Forecasting Breakthroughs:

Identifying Future Leaders in the Semiconductor Industry

A Technical Report Submitted to the Department of Systems Engineering

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia

Lauren Sullivan

Spring, 2024

Technical Project Team Members

Robert Brozey

Carter Dibsie

Ethan Kuzneski

Adam Rogers

David Underwood

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISOR

Michael D. Porter, Associate Professor, Department of Systems Engineering

Forecasting Breakthroughs: Identifying Future Leaders in the Semiconductor Industry

Adam Rogers
Systems and Information Engineering
University of Virginia
aer2gz@virginia.edu

Carter Dibsie
Systems and Information Engineering
University of Virginia
crd3uc@virginia.edu

Ethan Kuzneski
Systems and Information Engineering
University of Virginia
ejk9pk@virginia.edu

David Underwood
Systems and Information Engineering
University of Virginia
dau4bm@virginia.edu

Robert Brozey
Systems and Information Engineering
University of Virginia
rrb3zp@virginia.edu

Lauren Sullivan
Systems and Information Engineering
University of Virginia
les2uuw@virginia.edu

Donald Robinson
Titus Technology LLC
President

Michael D. Porter
Systems and Information Engineering
Data Science
University of Virginia

***Abstract*—Breakthrough technologies have the potential to disrupt markets and society. Anticipating such disruptions is crucial for policymakers, investors, and businesses in being proactive with regard to regulatory policies and in allocating resources effectively. This project aims to develop an analytical approach to identify companies that will lead in developing breakthrough technologies. The analysis focuses on the semiconductor industry, which has seen rapid growth in recent decades, surging from $139 billion in revenue in 2001 to $573.5 billion in 2022. Our systematic approach to predicting technological disruption in the semiconductor industry involves leveraging a combination of quantitative company data, human-centric elements, and feature engineering. Data was collected on 244 private semiconductor companies between 2012 and 2018, encompassing information about leadership profiles, research endeavors, media exposure, and financial performance. Two models were developed: a penalized regression model, and a boosted tree model, both aimed at forecasting the probability of a company achieving a valuation exceeding $500 million within five years of its first funding deal. Key variables such as the number of employees, year founded, total invested equity, number of active patents, and country of origin emerged as significant predictors of company success. This paper discusses the performance of our models and explores applying our findings to identify disruptive companies across industries.**

I. Introduction

A. Background

Technological disruptions are the events in which new, more efficient, or more cost-effective technological solutions alter the existing market landscape. They come in to replace established products, services, and business models, reshaping industries and businesses in the process. Technological disruptions have substantial impacts on society. They can exacerbate economic inequality, raise ethical and regulatory challenges, and even prevent accessibility to certain populations. They require thoughtful management and policies to ensure the benefits are widely distributed and negative consequences are mitigated.

The ability to predict technological disruptions is extremely profitable. Investors who can identify potential disruptors gain market share early on and capitalize on their skyrocketing revenues. However, there are several other benefits to predicting technological disruptions beyond this. First and foremost, it allows organizations and policymakers to mitigate risks associated with the displacement of existing technologies, industries, and jobs. Additionally, predicting technological disruptions can develop customer-centric approaches. Businesses can better understand the evolving needs of customers based on the market trends they form. Finally, educational institutions can adjust their curricula to better prepare students with the skills and knowledge required for the disrupted market.

B. The Semiconductor Industry

Due to the complexity and expansivity of technological disruptions, the scope of our analysis was narrowed down to focus on the semiconductor industry. This allows for a rigorous and thorough analysis of a ubiquitous industry that only continues to grow in profitability. The industry is largely fueled by the growing demand for chips in emerging technologies like 5G, artificial intelligence, and electric vehicles. Figure 1, a visual on the overall growth projections in the global semiconductor market from McKinsey & Company, is indicative of this. In the visual, the left vertical bar shows the market value of each sector in the year 2021 and the middle bar shows the projected market value of each sector

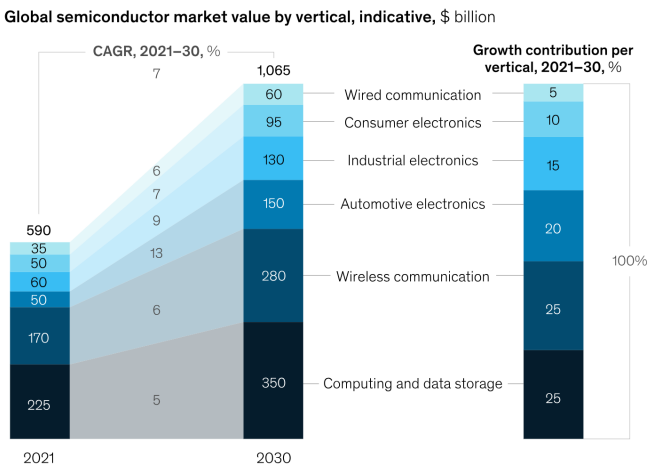

Fig. 1: Global Semiconductor Market Value

in the year 2030. These bars are primarily dominated by wireless communications and computing and data storage. The numbers on the top of the bars indicate the value of the entire industry at the time. McKinsey & Company estimated the global semiconductor market to be worth $590 billion in 2021 and projects it to be "a $1 trillion industry by [2030], assuming average price increases of about 2% a year and a return to balanced supply and demand after current volatility" [1]. In between those two bars is the compound annual growth rate (CAGR) for respective sectors and the overall CAGR of 7% for the entire semiconductor industry at the top. Industrial electronics and automotive electronics are poised to grow the most significantly year over year for the next decade. The bar on the right shows how each sector contributes to the growth of the semiconductor industry as a whole, with wireless communications and computing and data storage accounting for half of the overall growth over the decade. Hence, the semiconductor industry emerges as our prime focus, driven by its escalating significance and promising avenues for prosperity.

C. Objectives

The project's main objective is to construct a binary classification model that forecasts whether a company will succeed or fail within five years following its first recorded funding round. For this study, a company's success was characterized by attaining a valuation of over $500 million within the specified five-year period after the first recorded funding round. The term "unicorn" typically refers to startups valued at a billion dollars, but this benchmark has been adjusted to $500 million to broaden the model's applicability, acknowledging that valuations at this level also represent substantial investment opportunities. Conversely, a company is deemed to have failed if it does not achieve a $500 million valuation within the five years following the initial funding round. The models were developed with a focus on leveraging metrics that are most indicative of success according to this criterion.

II. Methodology

A. Data Description

PitchBook, a financial data provider with information on global mergers and acquisitions, private equity, venture capital, and other financial markets, was used to obtain data for analysis. Specifically, we pulled from the database any deal that included "semiconductors" as an industry variable and received funding in the past eleven years. The eleven-year time frame is enough time to be able to see the five-year outcome of more recent companies, but not too far where the "dot-com boom" or the 2008 financial crisis will skew the data. Deals were collected on semiconductor companies in design, manufacturing, supply chain, etc. to ensure robustness.

The raw dataset contained ninety-two different variables with the identifier variable being Deal ID. From there, the entry had information on the company involved in the deal (Description, Primary Industry Group, Current Financing Status, CEO Name, etc.), details of the deal itself (Deal Date, Deal Size, Deal Type, Pre-Money Valuation, etc.), specifics of stakeholders involved in the deal (Investors, Lenders, Sellers, Beneficiaries, etc.), and information on the financial state of the company receiving the deal (Revenue, Gross Profit, Net Income, Total Debt, etc.). The analysis categorized companies based on a binary outcome variable that denoted success (defined as reaching a $500 million valuation) or failure, including only those that had secured at least two funding deals, with the first post-deal valuation below $500 million and dated before 2018, ensuring a complete five-year outcome data for analysis. Derived from PitchBook and enhanced through network and feature selection techniques, the dataset encompassed a variety of variables, from funding details to company-specific information like CEO, location, and industry. Data was divided into training and testing sets, with all transactions from January 1st, 2012 to June 30th, 2017 for training and those from July 1st, 2017 to December 31st, 2018 for testing, maintaining a training-to-testing split of approximately 2/3 to 1/3. This resulted in 193 companies in the training set, of which 15.5% were deemed successful, and 51 in the testing set, with a success rate of 31.4%.

B. Building a Network of Stakeholders

Stakeholders wield significant influence over technological disruption through their financial resources, strategic decisions, and other supportive mechanisms. By actively participating in respective markets, they can shape the future of newly developed industries. Therefore, CEOs, investors, sellers, and companies' country/territory were analyzed in an effort to build a network of humans and corporations that stand at the forefront of technological disruption. Identifying these networks of successful stakeholders and monitoring their capital deployment should prove lucrative in theory.

Unfortunately, CEOs very rarely appeared more than once in any sort of semiconductor deal included in the dataset. Only three CEOs headed more than one company and the average

Country or Territory	Mean Post Valuation ($ Million)	Median Post Valuation ($ Million)	Ratio of Deals Exceeding $500 Million	Number of Deals in Dataset
United States	65.89	37.50	0.10	110
China	101.38	51.78	0.37	70
United Kingdom	28.78	6.19	0.10	10
Israel	55.89	34.04	0.33	9

Fig. 2: Post Valuation Analysis Based on Country/Territory

post valuation across their companies was significantly lower than that of companies headed by a CEO only appearing once in the dataset. The number of times that a CEO appears in a dataset and the actual name of the CEO proved to be inadequate for our network analysis because of the heterogeneity associated with the variable. Asset management firms like BlackRock, Vanguard Group, and Fidelity Investments appeared several times in both the investors and sellers sections of the dataset. However, there were no investors or sellers that drastically outperformed the pool. In fact, most groups that had invested or sold in companies receiving the greatest post valuations only invested or sold in that one company, and there was no deal that several of the top asset management names took part in. This pattern persisted outside of the semiconductor industry as well, across other industries including financial services, energy, consumer products, and even blockchain. This goes to show, that "even though investors generally treat technological disruption as a non-traditional risk, some leading investors do use organized approaches for handling it. These approaches differ substantially across investors" [2]. Very rarely is an algorithm similar across different stakeholders and these are methods used to encourage them to take their own actions on semiconductor deals. Companies' country/territory held the greatest differences in post valuations deeming them the most valuable in terms of building a network. Deals for companies headquartered in China were, by far, the most likely to achieve a post valuation exceeding $500 million, along with an exceptionally high average post valuation compared to other companies' country/territory headquarters. Figure 2 shows the average post valuation, median post valuation, ratio of companies exceeding a post valuation of $500 million, and number of occurrences of deals in each of their respective country/territory. Deals of the dataset are largely dominated by the United States and China.

C. Using Feature Engineering to Transform Raw Data

The model features obtained from PitchBook make up most of the attributes of the models. However, there were also other features both derived from PitchBook data and retrieved from external sources. One way in which data was derived from PitchBook was by using the CEO Education attribute which was unstructured text listing all of the education levels the CEO of the company has completed. The raw unstructured text was parsed into five columns consisting of Ph.D., MBA, Masters, BS, and BA in order to see if the level of education a CEO had showed any relation to the "success" of the company [3].

Data was also retrieved externally from OpenAlex which is an open-source research database that contains over 240 million scholarly documents like journal articles, books, datasets, and theses. OpenAlex collects these works from a variety of sources including Crossref, PubMed, institutional and discipline-specific repositories.

With the goal of being able to capture engagement surrounding these semiconductor companies prior to their initial funding round, the OpenAlexAPI was used to count the number of scholarly documents related to each company and their CEO from up to three years out to the day of their funding round. The search function used to query across OpenAlex's works looked for matching text across titles, abstracts, and full text. Both the counts for each year prior to the initial funding round and the percent increase/decrease between years were retrieved and turned into new attributes. The counts of the CEOs and companies appearing in works would help account for the weight/magnitude of publicity the company was getting prior to the funding round and the percent change between years captures trends within said publicity. This resulted in the creation of 10 new attributes: one attribute that counts the number of scholarly documents related to the company's name between 3 and 2 years before the initial funding round, another for the count between 2 and 1 years, and a third for the count between 1 year and the day of the initial funding round. Additionally, there are percentage change attributes that calculate the percentage increase or decrease between the 3-2 year count and the 2-1 year count, as well as between the 2-1 year count and the count from year 1 to the day of the initial funding round. This same process is repeated for the name of the CEO of that company [4].

D. Building Models to Predict Disruptions

The goal of modeling is to predict which semiconductor companies will reach a valuation of at least $500 million within five years from their initial funding deal. Lasso logistic regression was selected as one of the modeling methods, and it is particularly suitable for this scenario. The dataset features a significant number of variables, among which some may be irrelevant or redundant when predicting a company's success. Lasso logistic regression addresses this by integrating feature selection into the model fitting process, applying a penalty to the coefficients' absolute values. This penalty can reduce some coefficients to zero, effectively removing those variables from the model. This process helps to prevent overfitting, thereby simplifying the model and making it more interpretable [5]. Models were also built using XGBoost, a popular gradient-boosting machine-learning algorithm. XGBoost was used to help better capture the nonlinear relationships and interactions between the variables in the dataset [6]. Due to the large number of variables in the data, feature selection was done by utilizing feature importance scores and selecting the

variables with a non-zero score. Bayes optimization was used to optimize tuning parameters such as max depth, number of estimators, and the learning rate [7].

III. Results and Analysis

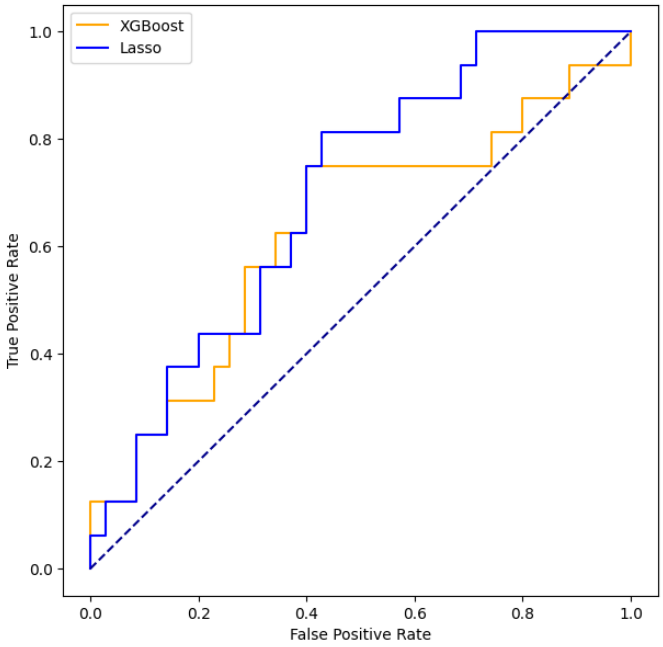

Fig. 3: ROC Curve Comparison

The AUC score was used as a metric for evaluating the models. Since every model was developed on the same training and testing sets, direct comparisons in performance were made. The Area Under the Curve (AUC) score is a metric used to evaluate the performance of binary classification models, derived from the Receiver Operating Characteristic (ROC) curve. The ROC curve plots the true positive rate against the false positive rate at various threshold settings, visually representing the trade-off between correctly identifying positive cases and falsely identifying negative cases as positive. The AUC score, ranging from 0 to 1, quantifies this trade-off by measuring the area under the ROC curve. An AUC score of 1 represents a perfect model that correctly classifies all positive and negative cases. Conversely, an AUC score of 0.5 suggests no discriminative power, equivalent to random guessing. Therefore, the closer the AUC score is to 1, the better the model is at distinguishing between the positive and negative classes across all thresholds. We use the AUC score as a metric because it provides a single number that summarizes the model's ability to rank predictions correctly, regardless of the specific classification threshold. This makes it particularly useful for comparing the performance of different models and for situations where the balance between sensitivity and specificity is important but may vary depending on the context or application.

In Figure 3, the ROC curve graphs show a visual comparison of the models' effectiveness in navigating the trade-off between false negatives and false positives. When evaluated on the testing dataset, the XGBoost model had an AUC score of 62.86% and the logistic regression model had a score of 77.68%.

Additionally, to further analyze the models, a threshold of 0.5 was set to evaluate the accuracy and frequency of type 1 and type 2 errors. It is important to note that depending on one's objective for using the models, the threshold can be adjusted to better meet their needs. A higher threshold is best for those who want to minimize the type 1 error rate, while a lower threshold is better for those looking to take on more risk. Figure 4 shows the confusion matrices for the two models. The Lasso model predicted failure, '0', correctly 34 times, and success, '1', correctly 2 times. However, it also incorrectly predicted success when the true value was failure once, and incorrectly predicted failure when the true value was success 14 times. The 95% confidence interval for the accuracy ranges from 56.17% to 82.51%, suggesting a moderate level of uncertainty in the accuracy estimate, which is a reflection of the small sample size. In terms of the logistic regression model's ability to identify each class, sensitivity (the true negative rate for class '0') is high at 97.14%, showing that the model is very good at identifying actual instances of class '0'. Conversely, specificity (the true positive rate for class '1') is very low at 12.5%, indicating the model struggles to correctly identify actual instances of class '1'. The XGBoost model on the other hand, was slightly better at correctly identifying instances of class '1' with a rate of 25%, while it struggled more than the Lasso model to identify instances of class '0', 91.42% at the threshold of 0.5.

Lasso Logistic Regression		
	Reference	
Prediction	0	1
0	34	14
1	1	2

XGBoost		
	Reference	
Prediction	0	1
0	32	12
1	3	4

Fig. 4: Confusion Matrices

Due to the uncertainty reflected in the models' performance metrics, its utility is primarily derived from the insights gained regarding which variables it identified as predictive. The outcomes from both models highlighted similar variables as significant in predicting the success of companies. These key variables, as shown in Figure 5, include the company's number of employees, founding year, total invested equity, the presence of early-stage venture capitalist deals, the number of active patents, pre-money valuation, and company location. The feature importance was determined by the average gain attributed to each feature during data splitting. This method takes into account both the frequency of a feature's usage in splits and its contribution to the model's predictive power.

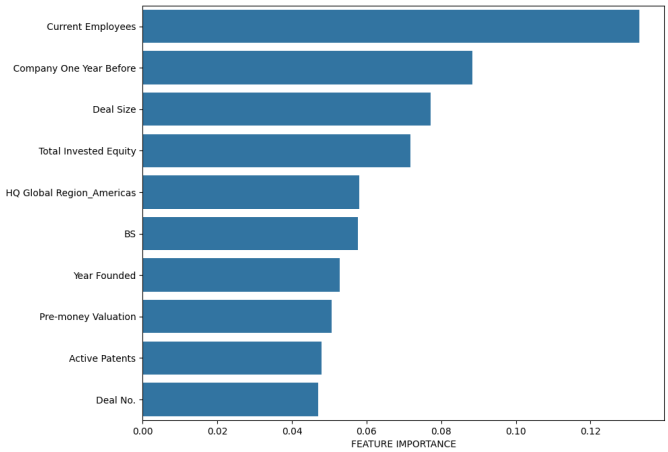

Fig. 5: Feature Importance Plot

The analysis suggests that larger companies, both in terms of workforce and pre-deal valuation, alongside those with substantial equity investments, have a higher probability of success. This outcome aligns with expectations, as larger organizations with considerable financial backing and market valuation are generally better positioned for growth and stability. Additionally, the significance of early-stage venture capital deals in predicting success shows the importance of initial funding and support for startups to kickstart operations and scale effectively. Given that this type of deal is designed to provide the necessary funding for companies to initiate their operations and establish themselves, its prediction of future success is logical. An interesting aspect of the findings is the superior success rate of Chinese companies; approximately 11% of the Chinese companies considered in this study achieved valuations exceeding $500 million, in contrast to just 5% of American companies reaching this valuation.

IV. Discussion

A. Greater Success of Semiconductor Companies in China

As previously mentioned, the semiconductor industry is one of the most globalized industries in the world and therefore one of the most strategically important. "The global model of semiconductor development has resulted in an asymmetric and interdependent relationship between China's critical role in semiconductor production and the United States which controls the key inputs into the value chain" [8]. In fact, 77% of all of the semiconductor deals represented in the study were concentrated in China or the United States. However, only twelve out of the one hundred and nineteen deals based out of the United States reached a post valuation within five years of over $500 million whereas twenty-six out of the seventy deals based out of China achieved this same feat. Furthermore, Chinese companies averaged a post valuation that was 1.53 times greater than the average post valuation of American companies and each of the top four valued post-valuation deals were based out of China. Thus the question is raised - why do semiconductor companies in China appear to have greater success than a peer competitor like the United States?

The Chinese government's desire for economic development likely takes credit. The country has organized itself as a "Socialist Market Economy," where the economy is realistically a "mechanism used by the government to achieve certain socialist goals that can be restricted by it if it fails to achieve them" [9]. With that, work in China is often cheaper due to low labor costs, economies of scale, government policies, infrastructure investments, and regulatory environments. Additionally, China has the largest population in the world making it easier for factories to employ large numbers of workers at low lease costs, charges for land use, electricity fees, permits, and so on. The country is known as "the world's factory." The "Made in China 2025" initiative put on by the Chinese government serves as a good example of this, it targeted unrealistic goals of 40% self-sufficiency of semiconductor manufacturing and production by 2020 and 70% by 2025. The ultimate goal is to "reduce reliance on foreign technologies [by] creating and developing companies that can innovate through research and development, dominate domestically, and produce competitive exports" [10]. Several sectors requiring these semiconductors are displayed in figure 6 along with their respective Made in China 2025 goals for 2020 and 2025. To put this into context, the country was only covering 20.5% of its overall semiconductor consumption by 2023. The plan is shunned internationally because its unrealistic economic goals take precedence over the people. "Local governments, charged with the unfunded mandate of enforcement, are unwilling to implement protective legislation at the sacrifice of economic development. The generation of employment and tax revenue take priority over decent working conditions" [11]. This goes beyond the government of China alone, it has also allowed private and foreign investors to exploit these workers as well. Companies can drastically cut their production costs because of the abundant labor supply, government incentives, and limited worker protections that China offers.

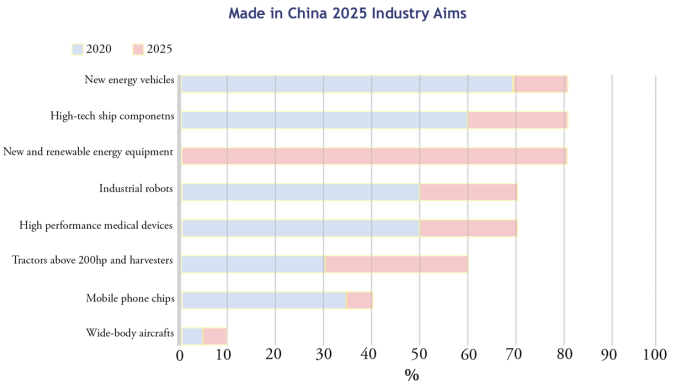

Fig. 6: Targets for Domestic Market Share of Chinese Products

B. Limitations

While the models have shown some success, especially when looking at Chinese companies within the semiconductor industry, there are some limitations to consider when interpreting the results and practical use of this model.

One limitation is since each company used in the training data takes 5 years to mature into either a "success" or a "failure", there will be at least a 5-year gap between the training data and the test data. This could lead to potential drift in the model where there may be important trends in the data that occur within those 5 years that we will be unable to apply to our test set. In the case of the model used for this project, data is only collected from companies receiving their initial round of funding between January 1st, 2012, and December 31st, 2018 (at the latest). This means if there was interest in investing in a company on January 1st, 2023, economic trends from Covid-19 (2020) and the government's 50 billion dollar investment in semiconductors (2022) would not be fully captured within the model.

Also, the initial/early-stage funding round that is used in the data is not necessarily the company's first funding round. It is the earliest round of funding that was available from the data collected from PitchBook between 2012-2018. This means that each company may be further along in the funding stages than others causing some discrepancies there within the data.

Another caveat to note is that investments do not fully capture a company's valuation since the company itself will take investments at arbitrary times. A company's valuation can change drastically over time but the data collected for the model only records the valuation at specific points in time. However, this is currently the closest information that can be obtained about a company's current value due to the lack of investor visibility for most private companies.

The outcome variable (over 500M within 5 years after the funding round) itself is an arbitrary value but has a major impact on model performance. The threshold of a success could have been shifted from 500M to another value and the "within 5 years" time constraint would have also been adjusted but these were chosen with thoughtful deliberation. Also instead of having the response variable measuring if a company is valued over 500M, a variable that measures the percent increase in valuation could have been used instead. However, that metric may not accurately identify market-disrupting companies because a substantial valuation increase could occur for companies with relatively small overall values. That being said, one could make an argument to have used some variation of a different response variable which would likely have resulted in different results.

C. Implications and Further Research

The implications of the models and the research conducted on predicting technological breakthroughs in the semiconductor industry extend beyond the immediate findings. This study highlights the complex interplay of various factors, including company characteristics, and funding dynamics, in shaping the success trajectories of semiconductor companies. The analysis identified certain variables (like number of employees, founding year, and total invested equity) as predictive of success in the semiconductor industry. This suggests that these factors can serve as important considerations when evaluating potential investments.

For entrepreneurs and business leaders, understanding the factors that contribute to a semiconductor company's success can guide strategic planning, operational improvements, and innovation efforts. Emphasizing patent development, securing early-stage venture capital, and leveraging geographical advantages could be beneficial considerations for emerging companies in the sector.

Additionally, the observed differences in success rates between semiconductor companies in different countries may reflect varying national policies, investment in R&D, and support for innovation. It's reasonable to infer that strategic government actions can influence industry success, but the specific policies and their impacts would require detailed policy analysis to fully understand their effectiveness.

While this study provides valuable insights into the semiconductor industry, similar methodologies could be used to uncover industry-specific success factors in other rapidly evolving sectors such as biotechnology, renewable energy, and artificial intelligence.

REFERENCES

[1] Burkacky, Ondrej, et al. "The Semiconductor Decade: A Trillion-Dollar Industry." *Semiconductors*, McKinsey & Company, 1 Apr. 2022.

[2] Barker, A., Monk, A., & Rook, D. (February 28, 2024). Technological Disruption and Long-Term Investors: Managing Risk and Opportunities.

[3] PitchBook. (2023). Data. Retrieved from https://pitchbook.com/data

[4] Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. ArXiv.

[5] Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical Learning With Sparsity: The Lasso and Generalizations. Boca Raton: CRC Press LLC.

[6] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). Association for Computing Machinery.

[7] Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. In Advances in Neural Information Processing Systems 25 (pp. 2951-2959). Neural Information Processing Systems Foundation.

[8] Grimes, S., & Du, D. (2022). China's emerging role in the global semiconductor value chain, Telecommunications Policy, Volume 46, Issue 2, 101959, ISSN 0308-5961.

[9] Ding, X. (April 2009). The Socialist Market Economy: China and the World. Guilford Press Periodicals, Volume 73, Issue 2.

[10] Backgrounder. "Made in China 2025." *Institute for Security & Development Policy*, June 2018.

[11] Josephs, H. K. (2009). Measuring progress under china's labor law: goals, processes, outcomes. *Comparative Labor Law & Policy Journal, 30(2),* 373-394.