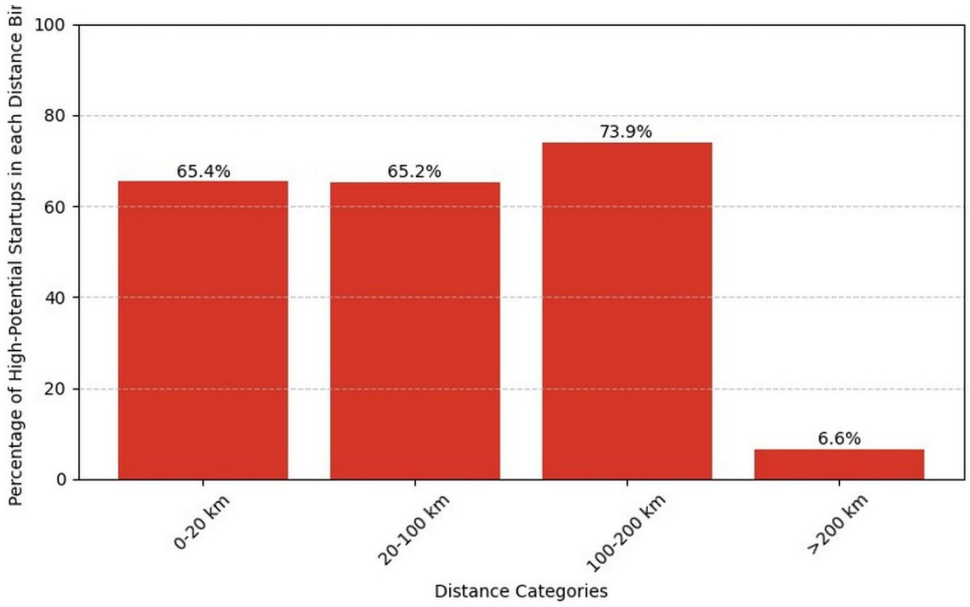

Figure 6. Proportion of high-potential startups by distance.

Third, geographic factors also play a major role in determining partnership potential. The variable dist_bin_no has a coefficient of −0.7589 and an odds ratio of 0.4682, indicating that an increase in distance reduces the likelihood of a startup being identified as high-potential by 53.2%. This aligns with other studies that distance weakens the effect of collaboration (Guertler and Sick, 2021; Hu et al., 2024), reflecting transaction cost economics (Williamson, 1979): Having alliances within a short distance minimizes transportation costs, logistical friction, and monitoring expenses (Mori et al., 2012), and results in "local search bias" (Fritsch and Kauffeld-Monz, 2010; Jeon et al., 2011; Meulman et al., 2018). In this case, the cut-off threshold is 200 kilometers, after which the chance of being considered

a suitable candidate significantly decreases. Among the startups further from this threshold, the company moves on with only 6.6% of them (Figure 6). This finding can also be observed in Figure 7, where the most promising candidates mostly cluster around Company A within a maximum of 200 kilometers. When mapped, it appears that almost all strictly lie inside Germany, indicating that cross-country alliances are less likely to be formed. This pattern demonstrates a strong preference for domestic partners, potentially because of differences in regulations, culture, or language.

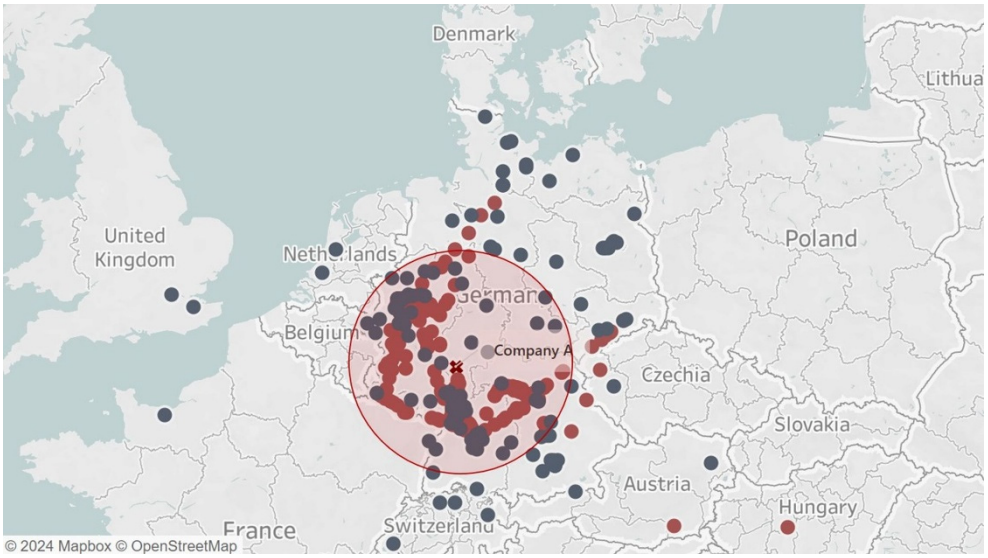

Figure 7. Distribution of high-potential startups (red data points) around company A. Map source: © 2025 Mapbox, OpenStreetMap contributors. Visualization created in Tableau.

Finally, another significant factor is business type (variable *b2c*), indicating the suitable customer type that the company is looking for. In this case, with a coefficient of -1.5406 and an odds ratio of 0.2143, the company appears to avoid B2C businesses and wants to find B2B firms instead. This preference represents a rational economic drive: B2B startups tend to have more stable revenue models, therefore reducing uncertainty. On the contrary, the feature *percent_female* does not appear to have a meaningful influence on the target variable after being added to the multivariate model. This is an interesting finding showing that its incremental effect has been fairly small, despite the high association with the target variable when being compared one-to-one. In other words, gender does not appear to affect partnership decisions, even though male leaders heavily dominate the German startup market.

In the same manner, the variable Hub Karlsruhe appears to have a very large coefficient (24.7601) and odds ratio (5.66e+10), but it is also not statistically significant with a p-value near 1.0. In other words, it is possible that startups registered in this hub could have a higher chance of being selected simply due to the hub's technological theme. Hub Karlsruhe is known to be Germany's AI hub, which aligns with Company A's interest in AI-focused startups, but this effect has been reflected in the variable *AI_DataAnalytics_PC2*. Although this variable is not significant in the current model, its logic aligns with other research for the consideration of indirect resources. Startups here might benefit from better access to technical resources or investor networks in the AI-DA ecosystem. For example, collaboration studies in the biopharmaceutical industry show that proximity to universities or

innovation clusters enhances the attractiveness of a potential partner due to indirect access to shared resources (Giunta et al., 2016). In other words, federal initiatives like these digital hubs can influence which startups gain visibility and resources, similar to how public policy shapes industry development in other contexts (Sun et al., 2024).

These results fulfill our second objective: Four major aspects forming a 'potential startup' include technological prowess, financial maturity, suitable business type, and geographic accessibility. In this case, startups with technologies like IoT, Smart Systems, AI, or Data Analytics are significantly more attractive due to their innovation resources. However, the technologies considered would depend on each company's unique strategic objectives (Meulman et al., 2018; Niederhauser et al., 2022).. Finally, startups with solid financial backing are preferred due to how financially vulnerable startups normally are (Öndas, 2021; Patel, 2015).

5.4. Implications

Our findings and interpretations are especially useful for startups that need collaborations to support and further expand their young business. From this dataset alone, 617 of 755 startups (approximately 82%) explicitly state that they are looking for partners, indicating the great demand from young firms for alliances. From the findings, ventures in the digital sector can focus on developing their technological expertise, funding stability, and the complementarity of their resources with what the potential partners are searching for (e.g., customer type). A key implication is how startups can focus on looking for allies around their geographical area since close and commutable distance is a key factor in shaping a potential partner. By understanding these factors, startups can better position themselves and focus on important areas.

The study also has practical applications for companies searching for partnerships. We utilize easy-to-use, straightforward methods like LR and publicly available information, which can be obtained from the web through methods, like web scraping, as used in this study. This method also minimizes bias coming from subjective local searches and guides companies to refine their selection criteria. First, from data on partnerships, this approach will automatically identify and return a list of key criteria. Second, it provides a ranked list of potential startups, ordered from most to least potential based on the probabilities of LR, which measures the likelihood of startups being suitable candidates, as shown in Table 7.

Table 7. Startup list ranked by predicted likelihood of being high-potential.

percent_ female	funding_no	Hub Karls-ruhe	b2c	dist_bin_ no	IoT_Smart-Systems_PC1	AI_Data-Analytics_PC2	predicted prob	actual
0	4	0	0	1	1.0561	0.6931	99.914%	1
0	4	0	0	1	0.8961	0.6931	99.885%	1
0	4	0	0	1	0.8865	0.6931	99.883%	1
0	4	0	0	1	0.8297	0.6931	99.870%	1
0	4	0	0	1	0.7785	0.6931	99.857%	1

After filtering out irrelevant or low-compatible startups, this list is helpful for decision-makers to move on to their third stage, where they investigate manually and comprehensively to make the final

decision. Finally, the information and data obtained later on about the partnership results will be updated in the original database to improve future partner suggestions. Figure 8 suggests a framework that can help with the usual partner selection process.

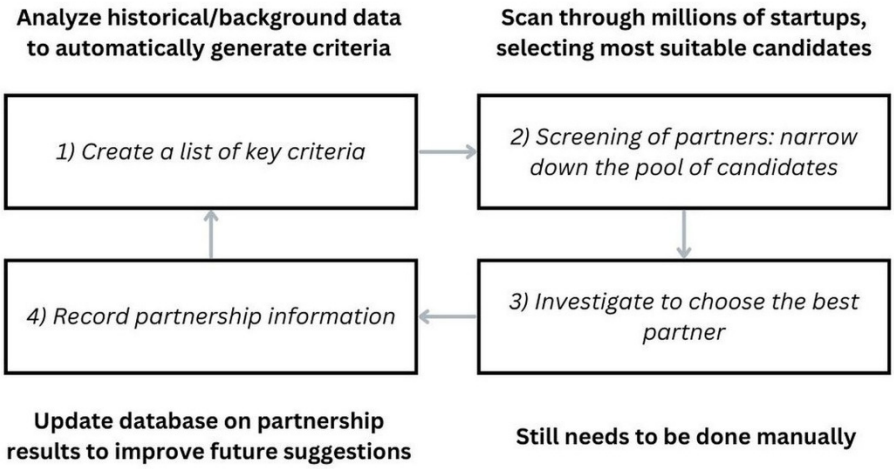

Figure 8. Suggested framework for startup partner selection.

We also derive valuable insights on startup-focused strategic alliances, which stand out from the literature featuring mostly university-company collaborations. Moreover, we address the class imbalance problem by utilizing SMOTE, significantly improving model prediction metrics from 14 to 70 percentage points, as shown in Table 4. Overall, we present a straightforward but effective framework that can be replicated by any company, even SMEs and family businesses, to upgrade their selection criteria. These contributions lay a foundation for future researchers to leverage broader datasets and more advanced techniques to enhance our understanding of strategic partnerships.

While the model demonstrates promising performance in identifying high-potential startups, its applicability across business settings should be considered with caution due to potential data heterogeneity. The features may reflect only Company A's implicit goals and may not generalize to companies in other industries (e.g., healthcare and manufacturing), countries, or business sizes (e.g., large corporates vs. SMEs). To address this, future users of the framework should retrain the model using their own partnership history or startup screening decisions, if available. Firms can utilize publicly available information via web-scraping and implement the data manipulation methods shown in this study. If the target variable (e.g., 'high-potential') is not readily available, semi-supervised learning, as mentioned, or expert annotation can be used to produce a labeled dataset aligned with organizational goals. Feature customization is also necessary to ensure that the model is relevant for the case at hand.

6. Conclusions

Entrepreneurs turn innovative ideas into reality, thus promoting progress, growth, and competitiveness. Around 15% of the entrepreneurs introduced an innovation onto the regional, German,

or world markets in 2017. One start-up in five is “digital”, i.e., digitalization or the use of digital technologies plays a crucial role in the realization of the business concept.

In this study, we provide a Machine Learning framework for classifying the potential startups for partnerships. Using this approach, we followed a structured methodology: Feature selection with the Chi-Square test, dimensionality reduction with PCA, negative sampling through heuristic-based clustering, handling class imbalance with SMOTE, and modeling with LR compared against other models. This procedure, fed by publicly available startup data, demonstrated that a simple and interpretable methodology can be applied to company partner screening and carried out even by resource-constrained firms.

With a precision score tested around 0.800, which can be tuned higher, an F1-score around 0.760, and ROC AUC scores around 0.922, the overall performance proved that the model can be applied for practical use. From the coefficients analysis, four factors appear to be the most influential:

- Technological expertise, especially GPTs, such as IOT, Smart Systems, AI, and DA, is a key consideration.
- Financial maturity, with later funding stages signaling greater stability.
- Geographic proximity, with startups within 200 km of the company showing a greater likelihood of being chosen.
- Business model alignment, with a preference for B2B startups.

The insights derived can be useful to different stakeholders. For companies, the pipeline offers a straightforward way to pre-screen startups efficiently. For investors, financial maturity and technological specialization provide rather trustworthy signals to reduce uncertainty when assessing startups. For policymakers, the importance of geographical proximity and hub associations emphasizes the role of innovation ecosystems in driving startup growth.

There are limitations to be mentioned. The first limitation of the study is that the data used is from the perspective of only one company and its criteria for selecting partners, which may not be representative enough for the whole industry. To address this, we focus on the interpretations behind a company’s decisions instead of developing large-scale, generalizable guidelines for all partnerships. More importantly, we leverage a straightforward data-driven approach and use publicly available information to create a replicable method that any company can use to fit its unique contexts. Another limitation lies in the data collection stage. Second, the labeling of negative samples is based on the assumption that startups not selected by Company A are fundamentally different from the positives, and these differences are embedded in the available variables. While this heuristic approach enables a quick and efficient labeling process, it may introduce bias if the assumption does not hold, especially when some unselected startups can still be potential partners under different conditions. The third limitation revolves around the data source, as the target variable shows only which startups are chosen for the next steps without any clear indicators of success (e.g., whether that partnership is successful). This might undermine our analysis, as the ones chosen by the focal company might not be suitable in the first place, causing later problems such as reputational damage, waste of resources, or delayed projects (World Economic Forum, 2018). The fourth limitation is that the gender data of founders is highly imbalanced and contains many missing values. Although imputation is applied, this may introduce further bias and prediction errors, which should be carefully evaluated in future applications of the model.

In future studies, researchers can enhance the results by adding multiple companies and more variables into the dataset, making the model applicable to a broader context. Moreover, to make the negative samples more reliable, researchers can test the use of soft labels, expert validation, or use multiple firms' historical data and compare this with the weak supervision implemented in this research. Additionally, gaining information on their success would be valuable, as it is crucial to validate the partner selection process from the collaboration initiator. Finally, it would be interesting to gain feedback and empirical studies on whether the approach utilized here increases the efficiencies and effectiveness when screening partners instead of manually picking among thousands of ventures on the market.

Author contributions

Dung Hai Dinh designed the research framework, supervised the overall research process and data analysis. He also contributed to the interpretation of results, coordinated the collaboration among authors, and performed final proofreading and revision of the manuscript.

Van Thanh Tran was primarily responsible for drafting the main text of the manuscript. She contributed to the literature review, methodology description, empirical analysis, and preparation of tables and figures.

Nicole Ondrusch contributed to proofreading and language editing of the manuscript, provided critical feedback on the structure and clarity of the paper, and supported refinement of the discussion and conclusion sections.

All authors read and approved the final manuscript.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is funded by Ministry of Education and Training under grant number B2023-VGU-04.

Conflict of interest

The authors declare no conflict of interest.

Reference

Aithal PS, Aithal S (2018) Study of Various General-Purpose Technologies and Their Comparison Towards Developing Sustainable Society. *Int J Manag Technol Soc Sci* 3: 16–33. http://doi.org/10.5281/zenodo.1409476

Allmendinger MP, Berger ESC (2020) Selecting Corporate Firms for Collaborative Innovation: Entrepreneurial Decision Making in Asymmetric Partnerships. *Int J Innov Manag* 24: 2050003. https://doi.org/10.1142/s1363919620500036

Attah RU, Garba BMP, Gil-Ozoudeh I, et al. (2024) Evaluating strategic technology partnerships: Providing conceptual insights into their role in corporate strategy and technological innovation. *Int J Front Sci Technol Res* 7: 077–089. https://doi.org/10.53294/ijfstr.2024.7.2.0058

Baiman S, Fischer PE, Rajan MV (2000) Information, Contracting, and Quality Costs. *Manag Sci* 46: 776–789. https://doi.org/10.1287/mnsc.46.6.776.11939

Bettinelli C, Bergamaschi M, Kokash R, et al. (2016) Process Innovation, Alliances, and the Interplay of Firm Age: Early Evidence from Italian Small Firms. *Int Bus Res* 9: 86. https://doi.org/10.5539/ibr.v9n5p86

Broekel T, Binder M (2007) The Regional Dimension of Knowledge Transfers—A Behavioral Approach. *Ind Innov* 14: 151–175. https://doi.org/10.1080/13662710701252500

Cardamone P, Pupo V, Ricotta F (2015) University Technology Transfer and Manufacturing Innovation: The Case of Italy. *Rev Policy Res* 32: 297–322. https://doi.org/10.1111/ropr.12125

Chang H, Gausemeier J, Ihmels S, et al. (2008) Innovative Technology Management System with Bibliometrics in the Context of Technology Intelligence, In: Castillo, O., Xu, L., Ao, S.-I. (Eds.), *Trends in Intelligent Systems and Computer Engineering, Lecture Notes in Electrical Engineering*. Springer US, Boston, MA, 349–361. https://doi.org/10.1007/978-0-387-74935-8_25

Chang MH, Liou JJH, Lo HW (2019) A Hybrid MCDM Model for Evaluating Strategic Alliance Partners in the Green Biopharmaceutical Industry. *Sustainability* 11: 4065. https://doi.org/10.3390/su11154065

Chawla NV, Bowyer KW, Hall LO, et al. (2002) SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res* 16: 321–357. https://doi.org/10.1613/jair.953

Cumming DJ, Grilli L, Murtinu S (2013) Governmental and Independent Venture Capital Investments in Europe: A Firm-Level Performance Analysis. *SSRN Electron J.* https://doi.org/10.2139/ssrn.2294746

Daiy AK, Shen KY, Huang JY, et al. (2021) A Hybrid MCDM Model for Evaluating Open Banking Business Partners. *Mathematics* 9: 587. https://doi.org/10.3390/math9060587

Desalegn G, Tangl A, Boros A (2024) The mediating role of customer attitudes in the linkage between e-commerce and the digital economy. *Natl Account Rev* 6: 245–265. https://doi.org/10.3934/NAR.2024011

Dube L, Verster T (2023) Enhancing classification performance in imbalanced datasets: A comparative analysis of machine learning models. *Data Sci Financ Econ* 3: 354–379. https://doi.org/10.3934/DSFE.2023021

Ellison Z (2023) Where Is Your Startup in Its Funding Lifecycle? *BuitIn.com.* Available from: https://builtin.com/founders-entrepreneurship/startup-funding-lifecycle.

Eom BY, Lee K (2010) Determinants of industry–academy linkages and, their impact on firm performance: The case of Korea as a latecomer in knowledge industrialization. *Res Policy* 39: 625–639. https://doi.org/10.1016/j.respol.2010.01.015

Federal Ministry for Economic Affairs and Climate Action (n.d.) START-UP FINDER.

Frei S, Chatterji NS, Bartlett PL (2022) Benign Overfitting without Linearity: Neural Network Classifiers Trained by Gradient Descent for Noisy Linear Data. *Conference on Learning Theory.* https://doi.org/10.48550/ARXIV.2202.05928

Fritsch M, Kauffeld-Monz M (2010) The impact of network structure on knowledge transfer: an application of social network analysis in the context of regional innovation networks. *Ann Reg Sci* 44: 21–38. https://doi.org/10.1007/s00168-008-0245-8

Geum Y, Lee S, Yoon B, et al. (2013) Identifying and evaluating strategic partners for collaborative R&D: Index-based approach using patents and publications. *Technovation* 33: 211–224. https://doi.org/10.1016/j.technovation.2013.03.012

Gigerenzer G, Gaissmaier W (2011) Heuristic Decision Making. *Annu Rev Psychol* 62: 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Giglio C, Corvello V, Coniglio IM, et al. (2025) Cooperation between large companies and start-ups: An overview of the current state of research. *Eur Manag J* 43: 142–153. https://doi.org/10.1016/j.emj.2023.08.002

Giunta A, Pericoli FM, Pierucci E (2016) University–Industry collaboration in the biopharmaceuticals: the Italian case. *J Technol Transf* 41: 818–840. https://doi.org/10.1007/s10961-015-9402-2

Gonzales JT (2023) Implications of AI innovation on economic growth: a panel data study. *J Econ Struct* 12: 13. https://doi.org/10.1186/s40008-023-00307-w

Guertler MR, Sick N (2021) Exploring the enabling effects of project management for SMEs in adopting open innovation – A framework for partner search and selection in open innovation projects. *Int J Proj Manag* 39: 102–114. https://doi.org/10.1016/j.ijproman.2020.06.007

Günsel A, Dodourova M, Tükel Ergün A, et al. (2019) Research on effectiveness of technology transfer in technology alliances: evidence from Turkish SMEs. *Technol Anal Strateg Manag* 31: 279–291. https://doi.org/10.1080/09537325.2018.1495836

He H, Garcia EA (2009) Learning from Imbalanced Data. *IEEE Trans Knowl Data Eng* 21: 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hostettler S, Bolay JC (2014) Technologies and Partnerships, In: Bolay, J.-C., Hostettler, S., Hazboun, E. (Eds.), *Technologies for Sustainable Development*. Springer International Publishing, Cham, 3–9. https://doi.org/10.1007/978-3-319-00639-0_1

Hu C, Zhu X, Liu R, et al. (2024) The impact of external technology acquisition on enterprise innovation performance: the moderating effect of geographical distance. *Technol Anal Strateg Manag* 36: 2875–2889. https://doi.org/10.1080/09537325.2023.2174359

Islahulhaq WW, Iis DR (2021) Classification of Non-Performing Financing Using Logistic Regression and Synthetic Minority Over-sampling Technique-Nominal Continuous (SMOTE-NC). *Int J Adv Soft Comput Its Appl* 13: 116–128. https://doi.org/10.15849/IJASCA.211128.09

Jeon J, Lee C, Park Y (2011) How to Use Patent Information to Search Potential Technology Partners in Open Innovation. *J Intellect Prop RIGHTS* 16: 385–393.

Joseph M Hilbe (2015) Practical Guide to Logistic Regression. *CRC Press - Taylor & Francis Group*. https://doi.org/10.1201/b18678

Kang I, Han S, Shin GC (2014) A process leading to strategic alliance outcome: The case of IT companies in China, Japan and Korea. *Int Bus Rev* 23: 1127–1138. https://doi.org/10.1016/j.ibusrev.2014.03.008

Kang J, Lee J, Jang D, et al. (2019) A Methodology of Partner Selection for Sustainable Industry-University Cooperation Based on LDA Topic Model. *Sustainability* 11: 3478. https://doi.org/10.3390/su11123478

Kopka A, Fornahl D (2024) Artificial intelligence and firm growth — catch-up processes of SMEs through integrating AI into their knowledge bases. *Small Bus Econ* 62: 63–85. https://doi.org/10.1007/s11187-023-00754-6

Lee K, Park I, Yoon B (2016) An Approach for R&D Partner Selection in Alliances between Large Companies, and Small and Medium Enterprises (SMEs): Application of Bayesian Network and Patent Analysis. *Sustainability* 8: 117. https://doi.org/10.3390/su8020117

Li J, Zheng K, Xu H, et al. (2019) The Strength of the Weakest Supervision: Topic Classification Using Class Labels, In: *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Presented at the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 22–28.

Li Y, Yan K (2025) Prediction of bank credit customers churn based on machine learning and interpretability analysis. *Data Sci Financ Econ* 5: 19–34. https://doi.org/10.3934/DSFE.2025002

Liu Y, Brown SD (2013) Comparison of five iterative imputation methods for multivariate classification. *Chemom Intell Lab Syst* 120: 106–115. https://doi.org/10.1016/j.chemolab.2012.11.010

Lo HW, Chang DS, Huang LT (2022) Sustainable Strategic Alliance Partner Selection Using a Neutrosophic-Based Decision-Making Model: A Case Study in Passive Component Manufacturing. *Complexity* 2022: 1–18. https://doi.org/10.1155/2022/9483256

Richardson M (2009) Principal Component Analysis. Available from: http://www.sdss.jhu.edu/~szalay/class/2024/etc/SignalProcPCA.pdf.

Meulman F, Reymen IMMJ, Podoynitsyna KS, et al. (2018) Searching for Partners in Open Innovation Settings: How to Overcome the Constraints of Local Search. *Calif Manage Rev* 60: 71–97. https://doi.org/10.1177/0008125617745087

Mockler RJ (1999) Multinational strategic alliances, Wiley series in practical strategy. Wiley, Chichester ; New York.

Mohnen P, Rõigas K, Varblane U (2018) Which firms use universities as cooperation partners? - A comparative view in Europe. *Int J Technol Manag* 76: 32. https://doi.org/10.1504/IJTM.2018.10009595

Mori J, Kajikawa Y, Kashima H, et al. (2012) Machine learning approach for finding business partners and building reciprocal relationships. *Expert Syst Appl* 39: 10402–10407. https://doi.org/10.1016/j.eswa.2012.01.202

Niederhauser L, Waefler T, Huber S, et al. (2022) Matching B2B-Partners in the Sharing Economy. *The 13th International Conference on Applied Human Factors and Ergonomics (AHFE 2022)*, 56: 10-16. https://doi.org/10.54941/ahfe1002246

Öndas V (2021) A Study on High-tech Startup Failure. https://doi.org/10.13140/RG.2.2.25524.37765

Partal MO, Gönel F (2025) Startup performance from an economic development perspective: impact evaluation after funding stage. *Glob Bus Econ Rev* 32: 357–376. https://doi.org/10.1504/GBER.2025.146493

Patel N (2015) 90% Of Startups Fail: Here's What You Need to Know About The 10%. *Forbes*.

Poslavskaya E, Korolev A (2023) Encoding categorical data: Is there yet anything "hotter" than one-hot encoding? https://doi.org/10.48550/ARXIV.2312.16930

Poyiadzi R, Bacaicoa-Barber D, Cid-Sueiro J, et al. (2022) The Weak Supervision Landscape. https://doi.org/10.1109/PerComWorkshops53856.2022.9767420

PWC (2009) Strategic Partnerships: The Real Deal?

Raghunathan TE, Lepkowski JM, Van Hoewyk J (2001) A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv Methodol* 27: 85–95.

Raghuvanshi J, Shukla D, Dahiya R (2022) Parameters for selecting the partners in locally owned renewable energy small-scale project for achieving energy security in Atlantic Canada. *Energy Sustain Dev* 68: 512–524. https://doi.org/10.1016/j.esd.2022.04.016

Ratner A, Bach SH, Ehrenberg H, et al. (2017) Snorkel: rapid training data creation with weak supervision. *Proc VLDB Endow* 11: 269–282. https://doi.org/10.14778/3157794.3157797

Riepe J, Uhl K (2020) Startups' demand for non-financial resources: Descriptive evidence from an international corporate venture capitalist. *Financ Res Lett* 36: 101321. https://doi.org/10.1016/j.frl.2019.101321

Rothaermel FT (2001) Complementary assets, strategic alliances, and the incumbent's advantage: an empirical study of industry and firm effects in the biopharmaceutical industry. *Res Policy* 30: 1235–1251. https://doi.org/10.1016/s0048-7333(00)00142-6

Rubin DB (1976) Inference and missing data. *Biometrika* 63: 581–592. https://doi.org/10.1093/biomet/63.3.581

Rybnicek R, Königsgruber R (2019) What makes industry–university collaboration succeed? A systematic review of the literature. *J Bus Econ* 89: 221–250. https://doi.org/10.1007/s11573-018-0916-6

Sebo P (2022) Are Accuracy Parameters Useful for Improving the Performance of Gender Detection Tools? A Comparative Study with Western and Chinese Names. *J Gen Intern Med* 37: 4024–4027. https://doi.org/10.1007/s11606-022-07469-6

Sinha RK (2024) Book review: Christoph Molnar. 2020. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Metamorph. *J Manag Res* 23: 92–93. https://doi.org/10.1177/09726225241252009

Solesvik M, Gulbrandsen M (2013) Partner Selection for Open Innovation. *Technol Innov Manag Rev* 3: 11–16. https://doi.org/10.22215/timreview/674

Steve F, Michael S, Susanne H, et al. (2025) 2025 technology industry outlook. Deloitte. Available from: https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-telecom-outlooks/technology-industry-outlook.html.

Sun H, Peng X, Tang X, et al. (2024) The mechanism of high carbon economy, local tax system and urban environmental pollution: Insights from China. *Indoor Built Environ*. https://doi.org/10.1177/1420326X241258869

Tello S, Latham S, Kijewski V (2010) Individual choice or institutional practice: Which guides the technology transfer decision-making process? *Manag Decis* 48: 1261–1281. https://doi.org/10.1108/00251741011076780

Trela K, Campbell Y, Dornbusch F, et al. (2021) How to Find New Industry Partners for Public Research: A Classification Approach. *IEEE Trans Eng Manag* 68: 1214–1231. https://doi.org/10.1109/TEM.2020.2992060

Weiblen T, Chesbrough HW (2015) Engaging with Startups to Enhance Corporate Innovation. *Calif Manage Rev* 57: 66–90. https://doi.org/10.1525/cmr.2015.57.2.66

Williamson OE (1979) Transaction-Cost Economics: The Governance of Contractual Relations. *J Law Econ* 22: 233–261. https://doi.org/10.1086/466942

Wongvorachan T, He S, Bulut O (2023) A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information* 14: 54. https://doi.org/10.3390/info14010054

World Economic Forum (2018) Collaboration between Start-Ups and Corporates A Practical Guide for Mutual Understanding (White Paper). World Economic Forum.

Wu C, Barnes D (2011) A literature review of decision-making models and approaches for partner selection in agile supply chains. *J Purch Supply Manag* 17: 256–274. https://doi.org/10.1016/j.pursup.2011.09.002

Wu C, Barnes D (2014) Partner selection in agile supply chains: a fuzzy intelligent approach. *Prod Plan Control* 25: 821–839. https://doi.org/10.1080/09537287.2013.766037

Wu L, Sun L, Chang Q, et al. (2022) How do digitalization capabilities enable open innovation in manufacturing enterprises? A multiple case study based on resource integration perspective. *Technol Forecast Soc Change* 184: 122019. https://doi.org/10.1016/j.techfore.2022.122019

Yan A, Ma H, Zhu D, et al. (2024) Digital transformation and corporate resilience: Evidence from China during the COVID-19 pandemic. *Quant Financ Econ* 8: 779–814. https://doi.org/10.3934/QFE.2024030

Zhang ML, Zhou ZH (2014) A Review on Multi-Label Learning Algorithms. *IEEE Trans Knowl Data Eng* 26: 1819–1837. https://doi.org/10.1109/TKDE.2013.39

AIMS Press