Using Predictive Models to Identify Trends Among Successful Dual-Use Startups

by

Samantha Ying

S.B. Computer Science, Economics, and Data Science and Business Analytics,
Massachusetts Institute of Technology, 2024

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN COMPUTER SCIENCE, ECONOMICS, AND DATA SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2025

Authored by:	Samantha Ying Department of Electrical Engineering and Computer Science January 17, 2025
Certified by:	Fiona Murray MIT Innovation Initiative Director, Thesis Supervisor
Accepted by:	Katrina LaCurts Chair, Master of Engineering Thesis Committee

Using Predictive Models to Identify Trends Among Successful Dual-Use Startups

by

Samantha Ying

Submitted to the Department of Electrical Engineering and Computer Science
on January 17, 2025 in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN COMPUTER SCIENCE, ECONOMICS, AND DATA SCIENCE

ABSTRACT

This study examines predictive models for assessing the success of dual-use startups in the United States. Utilizing data from the Small Business Innovation Research (SBIR) and Small Business Technology Transfer (STTR) programs, this research focused on startups founded post-2000 to reflect contemporary technological advancements. A key objective of this study was to create a rich and comprehensive dataset, addressing gaps in the dual-use startup literature and providing a foundation for future research. Machine learning approaches, including Logistic Regression, Random Forest, and Gradient Boosting Machines, were applied to evaluate critical success factors, with XGBoost identified as the most effective model. Despite the challenges of class imbalance, the study highlights the potential of data-driven methodologies to uncover trends and inform strategies for supporting dual-use startups. By integrating predictive modeling with the construction of a robust dataset, this research contributes both to the academic understanding of dual-use innovation ecosystems and to practical frameworks for fostering their growth.

Thesis supervisor: Fiona Murray
Title: MIT Innovation Initiative Director

Acknowledgments

I am deeply grateful to Katy Person and Professor Fiona Murray for their unwavering support and encouragement over the past two years. I would also like to thank Dr. AJ Perez for his guidance and expertise.

I extend my thanks to Gene Keselman and the team at Mission Innovation X for laying the foundation for this project and giving me the opportunity to work on it.

To my family, thank you for everything.

To all my friends who made MIT a difficult place to say goodbye to, especially the 宝贝, without whom I wouldn't have made it through MIT.

And to Bella, 고마워, for staying by my side and finishing this journey with me.

Contents

List of Figures

List of Tables

List of Equations

Chapter 1

Introduction

1.1 Introduction to Dual-Use

Many of the technologies today, from the Global Positioning System (GPS)[1] and touchscreens[2] to the Internet[3], began with government funding for military purposes, but later found commercial success. This phenomenon illustrates the concept of dual-use technology, meaning they serve both military and civilian applications. Dual-use startups are critical because they drive innovation in fields such as defense, aerospace, biotechnology, and information technology, with impacts that extend far beyond the original military applications.

Governments play a pivotal role in fostering these dual-use technologies by acting as key stakeholders in innovation ecosystems, as mentioned in MIT's Stakeholder Framework for Building & Accelerating Innovation Ecosystems [4]. Successful government initiatives, such as Singapore's ecosystem-focused innovation strategy and Israel's Yozma program, demonstrate how strategic interventions can catalyze entrepreneurship and venture capital growth while addressing societal and economic challenges [4]. By aligning public R&D investments with entrepreneurial needs, governments create opportunities for dual-use startups to thrive in both military and civilian markets.

1.2 SBIR Program

One of the most critical sources of early-stage funding for dual-use startups in the United States is the Small Business Innovation Research (SBIR) program. Created in 1982, the program provides non-dilutive funding to small businesses, offering grants and contracts to support research and development[5]. The SBIR program was established to encourage innovation by providing funding to small businesses developing dual-use technologies. This

program enables dual-use startups to get a head start on governmental applications, allowing them to focus on research and development without relying solely on revenue from sales.

Over the years, several successful companies have benefited from the SBIR program. Some notable SBIR winners include Qualcomm[6], a leader in wireless technology, and 23andMe[7], a leader in consumer genetics and personalized health services.

SBIR funding is structured in phases, allowing startups to gradually progress from initial research to prototyping and to production. In Phase I, startups receive grants to conduct feasibility studies, exploring the technical and commercial viability of their ideas. This stage is crucial for laying the groundwork for further development, enabling startups to refine their ideas and build a foundation for future work.

In Phase II, SBIR provides additional funding to help startups continue their research and development efforts. This phase focuses on building prototypes and conducting more extensive testing. With this support, startups can advance their technology, making it more market-ready and suitable for both government and commercial applications. Phase II funding can also help startups bridge the gap between research and commercialization.

To qualify for the SBIR program, the small business must be a for-profit entity based in the United States, with 500 or fewer employees, including affiliates. The Principal Investigator (PI) must be primarily employed by the small business during the project, and the small business must perform the majority of the R&D work, completing at least 66% of the work in Phase I and 50% in Phase II[8].

1.3 STTR Program

In addition to the SBIR program, the Small Business Technology Transfer (STTR) program was introduced by the US Small Business Administration in the early 1990s to foster collaboration between small businesses and research institutions. This collaboration leverages the human and social capital of academic inventors, which has been shown to significantly impact the success of entrepreneurial ventures. Academic inventors contribute not only technical expertise but also vital social networks that embed startups within broader scientific communities, enabling access to critical resources and partnerships[9]. Such embeddedness enhances the firm's capacity to innovate and navigate complex technological and market challenges.

The STTR program fosters structured, strategic partnerships between small businesses and research institutions, aligning their R&D efforts with broader innovation goals. An analysis by Frølund, Murray, and Riedel[10] highlights that strategic partnerships are more effective than ad hoc collaborations because they enable both parties to formalize their re-

lationships and align on shared objectives. While ad hoc partnerships are often initiated by individual researchers or engineers within companies to address specific R&D problems, strategic partnerships take a more holistic approach by integrating expertise across multiple organizational units. This approach fosters deeper synergy and allows for multifaceted interactions.

Universities have played pivotal roles in fostering collaboration and providing access to resources in the past. For instance, MIT and Harvard have played critical roles in the Greater Boston innovation ecosystem, contributing to the development of both defense and life sciences sectors [10]. Similarly, universities like Carnegie Mellon and the University of Pittsburgh have revitalized economically stagnant regions like Pittsburgh, leveraging strengths in robotics and computer science to drive innovation [4].

The STTR program has similar eligibility requirements to the SBIR program, but includes additional conditions to foster collaboration with research institutions. Like SBIR, the small business must be a for-profit entity based in the United States with 500 or fewer employees. However, STTR requires that the small business must formally collaborate with a U.S.-based non-profit research institution, such as a university or federal laboratory. The small business must perform at least 40% of the R&D work, while the research institution must perform at least 30%[11]. Additionally, the principal investigator in STTR projects can be employed either by the small business or the research institution, providing greater flexibility in project leadership while emphasizing partnership.

1.4 Challenges of Dual-Use

Dual-use startups operate in a distinct space, serving both military and civilian markets. Serving both markets introduces distinct challenges, particularly due to the complexities of government funding and procurement. Unlike traditional startups, dual-use companies must navigate the government budget cycles, compliance requirements, and procurement processes of the Department of Defense (DOD) or other governmental agencies, all while maintaining a presence in commercial markets.

One of the most significant hurdles has been named the "valley of death"[12], a term describing the gap between early-stage funding (provided through programs like SBIR and STTR) and achieving consistent revenue through long-term government contracts. The non-dilutive nature of SBIR and STTR funding is attractive, allowing startups to focus on innovation without the pressure of equity dilution. However, these programs often fund only the initial stages of research and development. Transitioning from prototype development to production-ready solutions in the military market can take years due to the government's

slow procurement processes, lengthy contract negotiations, and strict compliance standards. During this transition, startups often face funding gaps, which can jeopardize their growth and sustainability.

Furthermore, dual-use startups must face misaligned budget cycles within government agencies. Defense R&D funding doesn't always translate into recurring revenue, creating a gap that makes it difficult for companies to secure consistent support for scaling their technologies. Additionally, the DOD's unique requirements, such as specific use cases, security protocols, and rigorous testing standards, can create barriers that civilian-focused startups typically do not encounter.

Another critical challenge is balancing military and civilian market demands. While civilian markets may offer more predictable revenue opportunities and faster sales cycles, military markets require startups to tailor their business models to align with DOD priorities. Early and proactive engagement with stakeholders and leveraging partnerships with established defense contractors and research institutions can help dual-use startups bridge this gap.

This paper aims to explore the characteristics of dual-use startups and how they might contribute to success in civilian and military applications. This research will take a quantitative approach, using machine learning methods to uncover patterns among successful dual-use startups.

1.5 Related Work

The study of entrepreneurial success has garnered significant attention in the academic literature, with various works examining the characteristics that influence startup outcomes. A notable study on businesses in Massachusetts[13] explored predictors such as incorporation status, naming patterns, patent filings, and geographical location, concluding that these factors significantly impact the quality and performance of startups. The study also found that eponymous firm names—those named after founders—were less associated with high entrepreneurial growth. These names often signaled alignment with family-owned or 'lifestyle' businesses rather than innovation-driven ventures. Additionally, growth-oriented startups were more likely to have concise, distinctive names, often comprising no more than two words. While the analysis of eponymous firm names provides valuable insights, it is less applicable to innovation-focused startups. Building on the findings of the Massachusetts study, this research focuses specifically on dual-use startups, analyzing the unique factors that contribute to their success.

In the realm of dual-use startups, the relationship between SBIR funding and project success has been the focus of several studies. Audretsch, Link, and Scott[14] evaluate the

effectiveness of the SBIR program in fostering socially valuable research and development and facilitating the commercialization of innovative technologies. The study highlights that SBIR-funded projects achieve high social rates of return compared to private returns, underscoring the program's ability to address market failures. Social returns reflect the broader societal and economic benefits of funded projects, such as technological advancements and knowledge spillovers, while private returns focus on the direct financial gains of the funded companies. These findings emphasize the critical role of government funding in enabling projects that yield significant public value but may not attract private investment due to high risk or uncertain profitability.

Further insights into SBIR-funded projects were provided by Link, Swann, and van Hasselt[15], who conducted an assessment of failure rates among SBIR-funded Phase II projects. Their study identified that projects with additional non-SBIR developmental funding and university involvement had a significantly lower likelihood of failure. Firms with prior SBIR awards also benefited from accumulated human capital and technical expertise, which reduced the probability of project failure. However, the study found no significant difference in failure rates based on agency funding or ownership characteristics such as gender or minority status. These findings underscore the importance of external financial and academic support in driving the success of R&D initiatives, which aligns with the broader goals of dual-use startups seeking to transition from government-funded innovation to commercialization.

In addition, a study employing a probit regression model examined the commercialization success of SBIR-funded projects[16], revealing that factors such as additional funding, firm size, and a history of previous awards positively correlate with commercialization outcomes. Larger firms, with more resources and infrastructure, were better positioned to commercialize SBIR-funded technologies, while companies with additional developmental funding from non-SBIR sources, such as private investors or venture capital, demonstrated higher success rates. Similarly, firms with a history of prior SBIR awards benefited from accumulated expertise and improved capacity to navigate the complexities of government-funded projects, increasing their likelihood of commercialization. However, while these findings provide important insights into the factors influencing commercialization outcomes, they do not fully address the broader trajectories of company success, such as sustained growth, recurring government contracts, or transitioning from government funding to private-sector viability. This gap highlights the need for more predictive models tailored specifically to dual-use startups. These startups face unique challenges compared to traditional ones, including balancing the demands of dual markets and securing funding for technologies with both commercial and defense applications.

Other studies have developed predictive models to forecast startup success[17], identifying

key predictors such as the timing of the last funding round, the lag until the first funding round, and company age. For instance, research has shown that startups receiving later-stage funding rounds closer to their founding date are more likely to succeed, as this often indicates early validation of their business model or technology. Similarly, a shorter lag time between founding and the first funding round has been associated with higher entrepreneurial quality, reflecting the ability of startups to quickly attract external investment. Company age has also emerged as a significant predictor, with younger firms typically exhibiting greater flexibility and innovation but facing higher failure risks due to limited resources. However, while these models provide valuable tools for understanding entrepreneurial outcomes, they have not been specifically applied to dual-use startups. Dual-use ventures operate in a distinct space, often requiring both government support and private investment to commercialize technologies that straddle defense and commercial markets. As such, there remains a critical need for research that adapts these predictive frameworks to the unique characteristics of dual-use startups.

Chapter 2

Data

2.1 Scope

2.1.1 Defining Success

When exploring trends among successful dual-use startups, it is crucial to first establish a precise definition of success. Unlike typical startups, where success might be measured by an IPO or a merger, dual-use startups operate in a unique space that demands a more nuanced approach. To better understand this, interviews were conducted with founders from ten dual-use startups, asking them to define what success means to them. From these discussions, two primary themes emerged: self-sufficiency and mission achievement. Founders view self-sufficiency as the ability of the startup to generate sufficient revenue without relying on external funding. Mission success is defined by the startup's ability to solve the specific problem it was established to address.

Measuring self-sufficiency is straightforward, but quantifying mission success is more complex due to its varied nature across startups. This research will use government contracts as a proxy for both self-sufficiency and mission success. This research focuses specifically on contract IDVs (Indefinite Delivery Vehicles). IDVs, such as Indefinite Delivery/Indefinite Quantity (IDIQ) contracts or Government-Wide Acquisition Contracts (GWACs), are particularly valuable in this context because they represent recurring revenue and a sustained commitment from government agencies. IDVs allow for the delivery of goods or services over time, demonstrating that a startup's offerings remain relevant and integral to the government's needs. Ongoing alignment with government needs demonstrates both the startup's ability to consistently deliver value and its effectiveness in fulfilling its mission objectives.

In contrast, other award types are less effective for this purpose. Blanket Purchase Agreements (BPAs), while suitable for repetitive transactions, typically lack the scale and long-

term commitment seen in IDVs. Similarly, Purchase Orders (POs) and Delivery Orders (DOs) are usually tied to specific, one-time transactions or discrete deliveries, making them less indicative of recurring demand. Definitive Contracts, though comprehensive and detailed, are often finite in nature and do not necessarily reflect sustained government reliance on the startup's capabilities. By focusing on IDVs, this research highlights contracts that embody recurring revenue streams and long-term relationships.

2.2 Data Collection and Cleaning

2.2.1 Source Data

The data for this research comes from 3 different sources. The starting sample for this research is drawn from the SBIR and STTR database, which list all awardees since the inception of the programs. This database, accessible through sbir.gov, provides a comprehensive record of all businesses that have received SBIR or STTR awards, making it a rich source for identifying dual-use startups.

The next data source was Pitchbook, a widely recognized database that provides extensive financial and business information on companies. Pitchbook was chosen as a data source because it offers a comprehensive view of a company's financial health, funding sources, and business relationships. This information is critical to the analysis, as it helps identify key characteristics that may correlate with startup success, such as funding rounds, company demographics, and business models.

The last data source was IDV contracts downloaded from usaspending.gov. This dataset provided key details such as contract award dates, awarding agencies, and demographic information about businesses, including classifications like veteran-owned, woman-owned, and minority-owned. These attributes can be used to track the progression of dual-use startups and evaluate their success based on recurring government contracts, aligning with our defined success criteria.

2.2.2 Retrieving Companies from Pitchbook

The data collection process began with a comprehensive list of all unique SBIR/STTR awardees from 1982 to 2023 obtained from sbir.gov, totaling 31,774 companies. To focus the study on more recent startups, the sample was narrowed to include only SBIR/STTR awardees founded in 2000 or later. This decision aligns with our goal to analyze modern dual-use startups and understand their characteristics and success factors in the current market

environment. After filtering out older SBIR/STTR awardees, this reduced the sample to 21,022 companies.

Next, this refined list was uploaded to Pitchbook for matching to company profiles. Pitchbook's proprietary matching algorithm used the company name, headquarters city, and headquarters state provided in the SBIR/STTR database to identify corresponding profiles. Of the 21,022 companies, 11,890 were successfully matched to profiles on Pitchbook.

To further narrow the dataset, only companies founded in 2000 or later were kept, resulting in a final sample of 9,548 companies. Due to Pitchbook's download constraints, these profiles were downloaded in batches and later combined into one dataset. Duplicate company profiles in the dataset (seven in total) were identified due to Pitchbook's matching algorithm and subsequently removed, leaving the dataset with 9,451 unique profiles.

2.2.3 Adding Deal Data

Deal data was subsequently added to the dataset. Deals are defined as transactions in which a company receives funding from an investor and/or lender. Deal data was downloaded separately from the "Deal" tab on Pitchbook and matched to the dataset using the unique Pitchbook ID assigned to each company. Deal types on Pitchbook include venture capital funding, accelerators, crowdfunding, grants (including SBIR/STTR), and more. A full explanation of Pitchbook deal types is included in Appendix A.

Companies without any deal data (indicating they had not received any funding, including grant funding) were removed from the dataset. These companies were flagged by Pitchbook as "not actively tracked", which means that their profiles were not regularly updated. A total of 665 such companies were removed, narrowing the dataset to 8,876 profiles.

2.2.4 Location Discrepancies

After adding the deal data, 130 companies with non-U.S. headquarters or missing location data were identified. Since SBIR/STTR recipients must be based in the United States, these profiles were manually reviewed using team member details listed in Pitchbook and cross-referenced with contact names on SBIR/STTR award records. 48 companies were confirmed as SBIR/STTR recipients that had either moved their headquarters out of the U.S. or had multiple headquarters, while the rest (82 companies) were deemed false matches. Missing location data was updated, and false matches were removed, reducing the dataset to 8,794 companies.

2.2.5 Adding SBIR/STTR Award Data

To link SBIR/STTR award data with Pitchbook profiles, company names were standardized by converting them to lowercase, removing punctuation, and stripping common suffixes such as "Inc." or "LLC." Former and alternative company names listed in Pitchbook, which were often consolidated into a single entry, were split into separate entries to enhance the matching process. The SBIR/STTR award data was organized chronologically, ensuring that multiple awards for the same company were recorded in sequence.

The next step involved aligning variations in company names across the two datasets. While the Pitchbook dataset included former names, alternative names, and legal names, these variations were standardized to facilitate accurate matching. Chronological organization of the SBIR/STTR award data ensured that awards were linked in order, providing a clear historical record for each company.

To integrate the datasets, an iterative matching process compared standardized company names from the Pitchbook dataset with those in the SBIR/STTR award dataset. When a match was found, relevant award details, such as award dates and descriptions, were extracted and linked to the corresponding company in Pitchbook. Companies with multiple awards had each award recorded separately to prevent overwriting information.

The resulting dataset combined detailed SBIR/STTR award information with Pitchbook profiles, creating a chronological and comprehensive view of each company's SBIR/STTR award history. This enriched dataset served as a valuable resource for analyzing trends, funding patterns, and the impact of these awards on company development. It was structured for further analysis and saved for future use.

2.2.6 Profiles Without SBIR/STTR Matches

Of the 8,794 companies, 379 were not matched during the integration of SBIR and STTR award data. This was due to issues such as company name changes (not reflected in the former names listed on Pitchbook), slight variations in spacing or punctuation not caught during name standardization, or erroneous matches made by Pitchbook.

To address these edge cases, the 379 profiles were manually reviewed. Each profile on Pitchbook was cross-checked against the SBIR/STTR award dataset using additional information such as POC and PI names, headquarters city, and state. If no match was found, the company website or UCC filings were examined to determine if any individuals associated with the SBIR/STTR award were linked to the company. After investigation, 79 companies remained unmatched, indicating that Pitchbook's matching algorithm had incorrectly linked them. These profiles were removed, resulting in a final dataset of 8,715 companies.

2.2.7 Adding IDV Data

To add the success criteria to the dataset, an IDV dataset was created by downloading contract data year by year from 2000 to 2023 using the U.S. government's custom award data search tool[18], which allows only single-year downloads. These files were then merged into a single comprehensive dataset. Only the first IDV contract for each company was included in the dataset. This approach was taken to isolate the effects of other startup characteristics on the likelihood of receiving an IDV, as receiving an initial IDV may facilitate the acquisition of additional IDVs due to unobservable connections with government purchasing agencies or network effects.

To match IDV recipients to the companies in the dataset, company names were cleaned and standardized using the same methodology as mentioned above to enable accurate comparisons. Candidate company names from the Pitchbook dataset were evaluated for potential matches against company names in the IDV dataset.

Fuzzy matching library `RapidFuzz`[19] was used to compute similarity ratios between strings, taking into account both the textual similarity of the names and their contextual overlap. The contextual overlap was measured by comparing shared and unique terms in the names, with a penalty applied to discourage partial matches that lacked sufficient alignment. This approach allowed for a more flexible and robust comparison of company names, accounting for variations in spelling, formatting, and word order. By calculating similarity scores between candidate names, the fuzzy matching method enabled accurate identification of matches even when the names were not perfectly identical.

Using `RapidFuzz`, the highest-scoring matches exceeding a similarity threshold of 80 were selected. To avoid duplication, additional checks ensured that companies already matched were not rematched. When a match was confirmed, relevant information from the IDV dataset was added to the corresponding record in the Pitchbook dataset, enriching it with details from the government contracts.

The resulting dataset links companies from Pitchbook to their corresponding first IDV contract, providing a structured and accurate integration of contract data. This integration enhances the dataset by adding valuable information about companies' initial interactions with IDV contracts.

1,130 IDV contracts were matched to companies in the dataset. However, a review revealed that 276 of these contracts were awarded before the listed founding year of the respective companies in Pitchbook. These discrepancies, identified as invalid matches, were removed to ensure the accuracy of the dataset. Following this adjustment, the dataset now contains 854 companies with IDV contracts.

2.3 Descriptive Statistics

To provide a clearer understanding of the dataset created for this research, various descriptive statistics are presented to highlight key characteristics and trends within the data.

2.3.1 SBIR/STTR Dollars in Context

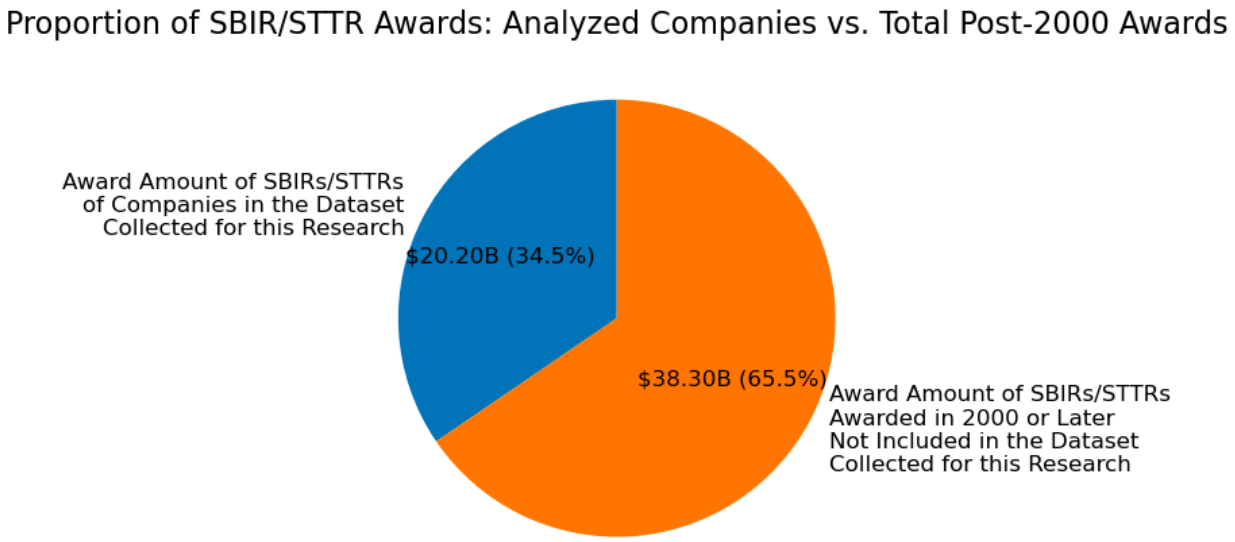

Figure 2.1: Pie chart depicting the proportion of SBIR/STTR award amount represented in the analysis compared to total SBIR/STTR award amount awarded after 2000.

Figure 2.1 compares the total SBIR/STTR award amounts allocated to the analyzed subset of companies with the total awards distributed since 2000. The subset accounted for $20.2 billion, representing 34.5% of the total SBIR/STTR award amount distributed since 2000. The remaining $38.3 billion, or 66%, was awarded to companies outside of the dataset collected for this research.

The primary differentiator between the two sets is if the company was found on Pitchbook. It is worth noting that companies are included in PitchBook if they are deemed significant enough for tracking, often based on factors such as funding, growth trajectory, or visibility in the market. However, the absence of a company on PitchBook does not definitively indicate insignificance. It could also reflect a lack of publicly available information or other barriers to PitchBook creating a profile for these companies. This means that while some companies outside the dataset may not have grown to a scale warranting inclusion, others

may have simply operated under circumstances where insufficient data was accessible for profiling. This ambiguity emphasizes the importance of analyzing the companies available on Pitchbook, as they likely represent a subset with relatively greater public visibility and commercial potential.

2.3.2 SBIR/STTR Award Amounts and Counts, Separated by Agency

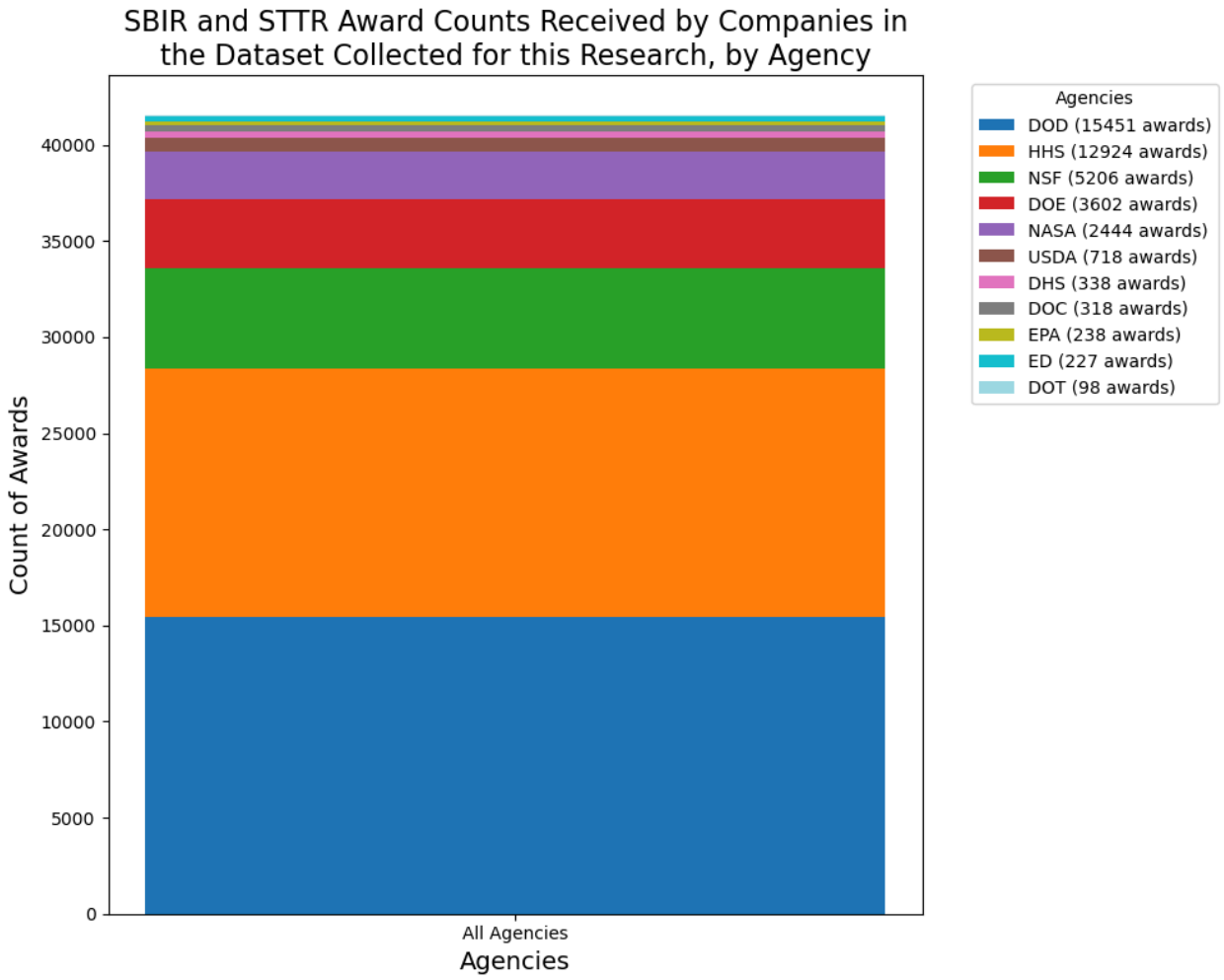

Figure 2.2: Stacked bar chart depicting number of SBIR/STTR awards given to companies represented in the dataset for this analysis, separated by agency.

Figure 2.2 highlights the distribution of SBIR and STTR awards received by companies represented in the dataset collected for this analysis, categorized by the U.S. government agencies that issued the awards. The Department of Defense (DOD) leads as the largest contributor, issuing 15,451 awards, followed by the Department of Health and Human Services (HHS) with 12,924 awards and the National Science Foundation (NSF) with 5,206

awards. The order of the top two agencies—DOD and HHS-aligns closely with their relative funding budgets, as outlined on the SBIR website[20]. What's interesting is that the dataset contains less Department of Energy (DOE) awards than NSF awards, despite the DOE having a larger budget for SBIRs and STTRs.

The DOE and NASA stand out as mid-level contributors, with 3,602 and 2,444 awards respectively, reflecting their strong focus on energy-related innovations and aerospace technology development. In contrast, agencies such as the U.S. Department of Agriculture (USDA), Department of Homeland Security (DHS), Department of Commerce (DOC), Environmental Protection Agency (EPA), Department of Education (ED), and Department of Transportation (DOT) represent a smaller share of awards, with counts ranging from 718 (USDA) to 98 (DOT).

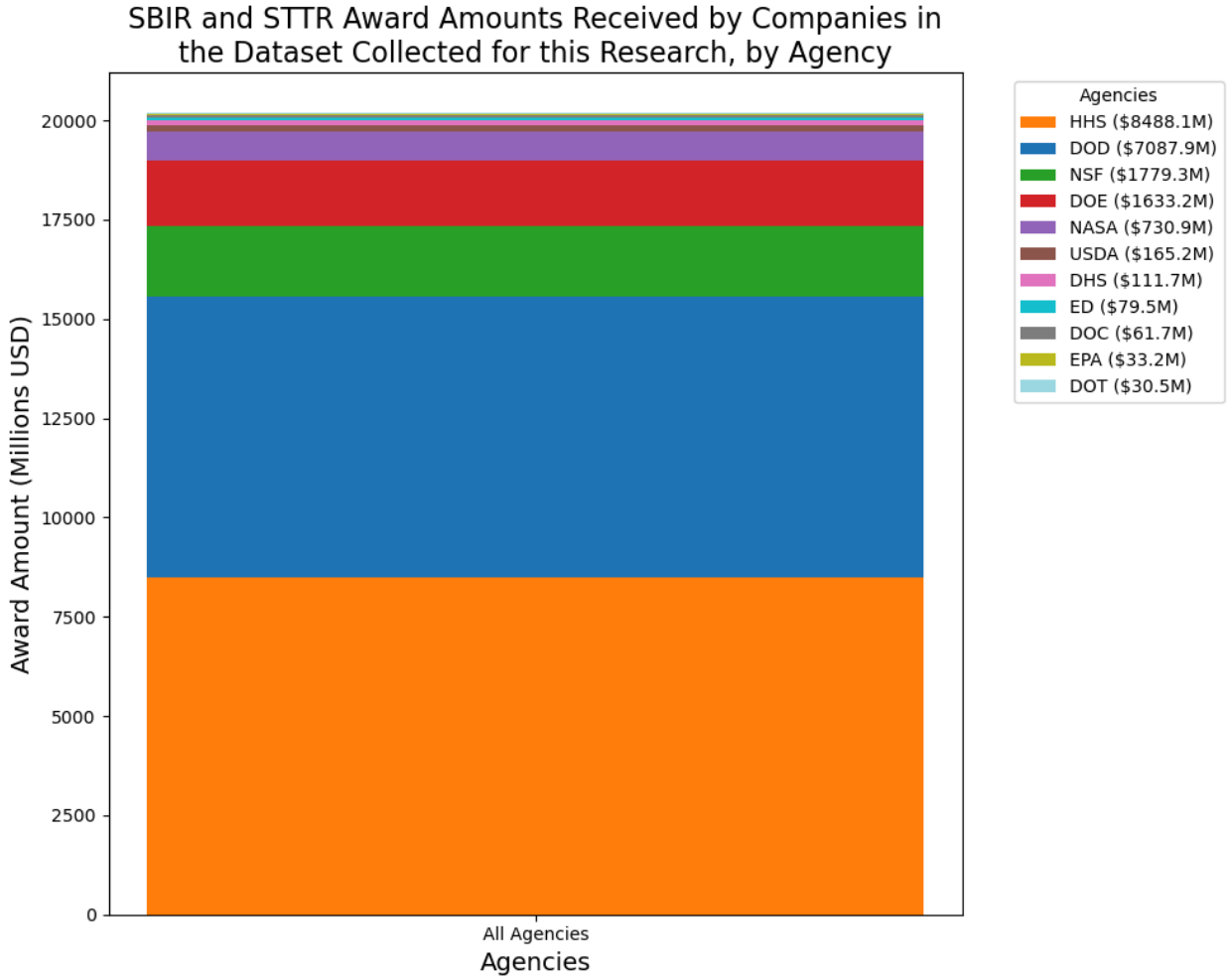

Figure 2.3: Stacked bar chart depicting the total SBIR/STTR award amounts given to companies represented in the dataset for this analysis, separated by agency.

Building on the previous figure, Figure 2.3 shifts the focus from the count of awards to the total monetary value of SBIR and STTR funding allocated by each agency to companies represented in the dataset. The order of the top three agencies—HHS, DOD, and NSF—has changed slightly compared to Figure 2.2.

HHS accounts for the largest share of total funding at approximately $8.5 billion, followed by the DOD at $7.1 billion, and the NSF at $1.8 billion. This distribution highlights an over-representation of HHS in the collected dataset relative to its actual budget for SBIR/STTR programs, where the DOD typically maintains the largest allocation. Conversely, the DOE appears underrepresented compared to NSF in the dataset, despite having a larger budget for SBIR/STTR programs according to official figures[20].

The disparity between the DOD and the HHS can be attributed to two main factors. First, the DOD typically issues smaller awards, particularly in Phase I, where amounts generally range from $50,000 to $250,000. In contrast, the HHS Phase I awards often start at $275,000, resulting in higher total funding amounts for the HHS despite fewer awards. Second, the dataset may disproportionately exclude certain DOD-funded companies, particularly those involved in military applications or classified projects, which are less likely to have publicly available profiles tracked by PitchBook.

A similar discrepancy is evident between the DOE and the NSF. Despite the DOE having a larger overall budget for SBIR/STTR programs ($315 million compared to $215 million for the NSF[20]), the NSF surpasses the DOE in both the number of awards and total dollar amounts represented in the dataset. This may result from the dataset's limitations, which likely underrepresent companies in energy-related fields, particularly those that lack public visibility or are not comprehensively tracked by PitchBook.

These trends highlight variations in the types of companies included in the dataset. The HHS's strong performance may reflect the prominence of health-related startups and biomedical research, which are well-represented in PitchBook. Conversely, the DOE's underrepresentation may be linked to fewer energy-focused companies appearing in the dataset, despite the agency's significant investment in energy innovation. These discrepancies underscore the need to contextualize findings within the limitations of the dataset, particularly its reliance on publicly available data sources.

2.3.3 SBIR/STTR Award Amounts and Counts in Context, Separated by Agency

Figure 2.4 depicts the total award amounts by agency within the dataset collected for this research, compared to the total award amounts by agency for all SBIR and STTR awards

given after 2000. This figure further illustrates the trends observed in the previous figures. The DOD, while accounting for the largest share of total SBIR/STTR funding since 2000, continues to show a significant gap in representation within the dataset used for this research. This aligns with the patterns highlighted in Figures 2.2 and 2.3, where the DOD's funding dominance is evident, yet its representation in PitchBook remains limited. This underrepresentation is likely driven by the nature of DOD-funded projects, which often focus on classified research or specialized military applications and may involve companies that lack the scale or public visibility necessary for PitchBook tracking.

Similarly, this graph underscores the contrast between the DOE and the NSF. As seen in earlier figures, the DOE lags behind the NSF in both award amounts and counts represented in the dataset, despite the DOE's larger SBIR/STTR budget. This underrepresentation may reflect the challenges associated with tracking energy-focused companies, which may prioritize research and development over commercial growth, reducing their likelihood of being captured in PitchBook. By comparison, the NSF's higher inclusion rate—capturing nearly 60% of its total award dollars—suggests the agency may place an emphasis on commercially viable projects and companies with greater public visibility, trends consistently observed in Figures 2.2 and 2.3.

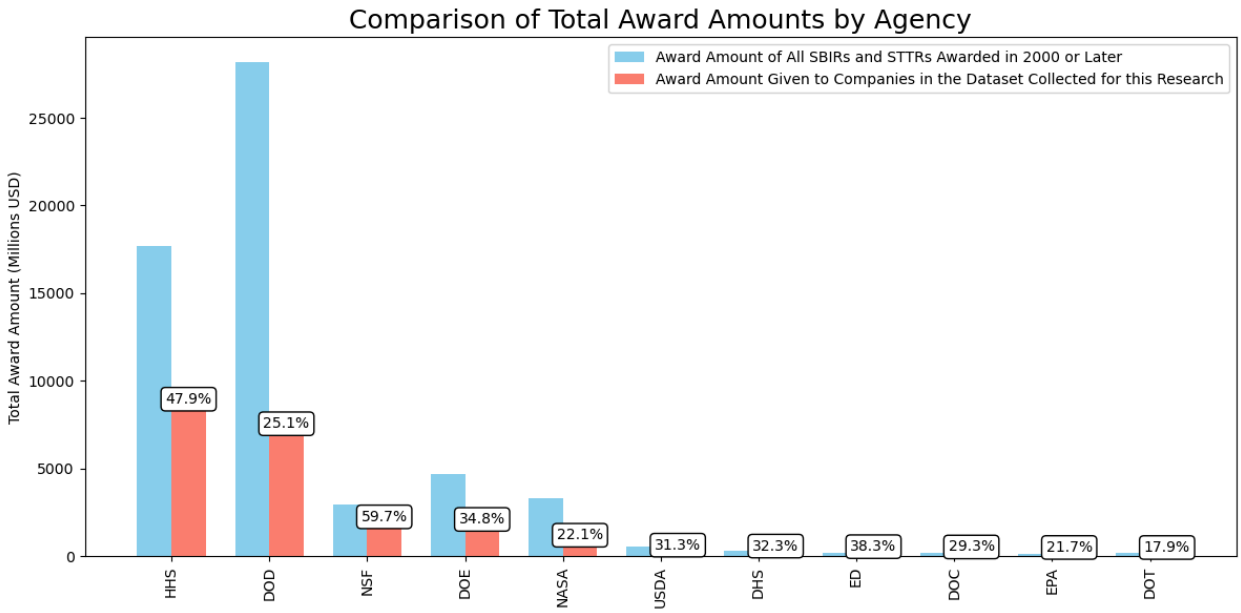

Figure 2.4: Grouped bar chart comparing total award amounts (in millions USD) by agency for all SBIR/STTR awards issued in 2000 or later (blue bars) versus the award amounts awarded to companies represented in the dataset used for this research (red bars). The graph is sorted in order of decreasing award amounts by agency given to the companies in the dataset collected for this research. Percentage labels indicate the proportion of the total award amounts in the dataset relative to all awards given after 2000 for each agency.

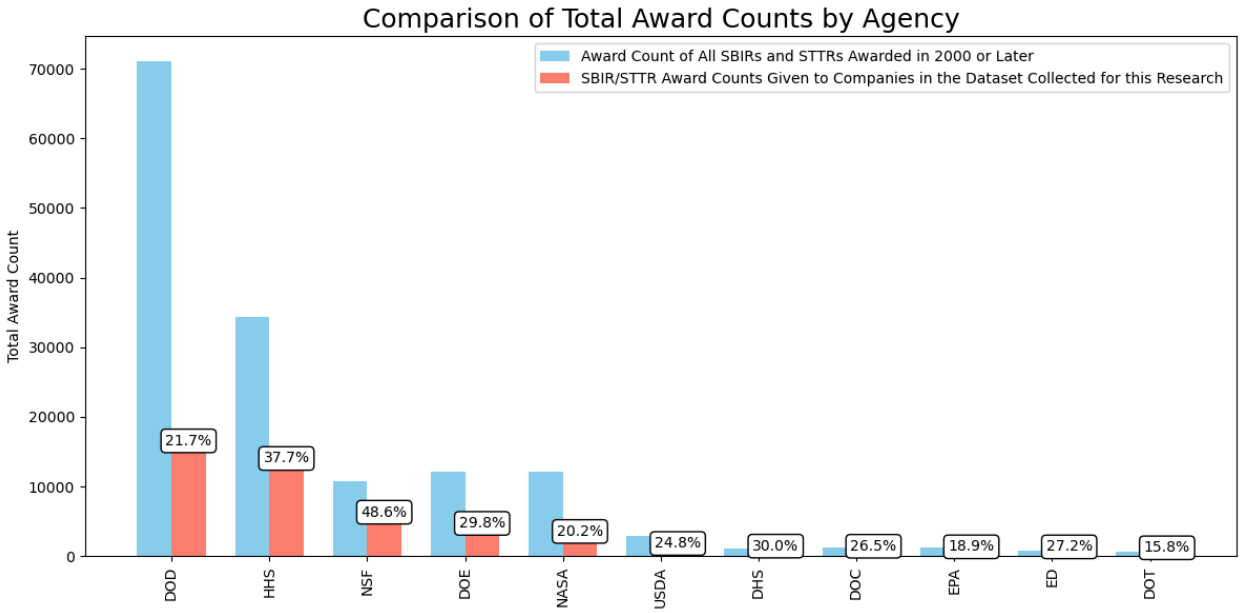

Figure 2.5: Grouped bar chart comparing total award counts by agency for all SBIR/STTR awards issued in 2000 or later (blue bars) versus the award counts given to companies represented in the dataset used for this research (red bars). The graph is sorted in order of decreasing award counts by agency given to the companies in the dataset collected for this research. Percentage labels indicate the proportion of the total award counts in the dataset relative to all awards given after 2000 for each agency.

Figure 2.5 focuses on the comparison of total award counts between the dataset for this research and the overall award counts given after 2000. The figure complements the trends observed in Figure 2.4. The DOD once again dominates in terms of total award counts, but the representation of its awards in the PitchBook dataset is disproportionately low, with only 21.7% of its awards captured. This reinforces the notion that many DOD-funded companies may remain small, operate in niche markets, or lack the public-facing characteristics needed for PitchBook inclusion.

Interestingly, the NSF maintains a consistent representation of nearly 50% across both award amounts and counts. This indicates that a significant portion of NSF-funded companies not only receive higher award amounts but also achieve commercial success or visibility that allows them to appear on PitchBook. On the other hand, agencies such as the DOE and NASA exhibit similar underrepresentation trends in award counts and award amounts, reflecting challenges in tracking companies in highly specialized fields like energy and aerospace.

2.3.4 SBIR vs STTR Distribution

Figure 2.6 provides a side-by-side comparison of SBIR and STTR funding distributions within the dataset collected for this research (left) and across all awards issued since 2000

(right). The dominance of SBIR funding is apparent in both cases, representing 88% of the total funding in the research dataset and 88.8% of the post-2000 funding overall. These proportions correspond to $17.78 billion and $51.95 billion for SBIR funding in the research dataset and overall distributions, respectively. In contrast, STTR funding accounts for a smaller but consistent share, contributing 12% ($2.42 billion) of the total in the research dataset and 11.2% ($6.56 billion) of the overall funding.

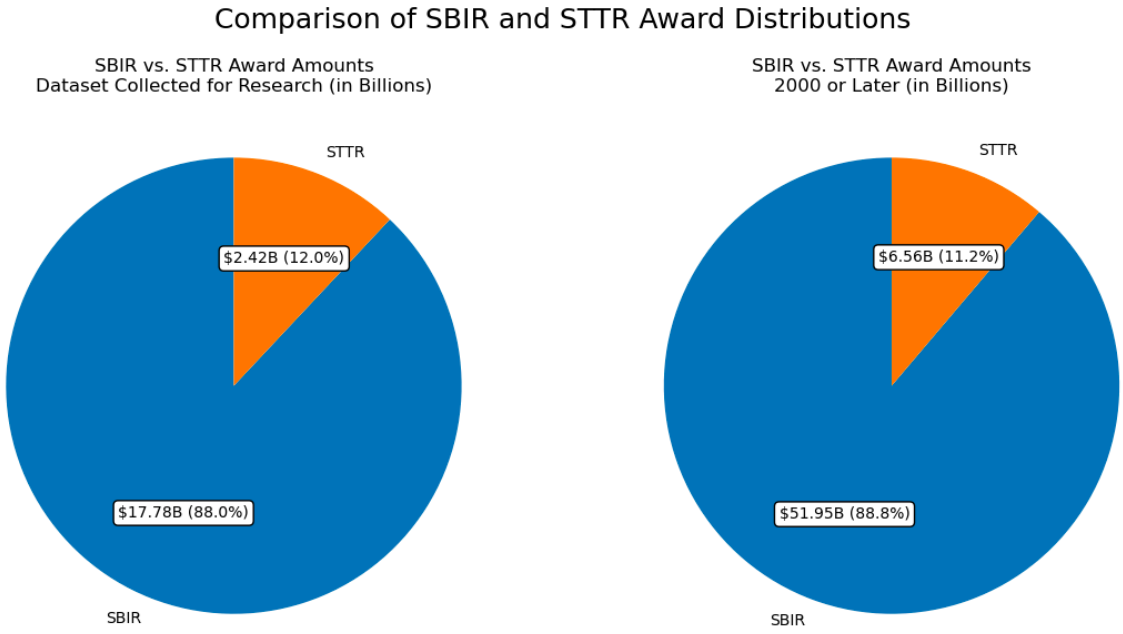

Figure 2.6: Comparison of SBIR and STTR Award Distributions. This figure presents two pie charts that illustrate the distribution of award amounts between SBIR and STTR programs. The left chart represents awards tracked in the dataset collected for this research, while the right chart depicts all SBIR and STTR awards issued in 2000 or later. The blue segments indicate funding allocated to SBIR awards, while the orange segments represent funding for STTR awards.

This close alignment between the two distributions suggests that the dataset collected for this research accurately reflects the broader trends in SBIR and STTR funding allocation. The similarity in proportions indicates that the dataset is representative of the overall award distributions, reinforcing its reliability for analyzing broader patterns in SBIR and STTR funding.

2.3.5 Phase I vs Phase II Distributions

Figures 2.7 and 2.8 illustrate the alignment between the distribution of SBIR and STTR awards in the dataset collected for this research and the overall distribution of awards given since 2000, measured both in terms of award amounts and counts across phases.

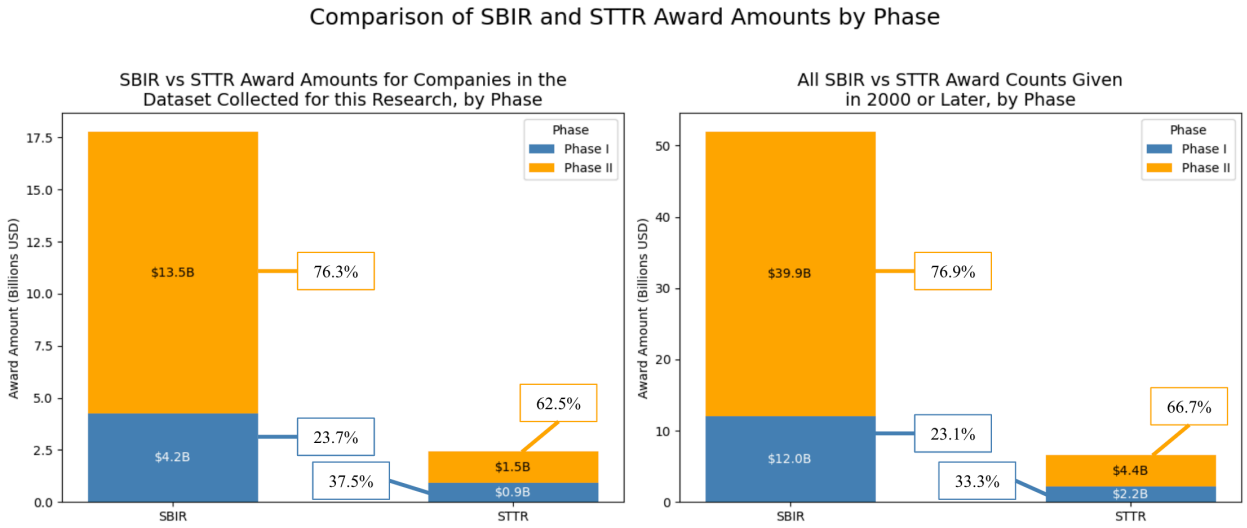

Figure 2.7: Stacked bar charts showing SBIR and STTR award amounts by phase for companies in the research dataset (left) and all awards since 2000 (right). Phase I is blue, and Phase II is orange.

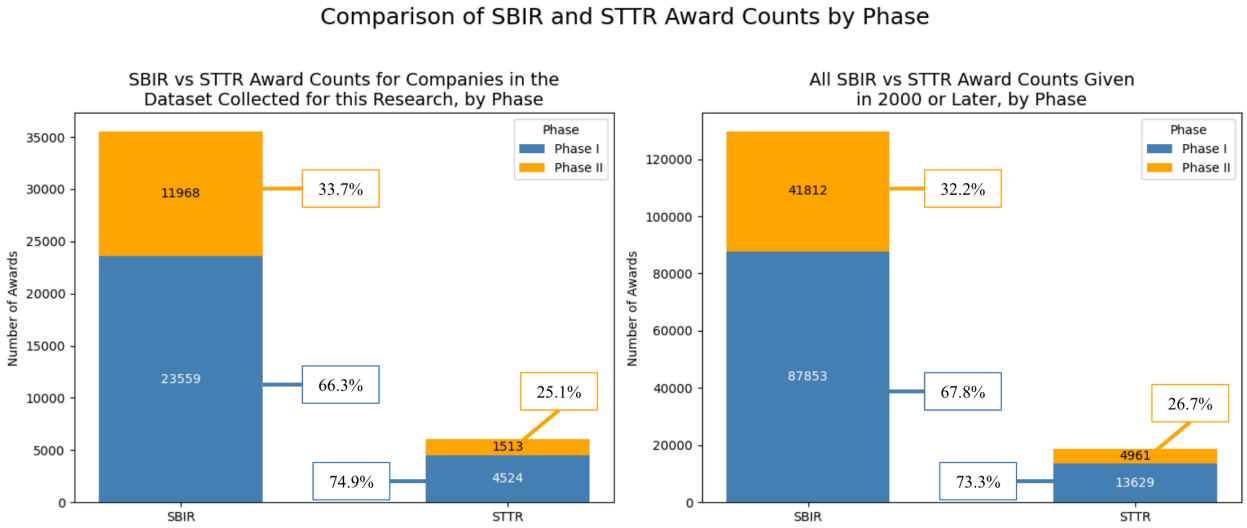

Figure 2.8: Stacked bar charts showing SBIR and STTR award counts by phase for companies in the research dataset (left) and all awards since 2000 (right). Phase I is blue, and Phase II is orange.

In Figure 2.7, the breakdown of award amounts highlights that SBIR consistently dominates both datasets. For companies in the research dataset, Phase I SBIR awards contribute \$4.2B (23.7%) of the SBIR total, while Phase II SBIRs account for \$13.5B (76.3%). The overall dataset that includes all awards from 2000 onwards exhibits similar proportions, with Phase I SBIRs making up \$12.0B (23.1%), and Phase II SBIRs contributing \$39.9B (76.9%).

For STTR, the award distribution follows a similar pattern, with Phase II contributing a larger portion of the total in both datasets—\$1.5B (62.5%) in the research dataset and \$4.4B

(66.7%) overall. The gap between Phase I and Phase II amounts is expected, as agencies may issue a Phase I award up to $314,363 and a Phase II award up to $2,095,748, over six times the Phase I limit[5]. This difference in funding limits explains the significant disparity in monetary contributions from Phase II awards.

Figure 2.8 reveals a similar alignment in the distribution of award counts by phase. In the research dataset, SBIR Phase I accounts for 23,559 awards (66.3%) compared to SBIR Phase II's 11,968 awards (33.7%). The overall dataset exhibits nearly identical proportions, with SBIR Phase I awards at 87,853 (67.8%) and SBIR Phase II at 41,812 (32.2%).

For STTR, the research dataset shows 4,524 Phase I awards (74.9%) and 1,513 Phase II awards (25.1%), while the overall dataset has 13,629 Phase I awards (73.3%) and 4,961 Phase II awards (26.7%). The smaller number of Phase II awards may be attributed to their more stringent eligibility criteria[21], which could lead to a natural reduction in the number of Phase II grants.

These consistent distributions across award counts and amounts reinforce the representativeness of the dataset collected for this research. It reflects broader trends in SBIR and STTR funding and award allocations by phase. The similarity in proportions suggests that the dataset captures the same funding priorities and award structures as the overall SBIR and STTR programs.

2.3.6 SBIR/STTR Awards in Research Dataset: Agency Distribution (First 20 Awards)

Figure 2.9 illustrates the percentage distribution of SBIR/STTR awards over the first 20 awards in the research dataset, broken down by agency. It reveals insightful trends in how agencies allocate their awards across repeated recipients.

One of the most striking observations is the initial dominance of the HHS, which grants over 30% of the first awards in the dataset. However, as the number of repeat awards increases, the DOD steadily rises, surpassing HHS between the 8th and 9th awards. By the 20th award, DOD is awarding about 50% of all awards, while HHS's share has significantly declined. This transition highlights DOD's tendency to fund repeat awardees over time, in contrast to HHS, which awards a larger proportion of new recipients.

Another noteworthy trend is the behavior of the NSF. The NSF contributes over 25% of the first awards, but experiences a sharp drop in its share, declining to about 10% by the third award. This suggests that the NSF may place a strong emphasis on early-stage projects and is less likely to provide repeated funding compared to the DOD or the HHS.

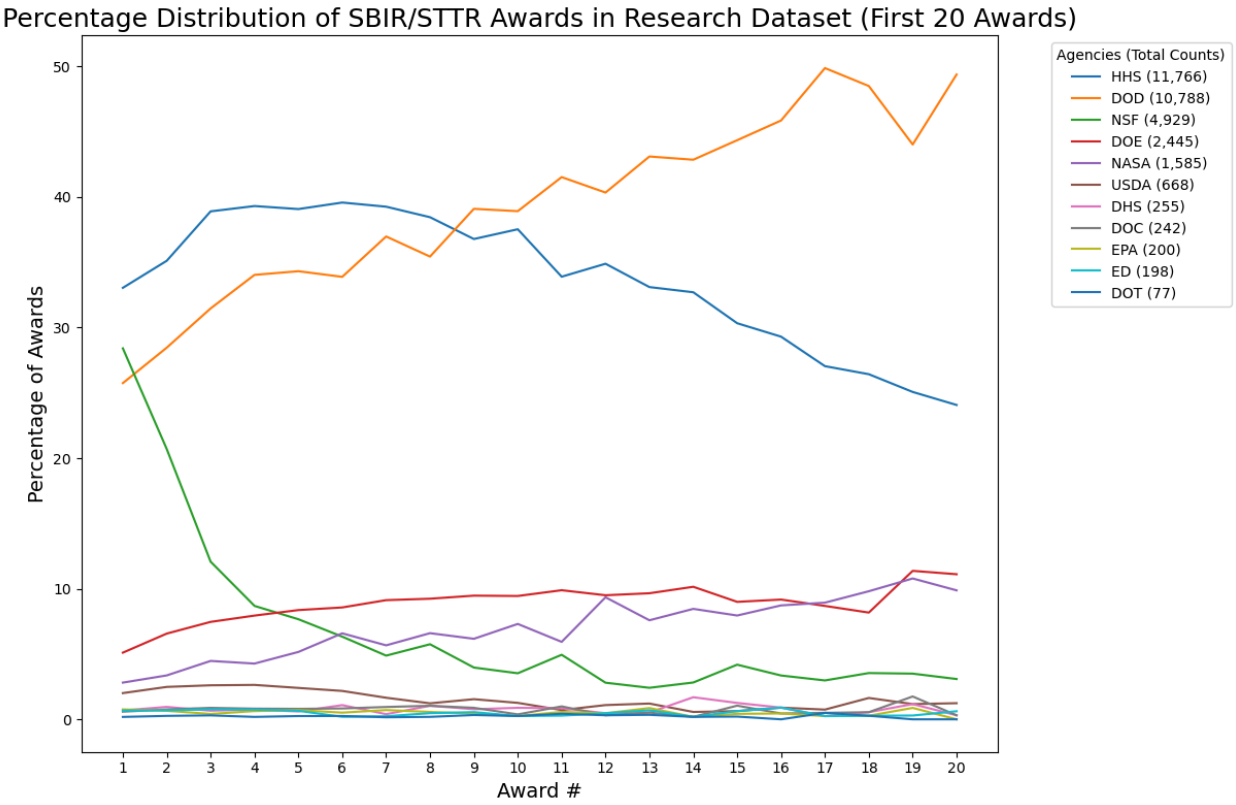

Figure 2.9: Line chart showing the percentage distribution of SBIR/STTR awards in the research dataset by agency for the first 20 awards. This line chart illustrates the percentage share of awards granted by each agency for the first 20 awards received by companies in the dataset, highlighting how agency contributions vary over these early awards.

These patterns point to the strategic value of targeting specific agencies based on their funding behavior. In particular, DOD may appear as a favorable agency for "SBIR mills", entities that consistently win a large number of awards. These entities often prioritize winning government funding over commercializing their innovations, using the grants primarily to sustain operations rather than transitioning to self-sufficiency in the commercial market.

According to the thresholds established in a State Science & Technology Institute analysis[22], a SBIR mill is an organization that has received 40 or more SBIR Phase I awards or 30 or more SBIR Phase II awards. The dataset collected for this research includes 41 such SBIR mills. The DOD's funding practices suggest a higher tolerance for awarding repeat recipients, making it an attractive target for organizations seeking sustained SBIR funding.

2.3.7 Funding Half-Lives, by Agency

Figure 2.10 provides a detailed view of the award frequency trends for different agencies over the first 20 SBIR/STTR awards in the research dataset, using half-life decay curves to model

the rate at which the number of awards diminishes over time. Observed data points for each agency are displayed alongside fitted curves, with the half-life ($t_{1/2}$) indicating the award number at which an agency's frequency of awards decreases by half. For example, agencies like NSF ($t_{1/2} = 0.89$) exhibit a rapid decline, while NASA ($t_{1/2} = 4.76$) shows a much slower rate of decline, granting a relatively steady number of awards across award numbers. The fitted curves provide insights into the funding behavior of each agency, highlighting how quickly their award frequency diminishes over time.

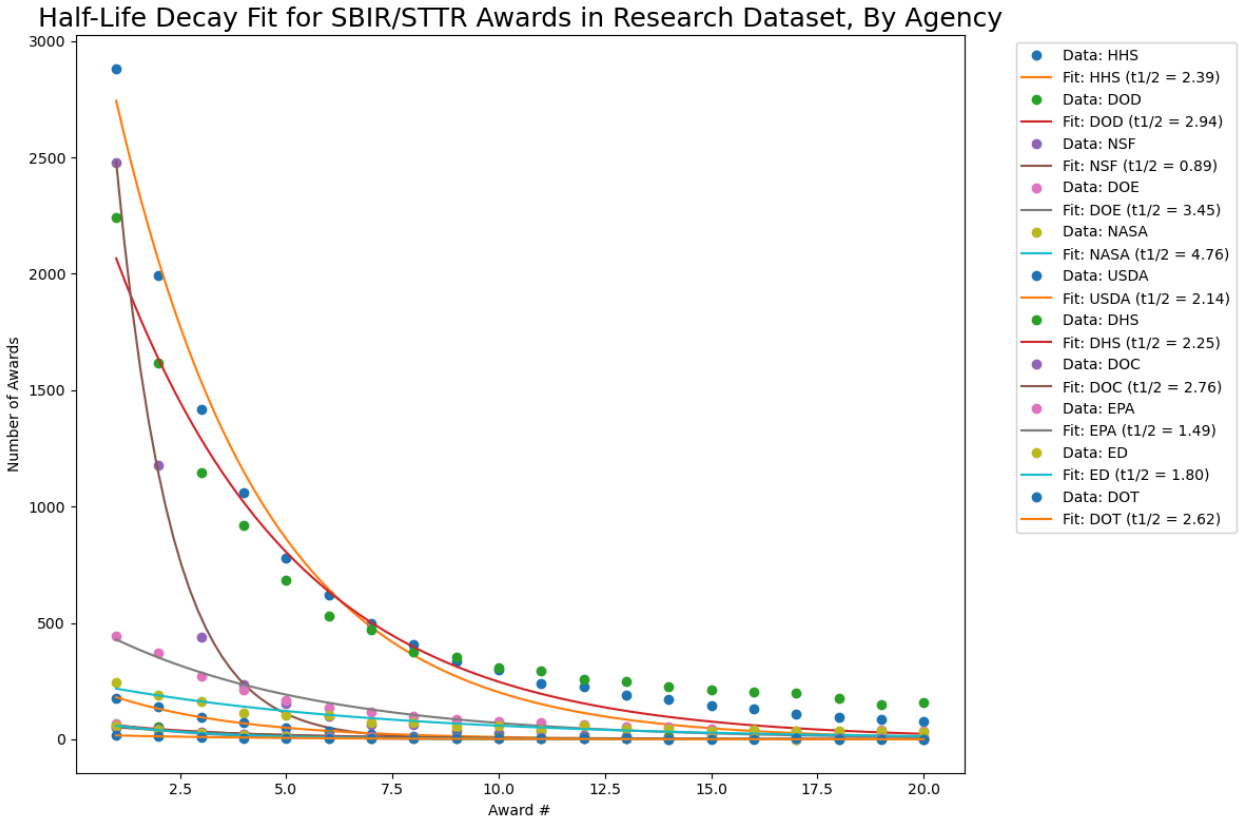

Figure 2.10: Line plot illustrating the half-life decay of SBIR/STTR awards over the first 20 awards in the research dataset, by agency. Observed data points are shown along with fitted half-life decay curves for each agency. The half-life ($t_{1/2}$) represents the number of awards at which the agency's award frequency decreases by half.

When compared to Figure 2.9, which showed that the DOD awards a larger share of repeat awards over time, it is notable that the longest half-life here belongs to NASA. Despite DOD's dominance in repeat awards, NASA's longer half-life ($t_{1/2} = 4.76$) suggests that it consistently grants a relatively steady number of awards, albeit in smaller absolute quantities.

In contrast, the DOD ($t_{1/2} = 2.94$) has a shorter half-life, indicating a more pronounced drop-off in award frequency as the award number increases. These findings align with the

observation that DOD is highly active in granting repeat awards but does so with a more front-loaded pattern compared to NASA's steady behavior.

This analysis underscores the variability in agency funding strategies. NSF's rapid decline in award frequency, indicated by its short half-life ($t_{1/2} = 0.89$), reflects its focus on providing initial funding rather than supporting repeated projects. On the other hand, NASA's slow decline suggests a steady but limited commitment to repeat funding. These patterns guide organizations in targeting agencies with high repeat funding tendencies or consistent award distribution.

2.3.8 State Headquarter Analysis

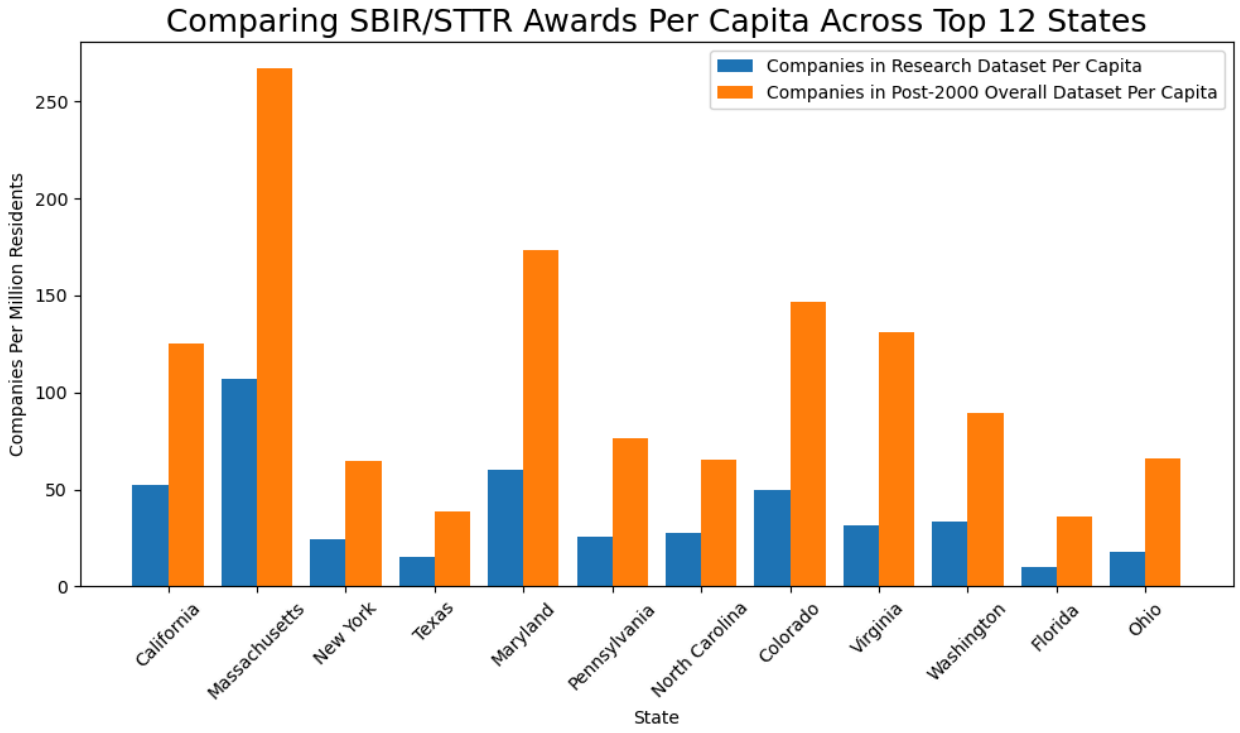

Figure 2.11: Bar chart comparing the per capita distribution of companies awarded SBIR/STTR grants in the top 12 most common states in the research dataset and the overall dataset (all awards after 2000). The metric, "Companies Per Million Residents" normalizes company counts relative to state populations, highlighting which states are leading in fostering SBIR/STTR-supported innovation per capita.

Figure 2.11 compares the per capita distribution of companies awarded SBIR/STTR grants across the top 12 most common states. Each state is evaluated based on two datasets: the research dataset and the post-2000 overall dataset (all SBIR/STTR awards given in 2000 or later), normalized to "companies per million residents". The per capita metrics were

calculated using the 2024 population estimates from the U.S. Census Bureau[23], ensuring that states with smaller populations are fairly compared to larger states.

The x-axis lists the states, ordered by the descending number of companies in each state within the research dataset, starting with California and ending with Ohio. The y-axis measures the number of companies per million residents. The blue bars represent companies from the research dataset per capita, while the orange bars represent companies from the post-2000 overall dataset per capita.

Massachusetts is underrepresented in the research dataset, as its per capita value is significantly lower than the per capita value of the overall post-2000s dataset. This discrepancy suggests that the research dataset might not fully capture Massachusetts' SBIR/STTR activity. Despite this, Massachusetts leads in per capita companies in both datasets, with over 250 companies per million residents in the post-2000 dataset. This leadership could be linked to the influence of renowned academic institutions like MIT and Harvard, which drive innovation through their research programs and strong ties to venture capital[10].

California maintains a high position in raw company counts which could be the result of the influence of Silicon Valley and world-class institutions like Stanford University. These factors contribute to its strong innovation ecosystem, though its per capita metric is moderated by its large population size. Maryland, another standout state, might be benefitting from its proximity to federal agencies such as the NIH and NASA, which drive SBIR/STTR-funded research in biotech and aerospace. Similarly, Virginia, Colorado, and other states in the top 12 leverage specialized industry strengths like defense, cybersecurity, and renewable energy to achieve high per capita metrics.

Figure 2.12 highlights the raw company counts for the same top 12 states across both datasets. The x-axis is ordered by descending company counts in the research dataset. While some states, such as Maryland and Virginia, are underrepresented in the research dataset, the same 12 states appear as the top performers in both datasets. The key difference lies in the ordering of these states, with variations in rankings between the datasets. This consistency across both datasets underscores the leading role of these states in SBIR/STTR funding, while the differences in ranking reflect the distinct characteristics and representation of each dataset.

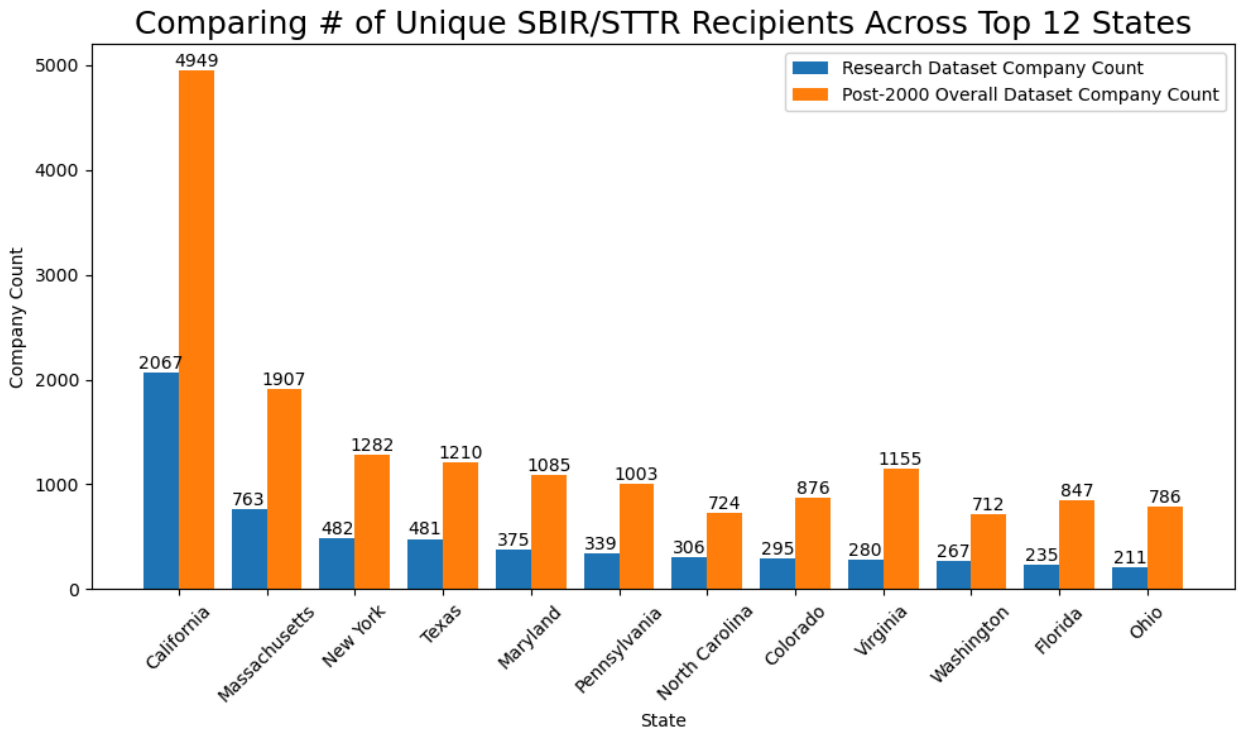

Figure 2.12: Bar chart comparing the number of unique companies awarded SBIR/STTR grants across the top most awarded 12 states, based on two datasets: the research dataset (blue bars) and the post-2000 overall dataset (orange bars). The x-axis is ordered by descending company counts in the research dataset.

Chapter 3

Methods

3.1 Model Selection

For this research, three models were selected: Logistic Regression, Random Forest, and XGBoost. Each model offers distinct advantages that suit the goals of this study.

Logistic Regression was selected as the first model due to its simplicity, interpretability, and effectiveness in binary classification tasks. As a generalized linear model, Logistic Regression predicts probabilities using a logistic function, making it ideal for determining the likelihood of a company receiving an IDV. Its straightforward implementation provides clear insights into the relationships between independent variables and the target variable. This interpretability is valuable for understanding the impact of specific features on the outcome.

Random Forest was chosen for its robustness, versatility, and ability to handle datasets with diverse characteristics effectively. As an ensemble learning method, Random Forest builds multiple decision trees during training and combines their outputs to improve predictive accuracy and reduce overfitting. This makes it well-suited for datasets with a mix of categorical and numerical variables. Additionally, Random Forest inherently performs feature selection by identifying the most important features during training, providing valuable insights into the factors driving predictions. Its ability to handle imbalanced datasets through built-in class weighting further enhances its applicability to this research.

XGBoost was selected for its ability to handle structured data with high-dimensional features and its ability to capture complex relationships between variables. As an implementation of gradient boosting, XGBoost tries to correct errors from previous models iteratively, making it highly effective for datasets with subtle patterns. This iterative learning process enables XGBoost to focus on misclassified instances, improving performance on imbalanced datasets. Its built-in regularization techniques mitigate overfitting, ensuring reliable generalization to unseen data. Additionally, XGBoost allows for the adjustment of class weights in

the objective function, enabling the model to account for class imbalances effectively. These characteristics make XGBoost a powerful tool for identifying and modeling trends among dual-use startups.

Several alternative classification models were considered but ultimately not selected due to their limitations in addressing the specific needs of this research.

Decision trees were not chosen because they are prone to overfitting, especially with deeper trees. Ensemble methods like Random Forest and XGBoost were preferred as they combine multiple decision trees to improve accuracy and robustness, mitigating the limitations of individual decision trees.

Support Vector Machines (SVMs) were also considered but deemed less suitable due to their lack of interpretability. SVMs focus on decision boundaries rather than feature importance, making it challenging to derive insights about the factors influencing startup success. Similarly, K-Nearest Neighbors (KNN) was ruled out because it struggles with high-dimensional data and does not provide insights into feature importance or variable relationships.

Finally neural networks, while powerful for complex relationships, were not selected due to their need for large datasets to perform effectively and avoid overfitting. Additionally, they lack interpretability, which is a critical requirement for understanding the drivers of success among dual-use startups.

This research focuses on models that are well-suited to handle the dataset's complexities while offering interpretability and accuracy.

3.2 Logistic Regression

3.2.1 Initial Feature Selection

An initial scan of the dataset was conducted to identify columns that had fewer than 1% missing values, which resulted in 71 columns deemed sufficiently complete. From these, variables believed to be both business-relevant and predictive were selected. The initial variables chosen are listed below, with the possible categories for categorical variables.

- Numerical Variables
  - **# of Deals:** The number of deals the company received (as found on Pitchbook).
  - **# of Awards:** The number of SBIR and STTR awards the company received.
  - **Founding Year:** The year the company was established.
- Categorical Variables

  - **Company Financing Status:** The type of investors that are backing the company.
    * Accelerator/Incubator Backed
    * Corporate Backed or Acquired
    * Corporation
    * Formerly Accelerator/Incubator backed
    * Formerly PE-Backed
    * Formerly VC-backed
    * Private Debt Financed
    * Private Equity-Backed
    * Venture Capital-Backed
  - **Primary Industry Sector:** The main sector in which the company operates.
    * Business Products and Services (B2B)
    * Consumer Products and Services (B2C)
    * Energy
    * Financial Services
    * Healthcare
    * Information Technology
    * Materials and Resources
  - **First Financing Deal Type:** The first type of financing received by the company. Deal types are listed in Appendix A.
  - **Award 1 Agency:** The federal agency granting the company's first SBIR/STTR award.
  - **Business Status:** Identifies a company's stage in its business or revenue lifecycle.
    * Clinical Trials - General
    * Generating Revenue
    * Out of Business
    * Profitable
    * Startup
  - **HQ State/Province:** The state or province where the company headquarters is located.
- Dependent Variable
  - **IDV:** An indicator for whether or not the company has secured an IDV contract from the government.
    * **IDV = 1:** The company has received an IDV contract.
    * **IDV = 0:** The company has not received an IDV contract.

Other fields were excluded because they either introduced excessive granularity or did not offer explanatory power. Highly correlated features, such as `# of SBIRs`, were also excluded in cases where they were subsets of other variables (e.g., `# of Awards`). Broader categorical variables were chosen over more granular variables as to limit the number of features (e.g. `Primary Industry Sector` was chosen over `Primary Industry Group`). To capture the effect of company age as a numeric feature, `Years Since Founding` was created by subtracting the founding year from 2024. This approach simplified analysis and improved interpretability compared to using raw founding-year values which ranged from 2000 to 2023.

3.2.2 Encoding Categorical Variables

Categorical variables required transformation into a numerical format suitable for Logistic Regression. This was achieved through one-hot encoding (dummy encoding) of the categorical features. The transformed encodings had the name of the original feature followed by an underscore and the value for the category (for example, `Primary Industry Sector_Healthcare`). For each categorical variable, the first category in alphabetical order was omitted to act as a baseline.

3.2.3 Addressing Multicollinearity

Although `Business Status` and `HQ State/Province` were initially included as dummy variables due to their potential predictive significance, they were removed. Multicollinearity diagnostics revealed that these variables exhibited high correlation with other predictors in the model. Specifically, the Variance Inflation Factor (VIF) analysis indicated that the inclusion of `Business Status` and `HQ State/Province` significantly inflated the VIF scores of the model's predictors, signaling severe multicollinearity.

Variance Inflation Factor (VIF)

The Variance Inflation Factor (VIF) is a statistical measure used to evaluate the presence and severity of multicollinearity in a regression model. Multicollinearity occurs when two or more predictor variables are highly correlated, leading to redundancy. This redundancy inflates the variance of the estimated regression coefficients, making them less reliable and harder to interpret. VIF quantifies the degree to which the variance of a regression coefficient is inflated due to multicollinearity[24]. Mathematically, for a predictor X_i, VIF is defined as:

$$VIF(X_i) = \frac{1}{1 - R_i^2} \tag{3.1}$$

R_i^2 represents the coefficient of determination obtained by regressing X_i on all other predictors. A VIF value exceeding 10 is commonly considered indicative of problematic multicollinearity, although the threshold can vary based on context.

Variance Inflation Factor (VIF) values were calculated for all predictor variables to assess the presence of multicollinearity within the model. It was found that the dummy variables `Business Status_Generating Revenue` and `HQ State/Province_California` exhibited exceptionally high VIF values, indicating a severe degree of collinearity with other variables in their respective categories. Upon further examination, these variables were determined to be highly correlated with other dummy variables within the same categorical groups. To simplify the analysis and mitigate the impact of multicollinearity, the entire `Business Status` and `HQ State/Province` categories were excluded from the model.

3.2.4 Addressing Class Imbalance

The distribution of the target variable, `IDV`, was found to be imbalanced, with a disproportionate number of observations having never received an IDV. Class imbalance can lead to biased model performance, as the classifier favor the majority class, reducing its ability to accurately predict the minority class. To combat this issue, sample weights were implemented to balance the influence of each class during model training. The weights were assigned inversely proportional to the class frequencies, calculated as follows:

$$w_1 = \frac{1}{2 \times N_1}, \quad w_0 = \frac{1}{2 \times N_0} \tag{3.2}$$

N_1 and N_0 represent the number of instances in classes 1 and 0, respectively. By applying these weights, the Logistic Regression model treats both classes with equal importance.

3.2.5 Fitting the Logistic Regression Model

The dataset was split into training and test sets with an 80/20 split, respectively. The Logistic Regression model was constructed using Statsmodels' `Logit` function. Statsmodels uses maximum likelihood estimation (MLE) to determine the regression coefficients that maximize the likelihood of observing the given data. This approach estimates the parameters by identifying the values that make the observed outcomes most probable under the logistic

model.

To incorporate regularization, L1 regularization (Lasso penalty) was applied using the `fit_regularized` method and a regularization strength (α) set to 0.01. L1 regularization introduces a penalty proportional to the absolute value of the coefficients. Mathematically, the objective function being minimized becomes:

$$-\log L(\beta) + \alpha \sum_{i=1}^{p} |\beta_i| \tag{3.3}$$

Where $\log L(\beta)$ is the log-likelihood of the model, β_i represents the regression coefficients, and p is the number of predictors. The inclusion of the L1 penalty encourages sparsity in the coefficient estimates by shrinking less important coefficients toward zero. This process effectively performs feature selection, as coefficients that are not strongly associated with the target variable may be reduced to zero, thereby simplifying the model and enhancing interpretability.

The regularization strength parameter (α) controls the extent of this penalty. A higher α value increases the penalty, leading to more coefficients being shrunk toward zero and potentially resulting in a simpler model with fewer predictors. Conversely, a lower α value reduces the penalty, allowing more coefficients to retain their original values.

The previously determined sample weights (`freq_weights=sample_weights`) were integrated into the model fitting process to ensure that each observation contributed proportionally to its assigned weight.

3.3 Random Forest

3.3.1 Initial Feature Selection

The process began with identifying features that had less than 10% missing values. A higher threshold was chosen because Random Forest has inherent feature selection capabilities, which allowed for the inclusion of a larger initial variable set compared to the Logistic Regression model. Categorical variables were then selected for inclusion in the initial model. During this step, First Financing Date and Last Financing Date were excluded due to their substantial impact on the dataset, collectively accounting for 1,231 rows with missing values. To further ensure data quality, rows containing blanks in any of the selected columns were removed, resulting in a reduction of 610 rows. The final set of variables used is shown in Table 3.1. The possible values for the categorical variables can be found in Appendix B.

Table 3.1: List of Features for Random Forest

Features	Features (Cont.)
Primary Industry Sector	Deal 1 Deal Type
Primary Industry Group	# of Deals
Primary Industry Code	Award 1 Agency
Company Financing Status	Award 1 Award Amount
Business Status	Award 1 Award Year
Ownership Status	Award 1 HUBZone Owned
Universe	Award 1 Phase
Year Founded	Award 1 Program
HQ State/Province	Award 1 Socially and Economically Disadvantaged
First Financing Deal Type	Award 1 Solicitation Year
First Financing Deal Class	Award 1 State
Last Financing Deal Type	Award 1 Women Owned
Last Financing Deal Class	# of Awards
Deal 1 Deal Date	# of SBIRs
# of SBIR Phase I	# of SBIR Phase II
# of STTRs Phase I	# of STTR Phase II
# of STTRs	

3.3.2 Data Transformation

The `Year Founded` variable was again transformed into `Years Since Founding` to represent the number of years since the company was established. Similarly, `Deal 1 Deal Date`, which represents the date when the company received its first funding, was converted into a new feature called `Years From Founding to Deal 1`. This variable captures the number of years it took for the company to secure its first funding (this includes grant and venture capital funding). Likewise, `Award 1 Award Year` was used to calculate the time it took for the company to receive its first SBIR or STTR award, creating a new feature called `Years From Founding to Award 1`.

State-based features were modified to address their high cardinality, as they contained over 50 categories, making them unsuitable for use in a Random Forest model. High cardinality features pose challenges in Random Forest because they can lead to sparsity, where categories with few instances provide little information for meaningful splits, increasing the risk of overfitting to specific categories. Additionally, these features can dominate the splitting process, reducing the model's focus on more meaningful features. To address this issue, the some states were grouped together. Table 3.2 summarizes the number of companies that applied for and received their first SBIR/STTR award in each state (plus District

of Columbia and Puerto Rico). The top 12 most common states, each with 200 or more companies, were kept as individual categories. The remaining states were grouped into two broader categories: `group 1` (states with ≥ 100 instances in the dataset) and `group 2` (states with < 100 instances in the dataset).

Table 3.2: Number of Companies per State for Award 1

State	Company Count	State	Company Count	State	Company Count	State	Company Count
CA	1891	MA	711	NY	455	TX	441
MD	351	PA	335	NC	281	VA	277
CO	258	WA	255	IL	211	FL	207
OH	191	MI	182	GA	147	NJ	144
IN	137	WI	124	AZ	123	OR	120
MN	109	UT	99	CT	97	MO	94
KY	64	AL	61	TN	59	NM	58
SC	55	IA	50	DC	46	NH	46
DE	43	OK	35	LA	32	AR	32
ME	31	MT	30	KS	30	RI	27
NE	20	VT	19	HI	19	ID	19
WY	18	NV	17	PR	15	WV	13
SD	10	AK	7	MS	6	ND	3

A similar process was applied to `HQ State/Province`, as listed in PitchBook. The headquarters state may differ from the state listed on the company's first SBIR/STTR award, due to the company relocating. Table 3.3 summarizes the number of companies headquartered in each state/province, including some that moved out of the United States. The top 12 states, each with 200 or more companies, were retained as individual categories. The remaining states were grouped into two broader categories: `group 1` (states with ≥ 100 instances in the dataset) and `group 2` (states with < 100 instances in the dataset).

Table 3.3: Number of Companies in Each HQ State/Province

State	Company Count	State	Company Count	State	Company Count
California	1919	Connecticut	93	Hawaii	19
Massachusetts	725	Missouri	89	Vermont	19
New York	450	Kentucky	68	West Virginia	13
Texas	449	Tennessee	60	Puerto Rico	12
Maryland	345	South Carolina	59	South Dakota	10
Pennsylvania	320	Alabama	56	Mississippi	8
North Carolina	293	New Mexico	54	Alaska	6
Colorado	268	District of Columbia	52	Jiangsu	5
Virginia	262	Iowa	46	England	5
Washington	251	Delaware	43	North Dakota	3
Florida	220	Oklahoma	33	Quebec	3
Illinois	196	New Hampshire	33	Scotland	1
Ohio	192	Maine	30	West Bengal	1
Michigan	176	Arkansas	29	Jiangxi	1
Georgia	152	Kansas	28	New South Wales	1
New Jersey	139	Louisiana	27	Ontario	1
Indiana	131	Rhode Island	27	Australian Capital Territory	1
Wisconsin	125	Nevada	26	Alberta	1
Oregon	117	Montana	26	Guangdong	1
Arizona	116	Nebraska	23		
Minnesota	111	Idaho	21		
Utah	94	Wyoming	19		

3.3.3 Training and Handling Class Imbalance

The dataset was split into training and test sets using an 80-20 ratio to ensure a representative distribution of the data. A baseline model without modifications was first tested.

To address the class imbalance observed in the dataset, four different methods were explored:

- **SMOTE:** Synthetic Minority Oversampling Technique (SMOTE) is a method used to address class imbalance by generating synthetic samples for the minority class. It works by selecting a minority class instance, identifying its nearest neighbors, and generating new synthetic samples along the line segments connecting these points. This technique helps balance the dataset by increasing the representation of the minority class without duplicating existing samples.
- **Undersampling:** Undersampling involves reducing the size of the majority class by randomly removing instances, thereby balancing the class distribution. While this approach reduces class imbalance, it can lead to the loss of potentially valuable information from the majority class if the dataset is small.
- **Built-in Class Balancing from Random Forest:** Random Forest provides an option to handle class imbalance inherently by assigning weights to the classes based

on their frequencies. This approach modifies the splitting criteria during training to give more importance to the minority class, ensuring a more balanced contribution to the model's decision-making.

- **Combined Approach:** A combined approach utilizing SMOTE, undersampling, and built-in class balancing was also attempted to leverage the advantages of each method.

These methods were evaluated to determine their effectiveness in mitigating the class imbalance and improving the model's overall performance. The results of each approach are discussed in subsequent sections.

3.3.4 Hyperparameter Tuning

Hyperparameter tuning was conducted using `RandomizedSearchCV` on the combined approach. Randomized search is a probabilistic technique that evaluates a fixed number of randomly sampled hyperparameter combinations from a specified search space.

The hyperparameters optimized included:

- **Number of Estimators (`n_estimators`):** The number of trees in the Random Forest, sampled randomly from integers between 100 and 500.
- **Maximum Depth (`max_depth`):** The maximum depth of each tree, sampled randomly from integers between 5 and 50.
- **Minimum Samples per Split (`min_samples_split`):** The minimum number of samples required to split an internal node, sampled randomly from integers between 2 and 20.
- **Minimum Samples per Leaf (`min_samples_leaf`):** The minimum number of samples required to be at a leaf node, sampled randomly from integers between 1 and 10.
- **Maximum Features (`max_features`):** The fraction of features considered for splitting, sampled from a uniform distribution between 0.1 and 1.0.
- **Class Weights (`class_weight`):** The weighting of classes to handle imbalance, tested with `"balanced"` and custom weights 0: 1, 1: 5 and 0: 1, 1: 10.
- **SMOTE Neighbors (`smote_k_neighbors`):** The number of nearest neighbors used by SMOTE to generate synthetic samples, sampled randomly from integers between 3 and 10.

The randomized search was performed using 100 iterations, with each iteration randomly selecting a unique combination of hyperparameter values from the defined ranges. A 5-fold cross-validation approach was applied to evaluate each combination. In 5-fold cross-validation, the dataset is split into five subsets, with the model trained on four subsets and

tested on the remaining subset. This process is repeated five times, ensuring that every data point is used for both training and testing. By averaging the results across folds, this method reduces overfitting and provides a reliable estimate of model performance.

The F1 score, a metric that balances precision and recall, was used as the scoring criterion to account for the imbalanced nature of the dataset. The best hyperparameter configuration identified through the randomized search was then used to train the final model.

3.3.5 Feature Importance

Feature importance was utilized to identify the most influential variables contributing to the performance of the optimized combined approach model. The feature importances were derived from the Random Forest classifier within the pipeline, which provides a quantitative measure of each feature's contribution to the model's predictive decisions. After the hyperparameter tuning process, the best model was selected using `RandomizedSearchCV`, and the `feature_importances_` attribute of the Random Forest classifier was accessed to extract importance scores.

3.3.6 Further Feature Selection

Further manual feature selection was conducted using feature importance scores derived from the optimized model. This additional pruning step was motivated by the need to refine the input features further, focusing exclusively on those that had a meaningful impact on the model's performance. Features with an importance score less than 0.005 were excluded, as they contributed minimally to the model's predictive power. This second round of feature selection aimed to enhance the model's interpretability and potentially improve its performance by removing remaining irrelevant or low-importance features that could introduce noise or lead to overfitting. The combined Random Forest model was then retrained on this pruned feature set, and `RandomizedSearchCV` was applied again to fine-tune hyperparameters.

3.4 XGBoost

3.4.1 Training and Handling Class Imbalance

The same methodology used for the Random Forest model was applied to train and evaluate the XGBoost model. The features selected for XGBoost were identical to those used in the Random Forest model to provide a consistent comparison. The dataset was initially split

into training and test sets using an 80-20 ratio to maintain a representative distribution of the data. A baseline XGBoost model was tested to establish a performance benchmark.

To address the class imbalance inherent in the dataset, SMOTE and undersampling techniques were employed. XGBoost was also tested with a combined approach of both SMOTE and undersampling. Unlike the Random Forest model, XGBoost does not include a built-in class balancing mechanism; however, its flexibility allows for adjustments to class weights directly in the objective function.

3.4.2 Hyperparameter Tuning

As with Random Forest, hyperparameter tuning for XGBoost was conducted using Randomized Search Cross-Validation (`RandomizedSearchCV`) on the combined approach. The following hyperparameters were optimized:

- **Number of Estimators (`n_estimators`):** Number of boosting rounds, sampled randomly between 100 and 500.
- **Maximum Depth (`max_depth`):** Maximum depth of each tree, sampled randomly between 3 and 15.
- **Learning Rate (`learning_rate`):** Step size shrinkage to control the contribution of each tree, sampled from a uniform distribution between 0.01 and 0.2.
- **Subsample (`subsample`):** Fraction of samples used for training each tree, sampled randomly between 0.6 and 1.0.
- **Column Subsample by Tree (`colsample_bytree`):** Fraction of features used for training each tree, sampled randomly between 0.6 and 1.0.
- **Gamma (`gamma`):** Minimum loss reduction required for a split, sampled randomly between 0 and 5.
- **Scale Positive Weight (`scale_pos_weight`):** Weighting to handle class imbalance, sampled randomly between 1 and 10.

A total of 100 iterations were performed, with a 5-fold cross-validation strategy used to evaluate each combination of hyperparameters. The F1 score was chosen as the scoring metric to balance precision and recall, given the class imbalance in the dataset. The best hyperparameter combination identified by `RandomizedSearchCV` was used to train the final XGBoost model.

3.4.3 Further Feature Selection

Initial feature selection was followed by further pruning based on feature importance scores, as determined by the combined approach XGBoost model. Features with importance scores

below 0.005 were excluded to reduce noise and improve generalization. The combined XGBoost model was retrained and fine-tuned on this refined feature set to ensure it focused on the most relevant variables.

Chapter 4

Results

4.1 Evaluation Metrics

4.1.1 Precision, Recall, and F1-Score

The performance of the Logistic Regression model was assessed using several standard metrics: precision, recall, F1-score, and accuracy. The formulas for precision, recall, and F1-score are given below:

- **Precision**: Precision measures the proportion of true positive predictions out of all positive predictions made by the model. It is calculated as:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \tag{4.1}$$

- **Recall**: Recall, also known as sensitivity or true positive rate, measures the proportion of true positive predictions out of all actual positive instances. It is given by:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{4.2}$$

- **F1-Score**: The F1-score is the harmonic mean of precision and recall, providing a balanced measure of both metrics. It is calculated as:

$$\text{F-1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.3}$$

4.1.2 Macro and Weighted Averages

The evaluation metrics are also reported in terms of macro average and weighted average to account for class imbalances:

- **Macro Average**: The macro average calculates the arithmetic mean of a metric (e.g., precision, recall, F1-score) across all classes, treating each class equally regardless of its size. It is useful for understanding how the model performs across all classes without considering class proportions.
- **Weighted Average**: The weighted average computes the average of a metric weighted by the number of instances in each class. This provides a more realistic measure of overall performance, particularly in datasets with class imbalances, as it gives greater importance to larger classes.
- **Average**: This term generally refers to the standard arithmetic mean of the metric without additional weighting.

4.2 Logistic Regression

Table 4.1: Coefficients with $p < 0.05$ (Logistic Regression)

Feature	Coefficient	Std. Error	p-value
const	-4.7886	0.4927	0.0000
# of Deals	0.0335	0.0122	0.0061
# of Awards	0.0160	0.0032	0.0000
Years Since Founding	0.1017	0.0092	0.0000
Primary Industry Sector_Energy	-0.6301	0.3128	0.0440
Primary Industry Sector_Healthcare	-0.6058	0.1563	0.0001
Primary Industry Sector_Materials and Resources	-1.1485	0.3209	0.0003
Company Financing Status_Corporation	0.5481	0.2309	0.0176
Company Financing Status_Formerly PE - Backed	0.9196	0.4181	0.0278
Company Financing Status_Private Debt Financed	1.0523	0.3597	0.0034
Company Financing Status_Private Equity – Backed	1.0737	0.2900	0.0002
First Financing Deal Type_Buyout/LBO	0.8413	0.3887	0.0304
First Financing Deal Type_Debt - PPP	1.0372	0.2815	0.0002
Award 1 Agency_Department of Commerce	1.5078	0.6278	0.0163
Award 1 Agency_Department of Defense	1.5659	0.4400	0.0004
Award 1 Agency_Department of Homeland Security	1.8057	0.5532	0.0011

Table 4.1 presents the coefficients, standard errors, and p-values for all variables found to be statistically significant ($p < 0.05$) in the Logistic Regression model. The coefficients in the model indicate the log-odds change in the outcome for a one-unit increase in the respective predictor, holding all other variables constant. Positive coefficients suggest that an increase in the variable is associated with higher odds of getting an IDV, while negative coefficients indicate the opposite.

4.2.1 Threshold Testing

Different decision thresholds were tested to observe their impact on precision, recall, and F1-score. The decision threshold determines the probability cutoff for classifying an instance as positive (predicting a company will receive an IDV). A higher threshold typically increases precision but reduces recall, as the model becomes stricter in assigning positive labels. Conversely, a lower threshold increases recall but may reduce precision by classifying more instances as positive.

4.2.2 Classification Reports For Logistic Regression

Threshold testing revealed trade-offs between precision and recall for the Logistic Regression model. As the decision threshold increased, precision improved, while recall decreased. This behavior is consistent with the model becoming more conservative in assigning positive labels, leading to fewer false positives but potentially more false negatives.

At a threshold of 0.3 (Table 4.2), the F1-score for `IDV` = 0 (not receiving an IDV) was 0.93, while for `IDV` = 1 (receiving an IDV), it was 0.36, resulting in a macro average F1-score of 0.65. The weighted average F1-score was also 0.88, reflecting the model's strong performance on the majority class (`IDV` = 0). However, the low F1-score for `IDV` = 1 highlights its struggle with the minority class, primarily due to the class imbalance, as highlighted by the Support column.

Table 4.2: Classification Report for Threshold = 0.3

Class	**Precision**	**Recall**	**F1-Score**	**Support**
`IDV` = 0	0.93	0.94	0.93	1567
`IDV` = 1	0.38	0.34	0.36	170
Accuracy			0.88	
Macro Avg	0.65	0.64	0.65	1737
Weighted Avg	0.88	0.88	0.88	1737

As the threshold increased to 0.4 (Table 4.3), the F1-score for IDV = 0 slightly improved to 0.94, while the F1-score for IDV = 0 dropped to 0.30. This reduction in the minority class's F1-score was due to a notable decrease in recall (0.22 compared to 0.34 at the 0.3 threshold). The macro average F1-score also declined to 0.62, with a weighted average F1-score remaining consistent at 0.88.

Table 4.3: Classification Report for Threshold = 0.4

Class	**Precision**	**Recall**	**F1-Score**	**Support**
IDV = 0	0.92	0.97	0.94	1567
IDV = 1	0.45	0.22	0.30	170
Accuracy			0.90	
Macro Avg	0.69	0.60	0.62	1737
Weighted Avg	0.87	0.90	0.88	1737

At a threshold of 0.5 (Table 4.4), the F1-score for IDV = 0 further increased to 0.95, while the F1-score for IDV = 1 dropped to 0.24 due to a significant decline in recall (0.15). This resulted in a macro average F1-score of 0.59 and a weighted average F1-score of 0.88. The increase in precision for IDV = 1 (0.55) did not compensate for the drop in recall.

Table 4.4: Classification Report for Threshold = 0.5

Class	**Precision**	**Recall**	**F1-Score**	**Support**
IDV = 0	0.91	0.99	0.95	1567
IDV = 1	0.55	0.15	0.24	170
Accuracy			0.91	
Macro Avg	0.73	0.57	0.59	1737
Weighted Avg	0.88	0.91	0.88	1737

At a threshold of 0.6 (Table 4.5), the F1-score for IDV = 0 remained at 0.95, indicating the model's robustness for the majority class. However, the F1-score for IDV = 1 further dropped to 0.18, driven by a recall of only 0.11. This resulted in the lowest macro average F1-score of 0.57, with a weighted average F1-score remaining steady at 0.88.

Table 4.5: Classification Report for Threshold = 0.6

Class	**Precision**	**Recall**	**F1-Score**	**Support**
IDV = 0	0.91	0.99	0.95	1567
IDV = 1	0.67	0.11	0.18	170
Accuracy			0.91	
Macro Avg	0.79	0.55	0.57	1737
Weighted Avg	0.89	0.91	0.88	1737

The relatively low scores for IDV = 1 observed in this analysis likely stem from two main factors. First, the high class imbalance in the dataset means the model struggles to correctly classify the minority class, which affects recall and F1-score. Second, Logistic Regression assumes linear relationships between predictors and the log-odds of the outcome. This limitation makes it poorly suited for capturing the complexities inherent in many real-world datasets, particularly when interactions or nonlinear relationships exist.

4.3 Random Forest

4.3.1 Baseline Model

The classification report for the baseline Random Forest model at the default threshold of 0.5 is presented in Table 4.6. The model performs well for the majority class (IDV = 0), with a precision of 0.92, recall of 0.99, and F1-score of 0.95 across 1,477 samples. However, its ability to classify the minority class (IDV = 1) is considerably weaker, achieving a precision of 0.54, recall of 0.10, and F1-score of 0.17 for the 144 samples in this class.

The overall accuracy of the model is 91%, which reflects the dominance of the majority class predictions. The macro-average metrics, which assign equal weight to each class, are noticeably lower (precision: 0.73, recall: 0.55, F1-score: 0.56), indicating the model's poor performance on the minority class. On the other hand, the weighted-average metrics, which account for the class imbalance, report higher values (precision: 0.88, recall: 0.91, F1-score: 0.88), driven by the model's strong predictions for the majority class.

Table 4.6: Classification Report for Baseline Random Forest Model

Class	**Precision**	**Recall**	**F1-Score**	**Support**
IDV = 0	0.92	0.99	0.95	1477
IDV = 1	0.54	0.10	0.17	144
Accuracy		0.91		
Macro Avg	0.73	0.55	0.56	1621
Weighted Avg	0.88	0.91	0.88	1621

4.3.2 SMOTE

The classification report in Table 4.7 shows the performance of the Random Forest model trained using SMOTE, using the default threshold of 0.5. Compared to the baseline model, there is a slight improvement in the recall for the minority class, which increases from 0.10 to

0.18. This indicates that the model is able to identify more samples from the minority class after the application of SMOTE. However, the precision for the minority class decreases to 0.43, and the overall F1-score for this class remains relatively low at 0.25.

The macro-average scores (precision: 0.68, recall: 0.58, F1-score: 0.60) show a slight improvement compared to the baseline, indicating that SMOTE provides a modest advantage for addressing class imbalance. The weighted-average scores remain consistent with the baseline, suggesting that the model's overall predictive performance is largely driven by the majority class.

Table 4.7: Classification Report for Random Forest Model with SMOTE

Class	**Precision**	**Recall**	**F1-Score**	**Support**
IDV = 0	0.92	0.98	0.95	1477
IDV = 1	0.43	0.18	0.25	144
Accuracy			0.91	
Macro Avg	0.68	0.58	0.60	1621
Weighted Avg	0.88	0.91	0.89	1621

4.3.3 Undersampling

Table 4.8: Classification Report for Random Forest Model with Undersampling Pipeline

Class	**Precision**	**Recall**	**F1-Score**	**Support**
IDV = 0	0.97	0.75	0.85	1477
IDV = 1	0.23	0.78	0.36	144
Accuracy			0.76	
Macro Avg	0.60	0.77	0.60	1621
Weighted Avg	0.91	0.76	0.81	1621

The classification report in Table 4.8 demonstrates the effects of applying an undersampling pipeline to address class imbalance (with the default threshold of 0.5). The most notable observation is the significant increase in recall for the minority class, which rises to 0.78. This indicates that the undersampling approach enables the model to correctly identify a much larger proportion of the minority class samples compared to both the baseline and SMOTE models. However, this improvement in recall comes at the cost of precision for the minority class, which drops to 0.23, reflecting a higher rate of false positives.

The overall accuracy decreases to 0.76, as the model sacrifices some predictive power on the majority class, with its recall reduced to 0.75. The macro-average recall reaches 0.77,

the highest among the methods evaluated, highlighting the effectiveness of undersampling in achieving better balance between the two classes. However, the macro-average precision and F1-score are relatively low, indicating trade-offs in model performance due to the reduction in the majority class's influence.

The weighted-average scores (precision: 0.91, recall: 0.76, F1-score: 0.81) reflect the overall shift in the model's behavior toward better treatment of the minority class at the expense of the majority class. These results emphasize that while undersampling can effectively improve minority class recall, it may not be the ideal approach for achieving a balance across all performance metrics.

4.3.4 Class Weights Pipeline

Table 4.9: Classification Report for Random Forest Model with Class Weights Pipeline

Class	**Precision**	**Recall**	**F1-Score**	**Support**
`IDV` = 0	0.92	0.99	0.95	1477
`IDV` = 1	0.57	0.09	0.16	144
Accuracy			0.91	
Macro Avg	0.74	0.54	0.55	1621
Weighted Avg	0.89	0.91	0.88	1621

The classification report in Table 4.9 evaluates the performance of the Random Forest model trained using a class weights pipeline with a threshold of 0.5. This approach assigns higher importance to the minority class during training to mitigate class imbalance. While the precision for the minority class improves to 0.57 compared to the baseline, the recall remains extremely low at 0.09, resulting in a limited F1-score of 0.16. This suggests that, although the model reduces false positives for the minority class, it still struggles to correctly identify most instances of this class.

The macro-average metrics (precision: 0.74, recall: 0.54, F1-score: 0.55) indicate a similar balance between the two classes compared to the baseline. The weighted-average scores (precision: 0.89, recall: 0.91, F1-score: 0.88) also remain similar to the baseline, emphasizing that the majority class predictions dominate overall performance.

4.3.5 Combined Approach

The classification report in Table 4.10 demonstrates the performance of the Random Forest model trained with a combined approach at the default threshold of 0.5. The precision for the

minority class was 0.50, which is higher than the precision achieved by undersampling and SMOTE pipeline individually. However, the recall for the minority class remains relatively low at 0.21, limiting the F1-score to 0.29. This suggests that while the combined approach achieves a better balance in reducing false positives for the minority class, it still struggles to correctly identify a substantial proportion of its instances.

Table 4.10: Classification Report for Random Forest Model with Combined Approach

Class	**Precision**	**Recall**	**F1-Score**	**Support**
`IDV` = 0	0.93	0.98	0.95	1477
`IDV` = 1	0.50	0.21	0.29	144
Accuracy			0.91	
Macro Avg	0.71	0.59	0.62	1621
Weighted Avg	0.89	0.91	0.89	1621

The macro-average scores (precision: 0.71, recall: 0.59, F1-score: 0.62) reflect an improvement over the baseline and other single techniques, showing that the combined approach strikes a more favorable balance between the two classes. The weighted-average metrics (precision: 0.89, recall: 0.91, F1-score: 0.89) remain similar to the baseline and other pipelines, indicating that the overall performance is still predominantly influenced by the majority class.

4.3.6 Optimized Parameters

Table 4.11 summarizes the best parameters identified through `RandomizedSearchCV` and the corresponding classification report for the combined approach Random Forest model using the default threshold.

The parameter `class_weight` specifies how much importance is assigned to each class during model training. By setting a ratio of {0: 1, 1: 5}, the minority class is given five times the weight of the majority class, helping the model to focus more on learning the minority class without being dominated by the majority class. The `max_depth` parameter defines the maximum depth of the decision trees in the Random Forest. A depth of 46 allows the model to learn complex patterns in the data while preventing overfitting that can occur with overly deep trees.

The `max_features` parameter, set to approximately 0.142, determines the fraction of features considered when searching for the best split at each node in a decision tree. This ensures that the model doesn't rely too heavily on specific features, improving generalization. The `min_samples_leaf` parameter, set to 3, defines the minimum number of samples required

to form a leaf node. This helps to prevent the model from creating overly specific splits that could lead to overfitting. Similarly, `min_samples_split`, set to 13, specifies the minimum number of samples required to split an internal node, enforcing a level of regularization to avoid splits based on very small subsets of data.

Table 4.11: Best Parameters and Classification Report from Randomized Search CV for Combined Random Forest Model

Best Parameters	**Values**
`classifier__class_weight`	{0: 1, 1: 5}
`classifier__max_depth`	46
`classifier__max_features`	0.1419990968922539
`classifier__min_samples_leaf`	3
`classifier__min_samples_split`	13
`classifier__n_estimators`	154
`smote__k_neighbors`	6

Class	**Precision**	**Recall**	**F1-Score**	**Support**
`IDV` = 0	0.95	0.91	0.93	1477
`IDV` = 1	0.35	0.47	0.40	144
Accuracy			0.87	
Macro Avg	0.65	0.69	0.67	1621
Weighted Avg	0.89	0.87	0.88	1621

The number of estimators (`n_estimators`), set to 154, refers to the number of decision trees in the Random Forest. A larger number of trees typically leads to more robust and stable predictions, although it increases computational cost. Lastly, the `smote__k_neighbors` parameter, set to 6, controls the number of nearest neighbors used by SMOTE to generate synthetic samples for the minority class. This value helps ensure that the synthetic samples are representative of the minority class distribution without being overly dependent on individual observations.

The classification report reflects the impact of these optimizations. For the minority class, the recall increased to 0.47, which is higher than the combined approach without these parameters, showing the model's improved ability to identify positive samples. However, the precision for this class remains relatively low at 0.35, resulting in an F1-score of 0.40. Despite the challenges with false positives, this F1-score for the minority class is the highest achieved among all the evaluated approaches, demonstrating the effectiveness of the hyperparameter tuning in balancing precision and recall for this class.

The macro-average metrics (precision: 0.65, recall: 0.69, F1-score: 0.67) demonstrate an improvement in overall class balance compared to the baseline and other approaches. The weighted-average scores (precision: 0.89, recall: 0.87, F1-score: 0.88) remain consistent with other pipelines, driven by the model's strong performance on the majority class.

4.3.7 Feature Importance

Table 4.12 lists the top 10 features ranked by their importance scores for the optimized combined approach. The feature names reflect the original variable and, for categorical variables, include the specific category as a result of one-hot encoding applied during preprocessing. The most influential feature is `Award 1 Agency_Department of Defense`, an indicator for whether or not the company received its first SBIR or STTR from the DOD. Other top features include `Years Since Founding` (0.0767) and `Award 1 Agency_National Science Foundation` (0.0763), highlighting the relevance of the funding agency and the time since the company's establishment in predicting the outcome.

Table 4.12: Top 10 Features by Importance in the Optimized Combined Random Forest Approach

Feature	**Importance Score**
`Award 1 Agency_Department of Defense`	0.108268
`Years Since Founding`	0.076656
`Award 1 Agency_National Science Foundation`	0.076333
`Award 1 Award Amount`	0.037183
`# of SBIR Phase II`	0.035227
`# of SBIRs`	0.030716
`Years From Founding to Award 1`	0.027952
`# of Awards`	0.026058
`Primary Industry Sector_Healthcare`	0.024737
`# of SBIR Phase I`	0.024314

Additional important features include numeric variables such as `Award 1 Award Amount` and `# of SBIR Phase II`, reflecting the impact of funding amounts and SBIR awards on classification. The inclusion of `Primary Industry Sector_Healthcare` as a top feature emphasizes the significance of the company's industry sector in the model's decision-making process.

4.3.8 Further Feature Selection

The results in Table 4.13 demonstrate the performance of the optimized Random Forest model after further feature selection (filtered for features with importance score greater than .005). Compared to the performance metrics reported in Table 4.11, which represent the model's performance prior to further feature selection, there are some improvements, particularly in metrics related to the minority class.

Table 4.13: Best Parameters and Classification Report from Randomized Search CV for Combined Random Forest Model after Further Feature Selection

Best Parameters	**Values**
`classifier__class_weight`	{0: 1, 1: 5}
`classifier__max_depth`	30
`classifier__max_features`	0.12287721406968567
`classifier__min_samples_leaf`	3
`classifier__min_samples_split`	8
`classifier__n_estimators`	340

Class	**Precision**	**Recall**	**F1-Score**	**Support**
`IDV` = 0	0.95	0.90	0.93	1477
`IDV` = 1	0.35	0.56	0.43	144
Accuracy			0.87	
Macro Avg	0.65	0.73	0.68	1621
Weighted Avg	0.90	0.87	0.88	1621

The feature selection process, combined with systematic hyperparameter tuning, resulted in the following optimized hyperparameters: `class_weight` remained at {0: 1, 1: 5}, emphasizing the importance of the minority class, while the `max_depth` was reduced to 30, limiting the complexity of each tree and mitigating overfitting. Additionally, the `max_features` parameter was reduced to 0.1229, ensuring the model considered only a fraction of the available features at each split, enhancing generalization. The parameters `min_samples_leaf` and `min_samples_split` were set to 3 and 8, respectively, introducing further regularization by requiring a minimum number of samples for splits and leaf nodes. Finally, the number of estimators (`n_estimators`) was increased to 340, leveraging a larger ensemble of decision trees to achieve more stable predictions.

The improvements in minority class performance are significant. The recall for the minority class increased from 0.47 to 0.56, and the F1-score rose from 0.40 to 0.43 after feature selection. This indicates a better balance between precision and recall, showing that the

model is now more effective at identifying minority class samples without a significant increase in false positives. The macro-averaged metrics (precision: 0.65, recall: 0.73, F1-score: 0.68) reflect an overall improvement in class balance, particularly in recall, which increased substantially from 0.69 before feature selection. The weighted-average metrics (precision: 0.90, recall: 0.87, F1-score: 0.88) remained consistent, driven by the model's strong performance on the majority class.

4.4 XGBoost

4.4.1 Baseline Model

The classification report for the baseline XGBoost model (using the default threshold), shown in Table 4.14, presents a performance profile that is comparable to the baseline Random Forest model (Table 4.6). For the majority class, XGBoost achieves a precision of 0.93, recall of 0.98, and F1-score of 0.95 across 1,477 samples. These metrics are nearly identical to those of the Random Forest model, which achieved a precision of 0.92, recall of 0.99, and F1-score of 0.95 for the majority class.

Table 4.14: Classification Report for Baseline XGBoost Model

Class	**Precision**	**Recall**	**F1-Score**	**Support**
`IDV` = 0	0.93	0.98	0.95	1477
`IDV` = 1	0.50	0.20	0.29	144
Accuracy		0.91		
Macro Avg	0.71	0.59	0.62	1621
Weighted Avg	0.89	0.91	0.89	1621

For the minority class, XGBoost shows a modest improvement in recall (0.20 compared to 0.10 for the Random Forest) but sacrifices precision, dropping to 0.50 from the 0.54 achieved by the Random Forest. This results in an F1-score of 0.29 for XGBoost, which is significantly higher than the Random Forest's F1-score of 0.17 for the minority class. The improved F1-score reflects XGBoost's slightly better balance between precision and recall when predicting the minority class, although its performance on this class remains suboptimal overall.

The overall accuracy of both models is 91%, reflecting their strong performance on the majority class. However, the macro-average metrics for XGBoost (precision: 0.71, recall: 0.59, F1-score: 0.62) are higher than those of the Random Forest model (precision: 0.73, recall: 0.55, F1-score: 0.56), primarily due to the improved recall for the minority class. The

weighted-average metrics are slightly higher for XGBoost (precision: 0.89, recall: 0.91, F1-score: 0.89) compared to the Random Forest model (precision: 0.88, recall: 0.91, F1-score: 0.88), driven by the slight improvement in minority class performance.

4.4.2 SMOTE

The classification report for the XGBoost with SMOTE at the default threshold is shown in Table 4.15. When compared to Random Forest with SMOTE (Table 4.7), XGBoost with SMOTE demonstrates similar performance on the majority class, achieving an F1-score of 0.95 and recall of .98, with XGBoost showing slightly higher precision (0.93 vs. 0.92).

For the minority class, XGBoost with SMOTE achieves a precision of 0.54 and a recall of 0.26, resulting in an F1-score of 0.35. This represents a significant improvement over Random Forest with SMOTE, which achieves a precision of 0.43, recall of 0.18, and F1-score of 0.25. The higher recall and F1-score of XGBoost with SMOTE indicate a better balance between precision and recall for the minority class, reflecting its enhanced ability to identify minority samples while managing false positives more effectively.

The macro-average scores further emphasize the advantage of XGBoost with SMOTE. XGBoost with SMOTE achieves a precision of 0.73, recall of 0.62, and F1-score of 0.65, outperforming Random Forest with SMOTE, which has a macro-average precision of 0.68, recall of 0.58, and F1-score of 0.60. This improvement highlights XGBoost's ability to handle class imbalance more effectively, particularly in terms of recall for the minority class. The weighted-average scores, which are influenced by the majority class, remain comparable between the two models, with XGBoost with SMOTE slightly outperforming Random Forest with SMOTE (F1-score of 0.90 vs. 0.89).

Table 4.15: Classification Report for XGBoost Model with SMOTE

Class	**Precision**	**Recall**	**F1-Score**	**Support**
`IDV` = 0	0.93	0.98	0.95	1477
`IDV` = 1	0.54	0.26	0.35	144
Accuracy			0.91	
Macro Avg	0.73	0.62	0.65	1621
Weighted Avg	0.90	0.91	0.90	1621

4.4.3 Undersampling

The classification report in Table 4.16 demonstrates the performance of the XGBoost model trained with undersampling at the default threshold, offering similar results to Random For-

est with undersampling (Table 4.8) with some differences. For the minority class, XGBoost achieves the same high recall of 0.78, indicating that both models are equally capable of identifying a significant proportion of the minority class samples. Similarly, the precision for the minority class is nearly identical (XGBoost: 0.24, Random Forest: 0.23), resulting in the same F1-score of 0.36.

Table 4.16: Classification Report for XGBoost Model with Undersampling

Class	**Precision**	**Recall**	**F1-Score**	**Support**
IDV = 0	0.97	0.75	0.85	1477
IDV = 1	0.24	0.78	0.36	144
Accuracy		0.76		
Macro Avg	0.60	0.77	0.61	1621
Weighted Avg	0.91	0.76	0.81	1621

For the majority class, XGBoost with undersampling has the same performance as Random Forest with undersampling, achieving a precision of 0.97, recall of 0.75, and an F1-score of 0.85. These results reflect the shared trade-offs of undersampling, which reduces the influence of the majority class to improve minority class recall.

In terms of overall metrics, XGBoost with undersampling exhibits no significant improvements over Random Forest with undersampling. The overall accuracy remains the same at 0.76 for both models, while the macro-average precision (0.60), recall (0.77), and F1-score (0.61) are almost identical. The weighted-average metrics, driven by the majority class, remain consistent as well (precision: 0.91, recall: 0.76, F1-score: 0.81).

Overall, XGBoost with undersampling performs on par with Random Forest with undersampling across all metrics. While undersampling enables both models to significantly improve recall for the minority class, XGBoost does not offer additional improvements compared to Random Forest in this specific setup.

4.4.4 SMOTE and Undersampling

The classification report in Table 4.17 highlights the performance of the XGBoost model trained with a combined approach of SMOTE and undersampling with the default threshold. For the majority class, the model achieves a precision of 0.93, recall of 0.98, and an F1-score of 0.95, consistent with the performance observed in the combined approach Random Forest model (Table 4.10). This demonstrates that the combined approach maintains robust performance for the majority class across both algorithms.

Table 4.17: Classification Report for XGBoost Model with SMOTE and Undersampling

Class	Precision	Recall	F1-Score	Support
IDV = 0	0.93	0.98	0.95	1477
IDV = 1	0.53	0.28	0.37	144
Accuracy			0.91	
Macro Avg	0.73	0.63	0.66	1621
Weighted Avg	0.90	0.91	0.90	1621

For the minority class, combined XGBoost outperforms the combined Random Forest model, achieving a precision of 0.53, recall of 0.28, and an F1-score of 0.37. Compared to the combined Random Forest model, which achieved a precision of 0.50, recall of 0.21, and F1-score of 0.29, combined XGBoost demonstrates a better ability to identify minority class samples while maintaining a similar level of precision. The higher recall contributes to the improved F1-score, indicating that XGBoost achieves a better balance between precision and recall for the minority class under this combined approach.

The macro-average metrics further illustrate the advantage of XGBoost. It achieves a macro-average precision of 0.73, recall of 0.63, and F1-score of 0.66, surpassing the combined Random Forest model's scores of 0.71, 0.59, and 0.62, respectively. This improvement highlights XGBoost's capability to handle class imbalance more effectively while maintaining strong overall performance. The weighted-average metrics for XGBoost (precision: 0.90, recall: 0.91, F1-score: 0.90) are slightly higher than those of the Random Forest model (precision: 0.89, recall: 0.91, F1-score: 0.89), reflecting the increased influence of the improved minority class performance on the overall results.

4.4.5 Optimized Parameters

The results in Table 4.18 show the performance of the optimized XGBoost model trained with a combined approach (SMOTE and undersampling) and tuned using `RandomizedSearchCV`. The results can be compared to the optimized combined Random Forest model results in Table 4.11. Both models achieved strong performance for the majority class, with XGBoost slightly outperforming Random Forest in terms of F1-score (0.94 vs. 0.93) due to identical precision (0.95 vs. 0.95) and higher recall (0.93 vs. 0.91).

For the minority class, XGBoost demonstrated a marginally higher F1-score (0.44 vs. 0.40) compared to Random Forest. This improvement is driven by XGBoost's slightly higher recall (0.49 vs. 0.47) and precision (0.39 vs. 0.35) for the minority class, indicating a better ability to identify minority class samples.

The overall accuracy of the XGBoost model was 0.89, slightly higher than Random Forest's accuracy of 0.87. Similarly, the macro-average metrics for XGBoost (precision: 0.67, recall: 0.71, F1-score: 0.69) outperformed those of Random Forest (precision: 0.65, recall: 0.69, F1-score: 0.67). These improvements indicate that XGBoost achieved a better balance between the two classes while maintaining high performance for the majority class.

The weighted-average metrics, which account for class imbalance, also show a slight edge for XGBoost, with precision, recall, and F1-score values of 0.90, 0.89, and 0.89, respectively, compared to 0.89, 0.87, and 0.88 for Random Forest.

Table 4.18: Best Parameters and Classification Report from Randomized Search CV for Combined XGBoost Model

Best Parameters	**Values**
`classifier__colsample_bytree`	0.9355734088277453
`classifier__gamma`	4.302023091558376
`classifier__learning_rate`	0.06005027210317281
`classifier__max_depth`	12
`classifier__min_child_weight`	6
`classifier__n_estimators`	370
`classifier__scale_pos_weight`	4.601906414112629
`classifier__subsample`	0.6508242050607539

Class	**Precision**	**Recall**	**F1-Score**	**Support**
`IDV` = 0	0.95	0.93	0.94	1477
`IDV` = 1	0.39	0.49	0.44	144
Accuracy			0.89	
Macro Avg	0.67	0.71	0.69	1621
Weighted Avg	0.90	0.89	0.89	1621

4.4.6 Feature Importance

Table 4.19 lists the top 10 features by importance in the optimized combined XGBoost model. The feature with the highest importance score is `Award 1 Agency_Department of Defense` (0.101565), making it the most significant contributor to the model's predictions.

The next two most important features are `Primary Industry Group_Pharmaceuticals and Biotechnology` (0.025380) and `Primary Industry Code_Biotechnology` (0.023588), followed by `Years Since Founding` (0.015593). Geographic features such as `Award 1 State Group_Group 2` (0.013219) and `Award 1 State Group_CA` (0.011980) also rank among the top contributors.

Temporal variables, including `Award 1 Solicitation Year_2012` (0.012429) and `Award 1 Solicitation Year_2013` (0.012307), along with `Award 1 Agency_National Science Foundation` (0.012196) and `Deal 1 Deal Type_Grant` (0.012192), round out the top 10.

Table 4.19: Top 10 Features by Importance in the Optimized Combined XGBoost Approach

Feature	Importance Score
`Award 1 Agency_Department of Defense`	0.101565
`Primary Industry Group_Pharmaceuticals and Biotechnology`	0.025380
`Primary Industry Code_Biotechnology`	0.023588
`Years Since Founding`	0.015593
`Award 1 State Group_Group 2`	0.013219
`Award 1 Solicitation Year_2012`	0.012429
`Award 1 Solicitation Year_2013`	0.012307
`Award 1 Agency_National Science Foundation`	0.012196
`Deal 1 Deal Type_Grant`	0.012192
`Award 1 State Group_CA`	0.011980

4.4.7 Further Feature Selection

Table 4.20 presents the best parameters and classification report for the XGBoost model after further feature selection with the default threshold. The feature selection process involved pruning features with low importance scores, resulting in a smaller, more refined feature set. This modification led to notable changes in both the best hyperparameters and model performance.

After further feature selection, the hyperparameters identified by `RandomizedSearchCV` shifted. For instance, the number of estimators (`n_estimators`) decreased from 370 to 249, while the maximum depth of trees (`max_depth`) was reduced from 12 to 3. These changes suggest that the reduced feature set allowed the model to achieve optimal performance with a simpler architecture, likely reducing the risk of overfitting. Other parameters, such as `colsample_bytree`, `subsample`, and `scale_pos_weight`, also adjusted to account for the updated feature space, reflecting a shift in the model's focus during training.

For the majority class, the precision, recall, and F1-score remained relatively stable between the two models. This indicates that the pruning process had minimal impact on the model's ability to predict the majority class accurately.

In contrast, the minority class saw notable improvements. The recall increased from 0.49 to 0.57, demonstrating the model's enhanced ability to correctly identify minority class instances after feature selection. However, the precision for the minority class decreased slightly from 0.39 to 0.35, reflecting a small trade-off of more false positives. The F1-score

for the minority class remained consistent at 0.44, showing that the improvements in recall offset the minor loss in precision.

Table 4.20: Best Parameters and Classification Report from Randomized Search CV for Combined XGBoost Model after Further Feature Selection

Best Parameters	**Values**
`classifier__colsample_bytree`	0.7424386903591385
`classifier__gamma`	4.932576243964898
`classifier__learning_rate`	0.13115496387137743
`classifier__max_depth`	3
`classifier__min_child_weight`	5
`classifier__n_estimators`	249
`classifier__scale_pos_weight`	3.253279771252825
`classifier__subsample`	0.7461427278696231

Class	**Precision**	**Recall**	**F1-Score**	**Support**
`IDV` = 0	0.96	0.90	0.93	1477
`IDV` = 1	0.35	0.57	0.44	144
Accuracy			0.87	
Macro Avg	0.65	0.73	0.68	1621
Weighted Avg	0.90	0.87	0.88	1621

The overall accuracy of the model decreased slightly, from 0.89 before feature selection to 0.87 after. The macro-average metrics also decreased, with F1-score decreasing from 0.69 to 0.68 and precision going from 0.67 to 0.65. The weighted-average metrics were largely unchanged, reflecting the dominance of the majority class in the overall performance.

Chapter 5

Discussion

5.1 Logistic Regression

The coefficients for the Logistic Regression model presented in Table 4.1 shed light on the key factors influencing the likelihood of success, as measured by obtaining IDV contracts. The intercept of -4.7886 suggests a low baseline log-odds of success when all predictors are at their reference levels. Among the numerical predictors, `# of Deals` and `# of Awards` were positively associated with success, indicating that companies that received more deals or more SBIR/STTR awards are more likely to succeed. These findings highlight the importance of financial backing and government recognition in enhancing a company's odds of achieving sustained government contracts. Additionally, `Years Since Founding` was positively associated with success, suggesting that older companies may benefit from an established presence, network, and industry maturity.

The influence of industry sectors on success varied significantly. Companies in the energy, healthcare, and materials and resources sectors exhibited lower chances of success compared to the baseline sector of business products and services (B2B). These negative associations suggest that certain industry-specific challenges could hinder success in these domains.

Certain financial statuses also play a pivotal role in predicting success. Companies classified as corporations, formerly private equity-backed, private debt-financed, or currently private equity-backed demonstrated substantially higher log-odds of success compared to the baseline category of accelerator-backed. These findings emphasize the importance of financial structure and access to strategic resources in navigating the complex government procurement landscape. Federal agency funding also appears to be a critical driver of success. Companies receiving their first SBIR/STTR awards from the Department of Commerce, Department of Defense, and Department of Homeland Security had significantly higher odds of success than the baseline agency, Department of Agriculture.

However, while the model provides valuable insights, its limitations must be acknowledged. The low recall and low precision for the minority class across thresholds indicate that the model struggles to accurately identify successful companies. This limitation suggests that the model may fail to capture nuanced patterns in the minority class, particularly given the significant class imbalance in the dataset.

5.2 Random Forest

The Random Forest model identified several features as strongly correlated with dual-use startup success, despite difficulties in accurately classifying the positive class. These predictors provide valuable insights into the characteristics that contribute to a startup's potential for success. Among the most significant features was the awarding agency of the first SBIR/STTR award, with the DOD emerging as the most important predictor and the NSF following closely behind. Startups receiving their first award from the DOD may benefit from the DOD's tendency to foster continued relationships, as seen in Figure 2.9.

Time-related features provided additional insights into the dynamics of government contract acquisition. The number of years since founding was positively correlated with success, potentially reflecting the accumulation of experience, stability, and refinement of a startup's operations over time.

Additionally, the number of SBIR Phase II awards and the total count of SBIR awards were significant. Phase II awards, which focus on prototype development and scaling, are particularly critical for bridging the gap between research and commercialization. The importance of Phase I awards further emphasized the foundational role of early feasibility studies in determining long-term outcomes.

Despite the valuable insights provided by the Random Forest model, there were several limitations. Class imbalance posed a significant challenge, as the minority class representing successful startups remained difficult to predict accurately. Even with techniques such as SMOTE and class weighting, many successful startups were still misclassified. This imbalance may have skewed the feature importance rankings toward characteristics of the majority class. Furthermore, while precision for the minority class improved slightly, the rate of false positives remained high, which could lead to inefficiencies in resource allocation if the model's predictions were applied in practice.

The interpretability of feature importance rankings also presents limitations. While the identified features provide a general understanding of the factors correlated with success, they do not capture complex interactions between variables or establish causal relationships. For example, the relationship between funding amount and success may depend on unobserved

factors such as market conditions or the quality of a startup's management team. Moreover, the model does not account for temporal changes, such as evolving government funding priorities or industry trends, which may impact the generalizability of the findings to future startups.

5.3 XGBoost

The results of the combined XGBoost model with hyperparameter tuning provide valuable insights into the factors influencing success in dual-use startups. Among all features, `Award 1 Agency_Department of Defense` (importance score = 0.101565) emerged as the most critical predictor, emphasizing the central role of DOD funding in shaping outcomes. However, beyond this feature, the importance scores of other predictors are relatively low, suggesting that the model relies heavily on the DOD feature to make predictions.

Despite achieving a high overall accuracy, the model's classification performance highlights important limitations, particularly for the minority class (class 1, representing successful startups). Examining the optimized combined approach after further feature selection, the recall for the minority class was 0.57, the precision was only 0.35. This indicates that the model identifies a reasonable proportion of successful startups but struggles with false positives. This trade-off reflects challenges inherent to the class imbalance in the dataset.

5.4 Comparative Analysis

Table 5.1 provides a comparison of the performance metrics for the minority class across the different models and methods used. The table highlights the best performance for each metric in bold. The method with the highest macro F1-score is highlighted in bold as well.

Logistic Regression demonstrated varying performance across different threshold levels. At a threshold of 0.3, it achieved its highest F1-score (0.36) for the minority class, but this was accompanied by relatively low recall (0.34). Increasing the threshold to 0.6 led to an improvement in precision (0.67), but at the expense of recall (0.11), resulting in a lower overall F1-score (0.18). These results indicate that Logistic Regression struggles to balance precision and recall effectively, making it less suitable for handling class imbalance without additional interventions.

Table 5.1: Minority Class Metrics for Each Model and Method

Model	Method	Precision (Minority)	Recall (Minority)	F1-Score (Minority)	Macro F1-Score
Logistic Regression	**Threshold = 0.3**	0.38	0.34	**0.36**	**0.65**
Logistic Regression	Threshold = 0.4	0.45	**0.22**	0.30	0.62
Logistic Regression	Threshold = 0.5	0.55	0.15	0.24	0.59
Logistic Regression	Threshold = 0.6	**0.67**	0.11	0.18	0.57
Random Forest	Baseline	0.54	0.10	0.17	0.56
Random Forest	SMOTE	0.43	0.18	0.25	0.60
Random Forest	Undersampling	0.23	**0.78**	0.36	0.60
Random Forest	Class Weights	**0.57**	0.09	0.16	0.55
Random Forest	Combined Methods	0.50	0.21	0.29	0.62
Random Forest	Optimized Combined	0.35	0.47	0.40	0.67
Random Forest	**Optimized Combined + Further Feature Selection**	0.35	0.56	**0.43**	**0.68**
XGBoost	Baseline	0.50	0.20	0.29	0.62
XGBoost	SMOTE	**0.54**	0.26	0.35	0.65
XGBoost	Undersampling	0.24	**0.78**	0.36	0.61
XGBoost	SMOTE + Undersampling	0.53	0.28	0.37	0.66
XGBoost	Combined Methods	0.53	0.28	0.37	0.66
XGBoost	**Optimized Combined**	0.39	0.49	**0.44**	**0.69**
XGBoost	Optimized Combined + Further Feature Selection	0.35	0.57	**0.44**	0.68

Random Forest exhibited significant improvements in recall when undersampling was applied, achieving the highest recall (0.78) among all models. However, this came at the cost of reduced precision (0.23), leading to a lower F1-score (0.36). When combined methods with optimized hyperparameters were used after further feature selection, Random Forest achieved a significantly better balance than Logistic Regression, with an F1-score of 0.43 (compared to 0.36 from Logistic Regression) and a macro-average F1-score of 0.68, demonstrating its potential as a robust model when properly tuned.

XGBoost consistently outperformed both Logistic Regression and Random Forest in identifying the minority class, achieving the highest F1-score (0.44) for the minority class under the optimized combined method with further feature selection. It also achieved the best macro-average F1-score (0.69). XGBoost's ability to handle class imbalance effectively through combined rebalancing approaches and optimized hyperparameters underscores its suitability for this dataset.

5.5 Limitations

While overall metrics were high, the performance metrics for correctly classifying the minority class remained low. These poor performance metrics underscore the limitations of relying solely on the available features to capture the nuanced drivers of success. A key limitation of this research was the availability and quality of data from PitchBook, which constrained the depth and breadth of the analysis. While PitchBook provides detailed data on funding, industry classifications, and startup characteristics, it lacks qualitative information on factors such as leadership quality, strategic partnerships, and innovation ecosystems—factors likely

critical for understanding dual-use startup success.

Additionally, the data in PitchBook may suffer from inconsistencies or gaps, particularly for startups that do not regularly report updates or for industries where comprehensive tracking is more difficult. For example, variations in the granularity and accuracy of industry classifications and funding information could lead to biases or inaccuracies in the model. Furthermore, the reliance on historical funding data from PitchBook means that external factors, such as changes in government priorities or economic conditions, are not adequately reflected, limiting the model's ability to capture the dynamic nature of dual-use startup ecosystems.

Future research should address these limitations by supplementing PitchBook data with additional sources, such as founder interviews, case studies, or publicly available datasets on government contracting and market trends. Incorporating longitudinal data from multiple sources would provide a more comprehensive view of how startups evolve and succeed over time. Moreover, improving the quality and consistency of input data—either through cross-validation with other databases or through manual verification—could help further refine the model's predictive power. These improvements would enable a more holistic understanding of the drivers of dual-use startup success and enhance the utility of the findings for policymakers and investors.

Chapter 6

Conclusion

This research provides valuable insights into the factors influencing the success of dual-use startups, specifically their ability to secure Indefinite Delivery Vehicle (IDV) contracts. Across all models evaluated—Logistic Regression, Random Forest, and XGBoost—the findings indicate that the awarding agency of the first SBIR/STTR grant has a measurable impact on startup success. This finding underscores the importance of early-stage funding mechanisms like SBIR/STTR awards and highlights the unique alignment between dual-use innovations and government priorities.

Among the models, XGBoost demonstrated the highest overall performance, achieving the best macro-average F1-score and the highest F1-score for the minority class (successful startups). Its robust handling of class imbalance enabled it to outperform Logistic Regression and Random Forest consistently. While Random Forest achieved the highest recall for the minority class, its lower precision led to reduced overall F1-scores, highlighting its limitations in balancing the trade-off between recall and precision effectively. XGBoost's superior performance underscores its capability to address the challenges posed by the dataset, particularly class imbalance.

The analysis also identified other features, such as the total number of awards, the time since founding, and the financial structure of startups, as important contributors to success. However, further research is needed to explore the nature of these relationships and understand why they influence outcomes.

Despite these insights, the study revealed significant limitations in the predictive models, particularly in accurately identifying successful startups within the minority class. Both Random Forest and XGBoost struggled with recall and precision for this class, underscoring the challenges of class imbalance and the need for more sophisticated methods to capture nuanced patterns in the data. Additionally, the reliance on PitchBook data, while providing a foundational dataset, highlighted key gaps in the analysis. The absence of qualitative

metrics, such as leadership quality, strategic partnerships, and market dynamics, limited the models' ability to fully capture the drivers of success. Inconsistencies in data quality and the lack of temporal context further constrained the generalizability of the findings.

By identifying critical predictors and highlighting the gaps in existing data and methods, this study lays the groundwork for further exploration into the dual-use startup ecosystem. It emphasizes the need for a multi-dimensional approach that accounts for both quantitative and qualitative factors, ultimately contributing to the development of more effective strategies for fostering innovation and commercialization in this unique domain.

6.1 Future Work

Further investigation is needed to delve deeper into the relationships identified in this study and to understand why they influence outcomes. For example, while the awarding agency of the first SBIR/STTR grant, the total number of awards received, time since founding, and financial structure were highlighted as important contributors to success, the underlying mechanisms driving these relationships remain unclear. The role of financial structure warrants further exploration to determine whether it acts as a stabilizing factor during the transition from grant funding to revenue generation or if it reflects inefficiencies that hinder scaling and long-term sustainability.

Future research should also address the limitations of the dataset by incorporating richer and more diverse datasets. A larger dataset would also allow for the exploration of more complex predictive models, such as neural networks, which require substantial amounts of data to perform effectively. Neural networks could capture intricate patterns and nonlinear relationships that simpler models might miss, potentially enhancing predictive accuracy.

Another area of exploration should involve the development of an end-to-end pipeline to continuously update and maintain the dual-use startup dataset that was created for this research, ensuring it remains a valuable resource for future studies on dual-use startups. Such a pipeline could leverage API integration to automate data extraction directly from key sources such as sbir.gov, Pitchbook, and usaspending.gov. An Extract, Transform, Load (ETL) process could standardize and clean the data, handling discrepancies in formats and ensuring consistency across datasets. To address issues of incomplete or inconsistent data, imputation techniques, such as regression-based methods or K-Nearest Neighbors, could fill in missing values while maintaining data reliability. For instances where automated processes fall short, manual validation interfaces should be included, enabling researchers to review and correct flagged anomalies or mismatches efficiently.

End-to-end workflow automation would tie these components together, creating a seamless

process to collect, clean, and integrate data into a centralized repository. Tools such as Apache NiFi or cloud-based platforms like AWS Glue or Google Cloud could enable scalable and automated workflows. By implementing this system, the dataset would not only remain current as new dual-use startups emerge but also become a robust foundation for longitudinal studies and deeper analyses. This research has established a critical dataset for dual-use startups, providing a basis for future investigations into their unique characteristics and success factors. The proposed pipeline will ensure that this dataset grows in value, supporting ongoing and future research into this dynamic and impactful sector.

Appendix A

Deal Definitions

Below is an explanation of the deal types that were included in the dataset for this research[25].

- **Accelerator/Incubator**: Participation in a program offering funding, office space, or mentorship, often in exchange for equity in the company.
- **Angel (individual)**: Capital provided by a high net-worth individual in the early stages of a company in exchange for a minority stake.
- **Bankruptcy: Admin/Reorg**: Bankruptcy proceedings involving the restructuring of debt under judicial oversight.
- **Bankruptcy: Liquidation**: A bankruptcy process in which a company's assets are sold to pay off debt, resulting in complete cessation of operations.
- **Buyout/LBO**: The purchase of a controlling interest in a company, often financed with borrowed money.
- **Capital Spending**: Financing for long-term investments in fixed assets like property or equipment.
- **Capitalization**: The use of founders' or management's own money to provide financial backing, often referred to as bootstrapping.
- **Convertible Debt**: Debt that converts into equity at a later stage, avoiding the need for new equity issuance during fundraising.
- **Corporate**: When a corporation injects capital into a private company in exchange for newly issued shares.
- **Corporate Asset Purchase**: A corporation acquiring a majority stake in an asset, such as patents or facilities.
- **Corporate Licensing**: Licensing intellectual property to or from a corporation as part of a deal.
- **Debt - Acquisition**: Debt raised to finance corporate acquisitions or mergers.

- **Debt - General**: New debt or loans obtained by a company that do not replace existing loans.
- **Debt - PPP**: Loans provided under the Payment Protection Program (PPP) to help businesses retain employees during the COVID-19 crisis.
- **Debt Refinancing**: New debt issued to replace existing debt with better terms.
- **Debt Repayment**: Funds used to pay off old loans or debt.
- **Debtor-In-Possession**: Loans made to companies under bankruptcy proceedings to support operations while restructuring.
- **Dividend Recapitalization**: New debt raised to pay dividends to private investors or shareholders.
- **Early Stage VC**: A Series A to Series B round that occurred within five years of the company's founding date. If no series is associated, the deal must also happen within five years of the founding date.
- **Equity Crowdfunding**: Financing received through a crowdfunding platform where individuals purchase equity in the target company.
- **Equity For Service**: Equity, warrants, or options provided in exchange for services rendered.
- **General Corporate Purpose**: Debt deals supporting general operations and routine business needs.
- **Grant**: Financing received that does not give the provider an economic interest or right in the company's assets or future cash flows.
- **IPO**: An investment open to public retail investors after meeting the registration requirements for new securities.
- **Investor Buyout by Management**: A management team acquiring ownership of their company from current investors.
- **Joint Venture**: A new entity or asset developed jointly by two or more companies, involving shared assets, funding, or stakes.
- **Later Stage VC**: Includes Series C to Series D rounds or any VC round more than five years after founding. Venture Growth deals are typically Series E+ or older companies with six or more VC deals.
- **Leveraged Recapitalization**: Restructuring using debt, often without changing ownership or distributing dividends.
- **Mezzanine**: A hybrid of debt and equity financing, often involving subordinated debt and a minority equity stake.
- **Merger of Equals**: Two firms of similar size merging to form a single new entity, with all original firms ceasing to exist.

- **Merger/Acquisition**: A corporation acquiring at least a controlling percentage of another corporation's capital stock.
- **Out of Business**: The complete cessation of all business operations.
- **PE Growth/Expansion**: A non-control equity investment by a private equity firm to support growth.
- **PIPE (Private Investment in Public Equity)**: A non-control equity investment by a private investor in a publicly traded company through newly issued securities.
- **Platform Creation**: Initial capital provided by a private equity team for the creation of a new company.
- **Product Crowdfunding**: Funding received from individuals via a crowdfunding platform in exchange for future products, typically before market release.
- **Project Financing**: Debt provided for specific projects, often in sectors like energy, infrastructure, or real estate.
- **Public Investment 2nd Offering**: Issuance of new stock by a company that has already conducted an IPO.
- **Restart - Early VC**: A significant down round in early-stage VC funding that dilutes existing investors.
- **Restart - Later VC**: A significant down round in later-stage VC funding that dilutes existing investors.
- **Reverse Merger**: When a public company is acquired by a private company, enabling the private company to go public more quickly.
- **Sale-Lease Back Facility**: A company selling assets, such as real estate, and leasing them back from the buyer.
- **Secondary Transaction - Open Market**: Stockholders selling shares on a public exchange or private marketplace.
- **Secondary Transaction - Private**: An investment where one investor buys a minority equity interest in a target company from another investor.
- **Seed Round**: Initial financing for a new enterprise in its earliest stages, typically designated as a "seed deal" in sources.
- **Share Repurchase**: Companies buying back their own shares from the market to reduce supply and increase share value.
- **Spin-Off**: A corporate realignment where a business division forms an independent entity.
- **University Spin-Out**: Companies originating from a university to utilize its intellectual property, typically with institutional ownership or licensing agreements.
- **Working Capital**: Debt used to support day-to-day operational expenses.

Appendix B

Variable Definitions

Below is an explanation of the variables that were not explained in Section 3.1.1.

- **Primary Industry Group:** A more granular description of the industry a company operates in.
  - Agriculture
  - Apparel and Accessories
  - Capital Markets/Institutions
  - Chemicals and Gases
  - Commercial Products
  - Commercial Services
  - Commercial Transportation
  - Communications and Networking
  - Computer Hardware
  - Construction (Non-Wood)
  - Consumer Durables
  - Consumer Non-Durables
  - Containers and Packaging
  - Energy Equipment
  - Energy Services
  - Exploration, Production and Refining
  - Forestry
  - Healthcare Devices and Supplies
  - Healthcare Services
  - Healthcare Technology Systems
  - Insurance
  - IT Services

  - Media
  - Metals, Minerals and Mining
  - Other Business Products and Services
  - Other Consumer Products and Services
  - Other Energy
  - Other Financial Services
  - Other Information Technology
  - Other Materials
  - Pharmaceuticals and Biotechnology
  - Restaurants, Hotels and Leisure
  - Retail
  - Semiconductors
  - Services (Non-Financial)
  - Software
  - Textiles
  - Transportation
  - Utilities
- **Primary Industry Code:** An even more granular description of the industry a company operates in.
  - Accessories
  - Accounting, Audit and Tax Services (B2B)
  - Aerospace and Defense
  - Agricultural Chemicals
  - Air
  - Alternative Energy Equipment
  - Animal Husbandry
  - Application Software
  - Application Specific Semiconductors
  - Aquaculture
  - Automation/Workflow Software
  - Automotive
  - Beverages
  - Biotechnology
  - BPO/Outsource Services
  - Building Products
  - Buildings and Property

- Business Equipment and Supplies
- Business/Productivity Software
- Clinics/Outpatient Services
- Clothing
- Coal and Consumable Fuels Equipment
- Commodity Chemicals
- Communication Software
- Computers, Parts and Peripherals
- Connectivity Products
- Construction and Engineering
- Consulting Services (B2B)
- Cultivation
- Database Software
- Decision/Risk Analysis
- Department Stores
- Diagnostic Equipment
- Discovery Tools (Healthcare)
- Distributors (Healthcare)
- Distributors/Wholesale
- Drug Delivery
- Drug Discovery
- Education and Training Services (B2B)
- Educational and Training Services (B2C)
- Educational Software
- Elder and Disabled Care
- Electric Utilities
- Electrical Equipment
- Electronic Components
- Electronic Equipment and Instruments
- Electronics (B2C)
- Energy Exploration
- Energy Infrastructure
- Energy Marketing
- Energy Production
- Energy Storage
- Enterprise Systems (Healthcare)

- Entertainment Software
- Environmental Services (B2B)
- Fiberoptic Equipment
- Financial Software
- Food Products
- Footwear
- Forestry Processing
- Gas Utilities
- General Purpose Semiconductors
- Government
- Holding Companies
- Home Furnishings
- Horticulture
- Hospitals/Inpatient Services
- Household Appliances
- Household Products
- Human Capital Services
- Industrial Chemicals
- Industrial Supplies and Parts
- Information Services (B2C)
- Insurance Brokers
- Internet Service Providers
- Internet Software
- Iron and Steel Mining
- IT Consulting and Outsourcing
- Laboratory Services (Healthcare)
- Legal Services (B2C)
- Leisure Facilities
- Logistics
- Machinery (B2B)
- Managed Care
- Marine
- Media and Information Services (B2B)
- Medical Records Systems
- Medical Supplies
- Metal Containers and Packaging

- Monitoring Equipment
- Movies, Music and Entertainment
- Multi-line Chemicals
- Multimedia and Design Software
- Network Management Software
- Office Electronics
- Oil and Gas Equipment
- Operating Systems Software
- Other Agriculture
- Other Apparel
- Other Business Products and Services
- Other Capital Markets/Institutions
- Other Chemicals and Gases
- Other Commercial Products
- Other Commercial Services
- Other Communications and Networking
- Other Consumer Durables
- Other Consumer Non-Durables
- Other Consumer Products and Services
- Other Containers and Packaging
- Other Devices and Supplies
- Other Energy
- Other Energy Services
- Other Equipment
- Other Financial Services
- Other Hardware
- Other Healthcare Services
- Other Healthcare Technology Systems
- Other Information Technology
- Other IT Services
- Other Materials
- Other Media
- Other Metals, Minerals and Mining
- Other Pharmaceuticals and Biotechnology
- Other Restaurants, Hotels and Leisure
- Other Semiconductors

- Other Services (B2C Non-Financial)
- Other Software
- Other Textiles
- Other Transportation
- Other Utilities
- Outcome Management (Healthcare)
- Paper Containers and Packaging
- Personal Products
- Pharmaceuticals
- Plant Textiles
- Plastic Containers and Packaging
- Practice Management (Healthcare)
- Precious Metals and Minerals Mining
- Printing Services (B2B)
- Production (Semiconductors)
- Publishing
- Raw Materials (Non-Wood)
- Real Estate Services (B2C)
- Recreational Goods
- Road
- Security Services (B2B)
- Social Content
- Social/Platform Software
- Software Development
- Specialized Finance
- Specialty Chemicals
- Specialty Retail
- Storage (IT)
- Surgical Devices
- Synthetic Textiles
- Systems and Information
- Telecommunications
- Therapeutic Devices
- Vertical Market Software
- Water Utilities
- Wireless Communications

  - Wireless Service Providers
  - Wood/Hard Products
- **Ownership Status:** Indicates the ownership under which a company is currently operating.
  - Acquired/Merged: A private company that was acquired by or has merged with another company and currently no longer operates as an independent business.
  - Acquired/Merged (Operating Subsidiary): A private company that is currently doing business as an operating subsidiary, after being acquired by or merged with another company. Public companies that are operating subsidiaries are not included in this category.
  - In IPO Registration: A company that has registered for an Initial Public Offering (IPO) with a recognized stock exchange, but that has not yet become public.
  - Out of Business: A company that is out of business and has ceased operations.
  - Privately Held (backing): A company that is currently privately owned or that is financially backed by an investor, such as a private equity or venture capital investor. This category does not include companies that were acquired by or merged with another company.
  - Privately Held (no backing): A company that currently has no financial backing.
  - Publicly Held: A company that is currently publicly listed. This category includes companies that have gone through a merger or acquisition but are still publicly listed.
- **Universe:** Group companies based on current or past financing statuses of the company. A company will be part of multiple universes if they have received backing from a variety of investor types throughout their lifecycle.
  - Debt Financed: Any company that has, at some point, received debt financing.
  - M&A: Any company that has been acquired by or received financing from a corporation.
  - Other Private Companies: All companies that are not in any other universe due to financing statuses including No Longer Backed, Pending Transaction, or Failed Transaction.
  - Pre-venture Companies: Any company that has received pre-venture capital funding from sources including angel investors, accelerators, incubators, or equity crowdfunding.
  - Private Equity Companies: Any company that has, at some point, been part of the portfolio of a private equity firm.
  - Publicly Listed: Any company that has been or is publicly traded.

  - Venture Capital Companies: Any company that has, at some point, been part of the portfolio of a venture capital firm.
- **Last Financing Deal Type:** A classification for the most recent deal a company has received.
  - Accelerator/Incubator
  - Angel (individual)
  - Bankruptcy: Admin/Reorg
  - Bankruptcy: Liquidation
  - Buyout/LBO
  - Capitalization
  - Convertible Debt
  - Corporate
  - Debt – General
  - Debt – PPP
  - Debt Refinancing
  - Debt Repayment
  - Dividend Recapitalization
  - Early Stage VC
  - Equity Crowdfunding
  - General Corporate Purpose
  - Grant
  - IPO
  - Joint Venture
  - Later Stage VC
  - Leveraged Recapitalization
  - Merger of Equals
  - Merger/Acquisition
  - Mezzanine
  - Out of Business
  - PE Growth/Expansion
  - PIPE
  - Product Crowdfunding
  - Public Investment 2nd Offering
  - Reverse Merger
  - Secondary Transaction – Open Market
  - Secondary Transaction – Private

  - Seed Round
  - Spin-Off
  - Working Capital
- **Last Financing Deal Class:** A broader classification of deal type for the most recent deal a company has received.
  - Bankruptcy
  - Corporate
  - Debt
  - Hedge Fund
  - Individual
  - Other
  - Out of Business
  - Private Equity
  - Public Investment
  - Venture Capital
- **Deal 1 Deal Date:** The date in which the company received its most recent deal.
- **Award 1 HUBZone Owned:** An indicator for if a business participated in the Historically Underutilized Business Zone (HUBZone) program during its first SBIR/STTR.
- **Award 1 Phase:** The phase of the first SBIR/STTR award.
  - Phase I
  - Phase II
- **Award 1 Program:** The program of the first award a company received.
  - SBIR
  - STTR
- **Award 1 Socially and Economically Disadvantaged:** An indicator for if a business was at least 51% owned by one or more socially and economically disadvantaged people, specifically during the first SBIR/STTR the company received.
- **Award 1 Solicitation Year:** The year of the solicitation for the first SBIR/STTR of a company.
- **Award 1 State:** The state that a company was operating when it received its first SBIR/STTR.
- **Award 1 Women Owned:** An indicator for if a company was at least 51% owned by women at the time of its first SBIR/STTR.
- **# of SBIR Phase I:** The number of Phase I SBIRs a company has received, until 2023.
- **# of SBIR Phase II:** The number of Phase II SBIRs a company has received, until

2023.

- **# of STTRs:** The number of STTRs a company has received, until 2023.
- **# of STTR Phase I:** The number of Phase I STTRs a company has received, until 2023.
- **# of STTR Phase II:** The number of Phase II STTRs a company has received, until 2023.
- **Award 1 Award Amount:** The dollar amount a company received for its first SBIR or STTR.
- **Award 1 Award Year:** The year a company received its first SBIR or STTR.

References

[1] C. Roulo. "What on Earth is the Global Positioning System?" U.S. Department of Defense. (Dec. 2018), URL: https://defense.gov/News/Feature-Stories/story/Article/1674004/what-on-earth-is-the-global-positioning-system/#:~:text=GPS%20was%20developed%20by%20the,in%20the%20world%20to%20use.%5C&text=The%20Gulf%20War%20was%20the,military%20used%20GPS%20in%20combat (visited on 11/28/2024).

[2] "University research key." (Oct. 2014), URL: https://www1.udel.edu/udaily/2015/oct/university-research-100814.html (visited on 11/29/2024).

[3] "A Brief History of the Internet." (n.d.), URL: https://www.usg.edu/galileo/skills/unit07/internet07_02.phtml (visited on 11/29/2024).

[4] P. Budden and F. Murray, "MIT's Stakeholder Framework for Building & Accelerating Innovation Ecosystems," MIT Lab for Innovation Science and Policy Working Paper, Tech. Rep., 2019.

[5] "About SBIR and STTR." (n.d.), URL: https://www.sbir.gov/about#:~:text=2014%20Appendices-,Congressional%20History,research%20and%20development%20(R%26D) (visited on 11/29/2024).

[6] "Qualcomm Inducted Into SBIR Hall of Fame." (Jun. 2022), URL: https://www.sbir.gov/success/qualcomm-inducted-sbir-hall-fame (visited on 11/29/2024).

[7] "SBIR-STTR Success: 23andMe." (Jun. 2022), URL: https://www.sbir.gov/success/sbir-sttr-success-23andme (visited on 11/29/2024).

[8] "Understanding SBIR and STTR." (n.d.), URL: https://seed.nih.gov/small-business-funding/small-business-program-basics/understanding-sbir-sttr (visited on 11/29/2024).

[9] F. Murray, "The role of academic inventors in entrepreneurial firms: Sharing the laboratory life," *Research Policy*, vol. 33, no. 4, pp. 643–659, 2004. DOI: 10.1016/j.respol.2004.01.013. URL: https://www.sciencedirect.com/science/article/pii/S0048733304000198x.

[10] L. Frølund, F. Murray, and M. Riedel, "Developing Successful Strategic Partnerships With Universities," *MIT Sloan Management Review*, 2017. URL: https://sloanreview.mit.edu/article/developing-successful-strategic-partnerships-with-universities/ (visited on 12/28/2024).

[11] "SBIR/STTR Program Overview." (n.d.), URL: https://www.defensesbirsttr.mil/SBIR-STTR/SBIR-STTR-Demo/#Guide (visited on 12/27/2024).

[12] J. M. Landreth, "Through DoD' s Valley of Death—A Data-Intensive Startup's Journey," *Defense Acquisition Magazine*, Feb. 2022. URL: https://www.dau.edu/library/damag/january-february2022/valley-death (visited on 12/27/2024).

[13] J. Guzman and S. Stern, "Nowcasting and Placecasting Entrepreneurial Quality and Performance," National Bureau of Economic Research, Working Paper 20954, 2015. DOI: 10.3386/w20954. URL: https://www.nber.org/papers/w20954 (visited on 12/27/2024).

[14] D. B. Audretsch, A. N. Link, and J. T. Scott, "Public/private technology partnerships: Evaluating SBIR-supported research," *Research Policy*, vol. 31, no. 1, pp. 145–158, 2002. DOI: 10.1016/S0048-7333(00)00158-X. URL: https://www.sciencedirect.com/science/article/pii/S004873330000158X (visited on 12/27/2024).

[15] A. N. Link, C. A. Swann, and M. van Hasselt, "An assessment of the US Small Business Innovation Research (SBIR) program: A study of project failure," *Science and Public Policy*, vol. 49, no. 6, pp. 972–978, 2022, ISSN: 0302-3427. DOI: 10.1093/scipol/scac049. URL: https://doi.org/10.1093/scipol/scac049.

[16] A. N. Link and J. T. Scott, "Government as entrepreneur: Evaluating the commercialization success of SBIR projects," *Research Policy*, vol. 39, no. 5, pp. 589–601, 2010, ISSN: 0048-7333. DOI: 10.1016/j.respol.2010.02.006. URL: https://www.sciencedirect.com/science/article/pii/S0048733310000545.

[17] C. Ünal and I. Ceasu, "A Machine Learning Approach Towards Startup Success Prediction," Humboldt-Universität zu Berlin, International Research Training Group 1792 "High Dimensional Nonstationary Time Series", IRTG 1792 Discussion Paper 2019-022, 2019. URL: https://hdl.handle.net/10419/230798.

[18] USA Spending, *Custom Award Data*, Accessed: 2024-12-29, 2024. URL: https://www.usaspending.gov/download_center/custom_award_data.

[19] RapidFuzz, *RapidFuzz: Fuzzy String Matching in Python*, Accessed: 2024-12-29, 2024. URL: https://rapidfuzz.github.io/RapidFuzz/.

[20] U.S. Small Business Administration, *Participating Federal Agencies*, Accessed: 2024-12-29, 2024. URL: https://www.sbir.gov/participating-agencies.

[21] *How to Apply*, https://www.sbir.gov/apply, Accessed: 2024-01-01, 2024.

[22] C. Edwards. "SSTI Analysis Reveals SBIR "Mills" Take Outsized Portion of Program' s Awards." Blog post, accessed: 2025-01-01. (2020), URL: https://ssti.org/blog/ssti-analysis-reveals-sbir-%E2%80%9Cmills%E2%80%9D-take-outsized-portion-program%E2%80%99s-awards.

[23] U.S. Census Bureau, *State Population Totals and Components of Change: 2020-2024*, Accessed: January 1, 2025, 2024. URL: https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html.

[24] Pennsylvania State University, *10.7 - Detecting Multicollinearity Using Variance Inflation Factors*, Accessed: 2025-01-02, 2018. URL: https://online.stat.psu.edu/stat462/node/180/.

[25] PitchBook, *List of deal types*, Accessed: 2025-01-15, 2024. URL: https://help.pitchbook.com/s/article/List-of-Deal-Types.