FOUNDERS AND FINANCIALS

MACHINE LEARNING ALGORITHMS IN VENTURE CAPITAL

VIKTOR LÅDÖ NAESS (25498)
EMRIK STÅL (25490)

Bachelor Thesis
Stockholm School of Economics
2023

Founders And Financials: Machine Learning Algorithms In Venture Capital

Abstract

The increased usage of machine learning (ML) algorithms in venture capital investment screening poses the question of how different characteristics influence predictions. The purpose of this study is to investigate how founder and financial characteristics influence ML predictions of the raising of more than one round for Swedish startups. A tuned random forest and logistic regression (logit model) are implemented on data for founder and financial characteristics for Swedish startups that have raised a minimum of one funding round. The target variable used for prediction is whether organizations have raised more than one round. SHAP values are used in an analysis of how different characteristics contribute to the ML predictions. For the random forest model, implemented on data from Sweden Tech Ecosystem and Serrano, financial characteristics impact the ML prediction more than founder characteristics. Further, a high share of female founders, a high distance from Stockholm and a lack of prior founder experience negatively contribute to the ML prediction. The importance of financials for the predictions is related to literature on founder replacement over time. The results for the founder characteristics are related to literature on geographical clustering in venture capital, gender bias among venture capitalists and the importance of entrepreneurial experience for future success. A positive contribution of prior founder experience to the prediction of a random forest model, implemented on data retrieved from EQT, is also found. The results inform a discussion on the effects of skewed data on ML predictions. Further, a discussion on how ML algorithms might institutionalize investor biases is conducted. If ML algorithms are not used constructively there is a risk that diversity becomes deprioritized in capital allocation.

Keywords

Sweden, Venture Capital, Founder Characteristics, Entrepreneurial Finance, Machine Learning

Authors

Viktor Lado Naess (25498)
Emrik Stål (25490)

Tutor

Marieke Bos, Deputy Director and Associate Professor, Swedish House of Finance

Examiner

Adrien d'Avernas, Assistant Professor, Department of Finance, SSE

Acknowledgements

We want to express our gratitude for all the feedback and support Marieke Bos has provided us with, Christian Sinding, Alexandra Lutz, Gustav von Sydow, Vilhelm von Ehrenheim, Anton Ask Åström and Linus Frosteryd for your time and valuable insights, and EQT for providing us with data.

Bachelor Thesis
Bachelor Program in Business and Economics
Stockholm School of Economics

I. Introduction

The contribution of venture capital to the growth and development of society spans decades. Through investments in innovative startups, venture capitalists have helped realize groundbreaking businesses such as Uber and Airbnb (Financial Times, 2020). However, behind many successful ventures are a team of talented founders. Gompers, Gornall, Kaplan and Strebulaev (2020) acknowledge the importance of the team in the selection of investments for venture capitalists. Despite the significance of the founders, to the best of our knowledge, we are short of studies that examine the relationship between founder characteristics and startup success in a Swedish context. Success is proxied as whether organizations have raised more than one funding round. The focus of the second round as the success metric, rather than the first, has inherent implications, not least the increased importance of financial characteristics relative to founder characteristics. As a business develops, financial characteristics will likely influence the possibility to raise capital at least equally as much as a strong team. However, the need for funding is unobserved and the focus on second round funding could lead to a selection bias in favor of successful startups.

This paper explores the importance of founder and financial characteristics in the context of Swedish startups and the raising of second round funding. We pose the question: *How do founder and financial characteristics influence machine learning predictions of the raising of more than one funding round for Swedish startups?*

Investigating which founder profiles that raise capital and thereby can build successful startups is relevant. Stockholm is ranked the fifth European capital with regards to the share of venture capital received. According to The Swedish Private Equity & Venture Capital Association Stockholm received on average 4 billion in venture capital between years 2017-2021. Stockholm is prominent in a European context in its attraction of venture capital. Three quarters of all venture capital raised in Sweden is concentrated in Stockholm[1] (Stockholm Chamber of Commerce, 2023).

Swedish venture capital is not only concentrated geographically. Male startup founders also dominate the share of venture capital received. Less than one percent of private venture capital is invested in female founded companies. The Deputy Prime Minister of Sweden and Minister of Energy, Business and Industry Ebba Busch says that improvements of conditions for women to start, build and own businesses are crucial to strengthen the future of the Swedish economy and competitiveness (The Government Offices of Sweden, 2023).[2]

Randomized controlled trials (RCT) would be the ideal method to investigate the impact of founder characteristics on the possibility to raise VC funding. A randomly generated group of startup founders could have been assigned founder characteristics such as gender and location in applications for first round funding. Then potential differences in the success to raise first round VC funding in an experiment could have been compared between this group and a control group not assigned the founder characteristics. We focus on how founder characteristics impact machine learning (ML) success predictions. Thus, conclusions cannot be made directly about decision making among venture capitalists. However, since ML is increasingly used in investment screening our approach still provides valuable insights. RCTs could reduce endogeneity issues associated with unobserved demand for funding and selection bias associated with our focus on second round funding.

---

[1] Omni (2020), Göteborgs Posten (2020) and Dagens Industri (2020) also highlight the dominance of Stockholm in the venture capital received in Sweden.

[2] SVT (2023), Dagens Industri (2022), Di Digital (2021) and Dagens Nyheter (2022) all address the low gender diversity in the Swedish venture capital space.

The paper consists of two separate result and analysis sections. In part one of the paper, five characteristics are investigated. These include the founder characteristics: the distance from Stockholm of the organization and prior founder experience in the team complemented with the financial aspects: net sales to average industry net sales ratio, return on equity (ROE) and the quick ratio. Moreover, the share of female founders in a team is added as a variable to the model at the end of part one. In part two a separate analysis is conducted and three characteristics are analyzed: share of female founders, distance from Stockholm of the startup and prior founder experience in the founding team. The success proxy used in both parts of the paper is whether organizations have raised more than one funding round. That the need for funding is unobserved is a limitation since a team with less capital needs could turn out more successful than a team that raises multiple capital rounds. Certain organizations that only raised one round might only have applied for a first round.

We attempt to answer the research question through a methodology based on a paper by Fuster, Goldsmith-Pinkham, Ramadorai and Walther (2022) focused on the reasons for inequality in household finance. The authors find that white and asian borrowers are favored relative to black and hispanic borrowers concluded by the usage of ML algorithms in the context of credits. Similarly to the authors we use a random forest model. However, the model is implemented on data for founder team, financial characteristics and funding round data. The authors propose the investigation of the effect of ML technologies on other financial markets than credit markets. In our paper a logit and tuned random forest model are trained for two separate datasets and sets of variables. Then the target variables are predicted for the test datasets. The models are then evaluated with the help of 7-fold cross validation with average accuracies computed. Using the superior tuned random forest model for each dataset the importance of each feature in the prediction of the target variable is extracted. It is interesting to investigate which characteristics will influence ML predictions the most. ML and random forest bring more versatility compared to for example logistic regression. Further, ML methods such as random forest scale better as different dimensions are added to the model (Hastie et al. (2009)).

Further, this methodology is combined with an analysis of SHAP values[3] based on Griffin, Hirschey and Kruger (2023). SHAP values contribute with an understanding of the marginal contribution of each feature to the ML predictions. Griffin et al. (2023) investigate the impact of characteristics of dealers on bond markups. SHAP values are used in the identification of the relative importance of features in the explanation of markups.[4] A substantially lower number of features compared to the authors are investigated in our paper due to data availability. Consequently, we acknowledge that this is a limitation of our study. However, similarly to the authors we graph average absolute SHAP values in the results section.

First, in part one the tuned random forest model is superior in terms of accuracy and cross validation train accuracy both with gender excluded and included. The financial metrics of firms display the greatest explanatory power for the predictions of the tuned random forest model across both datasets in part one. This is discussed through the lens of literature by Kaplan, Sensoy and Strömberg (2009) on founder replacement over time. Based on SHAP values, in both random forest models in part one, fully female teams and teams without prior founder experience negatively contribute to the ML prediction of raising more than one round. The SHAP values for

[3] Lloyd Shapley was awarded the Nobel Prize in Economic Sciences in 2012 for the Shapley value concept based on game theory and established in Shapley (1953).

[4] The method in the paper by Griffin et al (2023) builds upon work by Lundberg et al. (2020) on adopting SHAP values for tree based machine learning models.

the share of female founder variable relates to literature by Ewens and Townsend (2020) on the difficulty for women to get investor interest from male investors. The contribution of the prior founder experience variable to the ML prediction is analyzed through, for example, research by Gompers, Kovner, Lerner and Scharfstein (2010) on the importance of prior entrepreneurial experience for future success.

Second, in part two we find that the logit and tuned random forest model have cross validation test accuracies of 65.2% and 68.5% respectively. Both models display low F1 values of 0% for the logit model and 17.24% for the tuned random forest model. The random forest model maintains a high placebo test accuracy, indicative of a model that has not captured interaction. The share of female founders has the highest explanatory weight, followed by distance from Stockholm and the serial founder variable for the random forest model. Based on SHAP values for the random forest model prior founder experience positively contributes to ML predictions of raising more than one round. For gender, mixed teams have positive SHAP values whereas both fully male and fully female teams display negative SHAP values. Although the tuned random forest model in part two is more accurate compared to the logit model neither of the models are optimal due to the low recall levels. The recall levels seem to be influenced by the skewed nature of the data with most observations being classified as not successes. The omission of financial characteristics in part two naturally influences the results and is further addressed.

Gender biases in selection of investments among venture capitalists have been investigated in the literature. Ewens and Townsend (2020) provide empirical evidence that women-led startups face discernible disparities in funding opportunities compared to their male counterparts.[5] It is acknowledged that gender biases might be symmetric, meaning that investors invest in founders similar to themselves. The authors emphasize that the majority of early stage investors are male. Consequently, homophily is raised as a possible explanation for symmetric biases among investors. Currarini, Jackson and Pin (2009) defines homophily as the tendency of different individuals to relate with those they are similar to.

Decision making among venture capitalists has been investigated in prior literature. Gompers, Gornall, Kaplan and Strebulaev (2020) highlight the long time period for closing deals, with on average 118 due diligence hours spent in the sample of their study. Bonelli (2022) mentions the lengthy investment screening process in VC and addresses potential biases in investment selection. Consequently, VCs have started to adopt ML to improve and automate screening processes. For example, EQT Ventures has its own proprietary tool used to automate the screening process.

As a ML model is only as comprehensive as its input data, insights on founder team attributes relative to the raising of funding rounds are relevant. Biases in VC screening might become institutionalized as algorithms are trained on homogenous historic data. A potential illusion of objectivity and data driven decision making might for example lead male investors to continue to invest in those similar to themselves, that is male startup founders.

This paper is structured as follows. Section II presents related literature. Section III provides an institutional background of the Swedish VC industry. Section IV outlines the two datasets used, which include founder characteristics, financial measures of organizations and funding rounds raised. Section V describes the empirical implementation of the paper. Section VI displays the empirical results and analysis for part one and two of the paper. Section VII includes a robustness section. Section VIII concludes.

[5] For classical literature on statistical discrimination see for example Phelps (1972).

II. Related Literature

Firstly, previous literature study decision making among venture capitalists. Kaplan, Sensoy and Strömberg (2009) study 50 VC backed startups up until IPO and find that founder replacement together with going public is common. The results propose investors should focus more on the business compared to the team in investment screening. However, Gompers, Gornall, Kaplan, Ilya and Strebulaev (2020) uses a survey based methodology and mentions the team as most important in selection of investments among venture capitalists. The authors underline the importance of an active approach to generating deals through networks and referrals. Bubna, Das, and Prabhala (2020) highlight how different VCs apart from screening can co-invest in startups to reduce the risks associated with investments. Gompers, Mukharlyamov and Xuan (2016) also address VC syndication and its role for diversification, accumulation and pooling of resources and competences and to decrease information asymmetry relating to portfolio firms. The authors acknowledge that individual venture capitalists are more likely to cooperate with other venture capitalists with whom they are alike, for instance in terms of education and professional experiences, which relates to homophily.[6] Performance of investments is shown to be diminished for individual venture capitalists that are similar to each other.

Moreover, our study relates to articles about human characteristics in the finance literature. Ewens and Townsend (2020) find that female founders have a more difficult time in funding opportunities relative to male founders.[7] Our paper investigates characteristics of the founding team relative to the raising of funding rounds, where gender is one of the characteristics used. Moreover, we use a machine learning methodology unlike Ewens and Townsend. They focus on for example if an investor shares a startup's profile on AngelList as a proxy for interest in the startup.[8] However, we concentrate on whether organizations have raised more than one funding round as a success metric. Kaplan and Sorensen (2021) find certain personal characteristics that distinguish CEOs from others: more extreme levels of execution, general ability, strategic focus and charisma. The extension of our paper on Kaplan and Sorensen's is three-fold, firstly we focus on the attributes of founders of startups, whereas Kaplan and Sorensen view only the CEO which has the whole selection bias of hiring implemented. The second aspect is the set of attributes where our paper focuses on less qualitative characteristics, such as geographic location, prior founder experience and gender.[9] Lastly, we employ a ML method unlike the authors.

In addition, our study relates to literature about machine learning in finance research. Fuster, Goldsmith-Pinkham, Ramadorai and Walther (2022) find that white and asian borrowers are favored relative to black and hispanic borrowers due to the usage of machine learning algorithms in the context of credits. Our paper uses a similar methodology as the authors, however in another context, namely founder and financial characteristics and startup success. Bonelli (2022) showcases that as VCs adopt artificial intelligence (AI), investments are tilted toward startups with

[6] Ye Zhang (2023) also finds homophily, in an experiment where male US venture capitalists donate more money to male startup founders relative to female in a donation game.
[7] Kessler, Low and Sullivan (2019) does not find gender discrimination in the context of employers evaluating hypothetical resumes from Wharton graduates, where the importance of diversity for the employers are used as an explanation by the authors.
[8] Bernstein, Korteweg and Laws (2017) also use AngelList and conduct an experiment where it is found that investors' reactions are the strongest to information about the founding team of startups.
[9] Other characteristics not part of our scope are also relevant, for example Bengtsson and Hsu (2015) point out how shared ethnicity can strongly predict whether investors invest in startups.

similar business to already existing startups. The usage of historic data that are uninformative about innovative firms is raised as an explanation.

Our paper provides a machine learning based analysis of founder and financial characteristics that predict startups' possibilities to raise more than one funding round. Existing literature investigates for instance how certain founder characteristics influence startup attributes and the usage of machine learning in other contexts such as credit markets. However, we combine the analysis of founder attributes and financial measures with the usage of machine learning classification methods for startup success. To the best of our knowledge few such analyses exist at least in a Swedish context. Also our usage of Swedish data is relevant considering the unique position of Sweden in the European venture capital landscape. Lastly, our paper contributes with a feature importance analysis, nuanced by SHAP values based on previous literature, yet in a new context. That is the effect of founder and financial attributes on the raising of more than one round of funding.

III. Institutional Background

Sweden and Stockholm are unique in a European context in the support to startups via financing, according to Fredrik Ekström, CEO of Nasdaq. The Swedish ecosystem is higher on the political agenda compared to for instance the ecosystem in Norway and Denmark. Ekström expresses the high engagement in the community of private investors in Sweden, both directly through the stock exchange and indirectly via funds and the pension system. (Stockholm Chamber of Commerce, 2023).

Sweden represented 52% of the invested capital in the Nordic region during 2022 (PitchBook, 2023). The contribution of venture and growth capital to Swedish GDP is 1.5%, or 82 billion SEK. Swedish firms constitute 37 billion SEK of the invested venture capital in Sweden. Although, there is a high upside potential in VC investments the share of VC investments that generate a loss is 45%. Diversification is used to mitigate these risks in VC. Technology firms are the focus in the Swedish VC context and life science and ICT constitute 70% of Swedish venture and growth capital investments[10] (SVCA, 2022).

The high share of male VC partners (investors) has been raised as a potential explanation for the low share of venture capital invested in female founded startups. However, over the years female founded venture capital firms such as Backing Minds have emerged and are starting to alter the investing landscape (Dagens Industri, 2019). For instance, only 6.94% or 20 of 288 venture capital firms available on Sweden Tech Ecosystem have female partners as of 2023. Moreover, 68.4% or 197 of the 288 VC investors on the website are situated in Stockholm. Furthermore, 8.3% or 24 VC investors are located in Skåne County and 7.29% or 21 investors are located in Västra Götaland County. There are few VC investors in each of the rest of the Swedish counties (Sweden Tech Ecosystem, 2023).

There is a low share of female VC partners and venture capital firms situated outside of three geographical clusters. Consequently, it is interesting to investigate if certain founder profiles are benefitted or not in the machine learning predictions of raising more than one round. This is for example discussed relative to literature on homophily, VC clustering and prior entrepreneurial experience in section VI.

IV. Data

The ideal dataset would include data for founders in a control group and treatment group. This relates to the discussion in the introduction about randomized controlled

[10] Simultaneously these sectors account for only 9% of the total Swedish economy (SVCA, 2022)

trials as the ideal method. Ideally financial metrics, including for example forecasted sales growth and other potential metrics included in pitch decks presented to VCs would be included in the dataset. Considering that future potential is important in the VC context, our usage of historical financial data relative to the second round year in part 1 might be a limitation. Further, in the ideal dataset different founder characteristics would be randomly assigned to organizations in a treatment group in an experiment. Ideally, the experiment would focus on pitches for first round financing. This would ensure all startups included, actively seeked financing. Our data includes the actual founder characteristics of teams and historical financials. This data is used as input in a random forest model. Thereby, our results cannot directly provide evidence of potential biases among VC investors, which an experiment could provide. However, our method provides insights about which founder and financial characteristics influence ML predictions of the raising of more than one round. Again this is relevant considering the increased usage of ML algorithms in the VC space (Bonelli, 2022).

Part 1. Serrano And Sweden Tech Ecosystem Part

Firm characteristics retrieved from the Serrano database have been matched with data scraped from the website Sweden Tech Ecosystem.[11] Location of startups, the number of rounds raised and whether teams have serial founders have been scraped for years 2010-2012. The time period has been chosen to allow enough time for all organizations to raise multiple rounds of funding. Again success has been defined as raising more than one round. Importantly, angel, seed, convertible, early VC, growth equity VC, late VC, media for equity and series A-G rounds have all been treated as funding rounds in the data scraping process.[12] Moreover, for transparency rounds described ambiguously on the Sweden Tech Ecosystem website have been treated as funding rounds. These rounds occasionally include for example smaller rounds financed by incubators and accelerators. Debt, grants, acquisitions and IPOs have not been treated as funding rounds. Implications of for instance the inclusion of early angel and seed rounds are discussed in section VI.

An aim of this paper is to contribute to the literature on decision making among venture capitalists. Although, raising multiple rounds is not a direct measure of success it can still be an indication of interest among venture capital investors. Gompers (1995) highlights how venture capital investors can stage investments and inject more capital in second funding rounds as a business has proven successful. Consequently, raising more than one round can be a success indicator for startups. Simultaneously, the inclusion of rounds raised from non professional VC firms such as angels can be a limitation. For certain firms the second round might thus be the first round raised from a professional VC. Therefore, this might impact the possibility to draw conclusions related to raising multiple rounds as an indicator of success.

We also acknowledge that in practice an organization could raise multiple rounds and still fail.[13] Also a startup might raise only one round of funding and still be relatively successful.[14] Further, since Sweden Tech Ecosystem is a website with the purpose to for instance support VCs in investment screening it is relevant to analyze

[11] Sweden Tech Ecosystem is operated by Tillväxtverket, Svenska Institutet, Vinnova and Business Sweden.
[12] Based on available categories for filtering on the Sweden Tech Ecosystem website.
[13] For example, Jawbone raised multiple rounds of funding from top VCs and still failed (Financial Times, 2017).
[14] Startups might strategically use bootstrapping whereby business growth is financed by internal cash flows rather than external capital (Forbes, 2019).

data from the website. Again in the paper all organizations have raised a minimum of one funding round. However, it could be interesting to also investigate organizations that did not raise any funding round. At the same time retrieving financial information from the Serrano database for very small organizations that have not raised any funding rounds could also be difficult. Also it would be relevant to take into account the size of each funding round. However, because of scarcity of this data especially for earlier rounds on the Sweden Tech Ecosystem website this aspect will not be taken into account. This is a limitation since in reality raising two large rounds differs from raising two smaller rounds. Data on the Sweden Tech Ecosystem platform is collected through the usage of AI and API and public data sources are continuously scanned (Vinnova, 2021). Therefore, there might be a selection bias for the startups included on the Sweden Tech Ecosystem website since algorithms are used to aggregate data. Consequently, all startups founded 2010-2012 that sometime have raised a minimum of one round might not be visible on the website. However, given that a VC investor might use the data on the website for decision making it is still interesting to analyze.

Firstly, filters have been set on the Sweden Tech Ecosystem website to show the funding rounds for firms founded in 2010, 2011 and 2012 respectively. A filter to not show non-tech firms has been set to ensure firms that are primary subjects for venture capital investments are only shown, relating to the institutional background.[15] Data have subsequently been scraped for the organization name, location, founding year and the number of rounds raised.[16] Then the distance in kilometers from Stockholm has been retrieved for the cities, with the help of Google Maps. For organizations where location information has been missing for city level from Sweden Tech Ecosystem this information has been retrieved through Crunchbase. If a firm has raised more than one round this has been numerized as 1, and otherwise the variable has been set to 0. Then, a filter to show organizations with serial founders has been set. The variable serial founder has been set to 1 for the organizations displayed on the Sweden Tech Ecosystem website when the serial founder filter has been present.[17]

Two characteristics of the founding team were initially used in part one. Gompers, Gornall, Kaplan, Ilya and Strebulaev (2020) have shown that VC investors view the team as important for investment decisions. The authors also acknowledge the importance of entrepreneurial experience and even a prior relationship to a VC for founders. Additionally, Gompers, Kovner and Lerner (2010) find evidence of the importance of geographic location relative to VC investments. Consequently, the variables distance from Stockholm and prior founder experience are relevant to include. These founder characteristics variables have been chosen both because of their availability and relevance in relation to prior literature.

---

[15] This initial filtration could induce a selection bias that could impact the share of female founders. Ewens and Townsend (2020) address how female founders are less unfavored with male investors when in female focused industries. This potential selection bias is, however, not discussed further.

[16] The data for the organizations were scraped from the following website during the period retrieved 2-5 November 2023 (Sweden Tech Ecosystem is continuously updated and new firms not visible on the website at the time of the data scraping process might have been added to the website): https://sweden.dealroom.co/transactions.rounds/f/growth_stages/anyof_late%20growth_early%20growth_seed_not_mature/launch_year_max/anyof_2012/launch_year_min/anyof_2010/rounds/not_GRANT_SPAC%20PRIVATE%20PLACEMENT/slug_locations/allof_sweden/tags/not_outside%20tech

[17] Again this data was scraped during the period 2-5 november 2023 https://sweden.dealroom.co/transactions.rounds/f/founders_is_serial_founder/anyof_yes/growth_stages/anyof_late%20growth_early%20growth_seed_not_mature/launch_year_max/anyof_2012/launch_year_min/anyof_2010/rounds/not_GRANT_SPAC%20PRIVATE%20PLACEMENT/slug_locations/allof_sweden/tags/not_outside%20tech

It is also relevant to complement data on the founding team with financial information about each organization. The Serrano database retrieved from the Swedish House of Finance has been used to retrieve information about financial information. The financial characteristics included are a net sales ratio, ROE and quick ratio. For example, this relates to Kaplan, Sensoy and Strömberg (2009) that highlights the importance of the business vis-à-vis the team in investment decisions. These variables have been extracted for the year prior to the second funding round. When the second funding round is the same year as the founding year, data for the founding year is selected. VC investors might base decision making partly on recent historical financial performance of the firms. However, financial forecasts for the future might also affect decision making. For example, Puri and Zarutskie (2012) highlight the importance of the potential for scale for VC backed firms. Concurrently, the year prior to the year of the second round is the closest proxy for financial performance based on the availability in the Serrano database.

When a firm visible on the Sweden Tech Ecosystem website has not been available in Serrano the observation has been removed. The potential selection bias that arises is discussed later. Further, when information about the target variables have been missing for any firm, this observation has also been omitted. It is acknowledged that this is a limitation, yet it is needed to enable machine learning analysis. For example, this follows Fuster et al. (2022) that drop missing values from the dataset in their paper.

After the completion of the data cleaning 210 observations are part of the dataset for years 2010-2012 in part one excluding gender. The small size of our dataset is a further limitation of our study. The data is split into features (x) and target (y) variables as seen below in table 1 and subsequently it is split into a training and test data set. The training set is 70% of the data and the test set is 30% of the data.

Table 1- Variables Part 1

Sweden Tech Ecosystem	**Serrano**
Model specific y: More than one round (1=yes, 0=no) x1: Distance from Stockholm in kilometers (continuous) x2: Prior founder experience (1=yes, 0=no)	x3: Net sales ratio (year before second round $= \frac{Net\ Sales\ (year\ before\ second\ round)}{Average\ Industry\ Net\ Sales\ \ (year\ before\ second\ round)}$ x4: ROE (year before second round) $= \frac{Adjusted\ Net\ profit/loss}{Adjusted\ Equity}$ x5: Quick ratio (year before second round) $= \frac{Total\ Current\ Assets - Inventory}{Total\ Current\ Liabilities}$
Other variables Organization Name Year founded Location	
Linkedin	
Added Part 1 Including Gender *x6: Share of female founders (continuous 0-1)*	

Variables and variable definitions for the logit and random forest model

Table 2- Summary Statistics- Part 1 Excluding Gender

Column N = 210	**Mode**	**Median**	**Mean**	**Q1**	**Q2**	**Q3**	**Q4**	**Stan dev.**
More Than One Round	1	1	0.680952	0	1	1	1	0.467221
Founded on Year	2012	2011	2011.085714	2010	2011	2012	2012	0.806025
Final Second Round Year	2015	2014	2014.538095	2013	2014	2016	2021	2.202864
Distance From Stockholm	0	71	233.315714	0	71	469	929	263.070826
Serial Founder	0	0	0.276190	0	0	1	1	0.448181
Quick Ratio	1.212903	1.1780836	3.954649	0.887594	1.780836	3.994718	45.266272	6.330634
Return on Equity	0	-0.244695	-0.970631	-0.908996	-0.244695	0.032905	1.246581	2.335764
Net Sales	0	658	5736.542857	32.25	658	3767	262777	20892.039069
Net Sales Ratio	0	0.229683	1.170728	0.012689	0.229683	1.066603	21.388328	2.764545

Summary statistics of part 1, merged Sweden Tech and Serrano dataset excluding gender

Summary statistics for the merged Serrano and Sweden Tech Ecosystem Data is displayed in table 2. This is the part one data excluding gender. First, it can be observed that more than half of the organizations in the dataset have raised more than one round. This is reflected by the mean of 0.68 for the more than one round dummy variable. 143 of the 210 observations in the dataset raised more than one round. The high level of success in raising more than one round is likely influenced by the initial conditions for the sample. All organizations on the funding round section of the Sweden Tech Ecosystem website used, have raised a minimum of one round. If organizations that had not raised any round had been taken into account the success level would likely have been lower. Second, only 27.62% of the organizations in the dataset have serial founders. Third, for distance from Stockholm the average distance from Stockholm in the dataset is 233.32 kilometers. The limitation in terms of the few observations per city in the sample is hereby noted (see Appendix A for details). The majority of the organizations in the sample are located in Stockholm, that is at zero distance from Stockholm. The bottom three quartiles of the organizations in the dataset raised their second round any time before 2016. This relates to our somewhat later stage focus of our paper.

Based on the Serrano data it can be discerned that the firms on average have positive sales and liquidity. This is represented by positive mean values for the quick ratio and net sales numbers. However, simultaneously the profitability of the firms in the sample is on average negative, reflected by a negative average for the ROE variable. This is likely a reflection of the difficulties of startups to be profitable in their initial stages (Davila and Foster, 2007). Again the success metric used is whether startups have raised more than one funding round. The full dataset is relatively balanced in terms of the observations per year, with 60, 72 and 78 organizations founded in years 2010, 2011 and 2012 respectively. See Appendix B for branch sector definitions and sales data for the different observations in our dataset in part one excluding gender per sector. It should be noted again that an initial filtration to not show non tech firms was made in the first data scraping of the Sweden Tech Ecosystem. However, this only removed three firms. This is likely since Sweden Tech Ecosystem in itself is a website focused on technology firms.

Gender has then been scraped for the founders of the organizations in the dataset used in part one using Linkedin. The gender for all founders found on Linkedin have been noted. If no founders have been available on Linkedin other websites such as the companies own websites have been used to determine the gender of the founders.

Organizations where the founders have not been found either through a Linkedin search or through other websites such as a company's own website and announcements about the firm, have been dropped. The original dataset in part one has been reduced to 191 observations after gender has been added to the dataset. That is 9% of the observations from part one excluding gender have been dropped. This is a limitation that reduces the reliability of our study. The dataset in part one including gender will be analyzed separately to the dataset in part one excluding gender. The inclusion of gender data relates literature by Ewens and Townsend (2020) and their investigation of male dominance and gender discrimination in the VC space.

Table 3- Summary Statistics- Part 1 Including Gender

Column N = 191	**Mode**	**Median**	**Mean**	**Q1**	**Q2**	**Q3**	**Q4**	**Stan dev.**
More Than One Round	1	1	0.696335	0	1	1	1	0.461048
Founded on Year	2012	2011	2011.094241	2010	2011	2012	2012	0.795724
Final Second Round Year	2015	2014	2014.560209	2013	2014	2016	2021	2.258385
Distance From Stockholm	0	16.7	220.759162	0	16.7	469	929	260694917
Serial Founder	0	0	0.293194	0	0	1	1	0.456423
Quick Ratio	1.212903	1.782296	3.874490	0.909262	1.782296	4.041283	45.266272	5.925131
Return on Equity	0	-0.289300	-0.978976	-0.901756	-0.289300	0.074449	1.246581	2.411250
Net Sales	0	621	5772.267016	32.5	621	4074	262777	21721.599945
Net Sales Ratio	0	0.221997	1.191309	0.012931	0.221997	1.085014	21.288328	2.849457
Share Female Founders	0	0	0.100087	0	0	0	1	0.254868

Summary statistics of part 1, merged Sweden Tech and Serrano dataset including gender

After gender has been scraped for founders for the organizations in the dataset in part one, some observations are dropped. This is because it has been difficult to find the founders for these organizations online. Based on the gender statistics for the new version of the dataset in part one we see that the mean share of female founders is 0.10. Consequently, the data reflects highly male-dominant organizations on average. There are no substantial differences in the mean values of the other different variables in part one including gender compared to excluding gender.

Firstly, data for organization name, location, funding round data and founding year was scraped from Sweden Tech Ecosystem. The data for the 370 organizations founded years 2010-2012 that have raised a minimum of one funding round and were listed on Sweden Tech Ecosystem were initially scraped.[18]

[18] See footnote 15 for the section of the website used. Again the website is updated continuously and more organizations could be visible over time compared to the data scraping period 2-5 November 2023.

Table 4. Filtering Decisions Overview

Filtration Step	Percentage Removed from total set	Percentage Removed from previous set	Cumulative N Left 370
1- Serrano Matching	38.92%	38.92%	226
2- Year and Branch	2.16%	3.54%	218
3- Extreme values removed	2.16%	3.67%	210
4- share of female founders	5.14%	9.05%	191

The table displays the filtration process for the part 1, merged Sweden Tech and Serrano dataset. First, a matching of startups with the Serrano database was conducted and organizations not found in Serrano were removed. Second, data lacking funding year or branch sector were removed. Third, extreme values with a Z-score greater than three standard deviations from the mean were removed. The reasoning behind this is to remove absolute extremes while retaining the variability and oddities that are present in start-up financials. Fourth, organizations without gender data successfully scraped were dropped.

The filtration process for the part one dataset displayed in Table 4 implies that in total 43.24% of the 370 organizations scraped from Sweden Tech Ecosystem have been dropped for the dataset in part one excluding gender. In the dataset in part one including gender 48.38% of the 370 observations have been dropped because of missing data.

Naturally the reduction of the dataset because of missing data has implications for the interpretations of our results. Data might especially be missing for organizations that have not succeeded in raising multiple funding rounds. Consequently, an unintentional selection bias is inferred. This could be part of explaining the majority of organizations categorized as having raised more than one round in the data. Thus, it should be noted that the share of organizations having raised more than one round might have been lower if data availability would have been higher.

Part 2. EQT Part

Firstly, we want to emphasize that part two is separate from part one and the data in part two is different to and has been treated fully separate from that in part one. In part two an aggregated dataset retrieved with the help of EQT is used for further analysis.

Exact details on the underlying data sources and method of data aggregation cannot be provided for part two. This is a limitation since it reduces the replicability of our study. While we do not have access to information about the data aggregation process of the EQT data there is a risk of selection bias. Naturally, any automated aggregation of different data sources might imply a filtration of the dataset. Consequently, the dataset might not be representative of every startup in Sweden during the chosen years. Unlike in part 1 we cannot provide the same degree of detail about the definitions of the different variables. The access to exhaustive founder data underlines the value of the collaboration with EQT. Further, considering that an aim of our study is to contribute to the literature on decision-making among venture capitalists, usage of data from Sweden's largest venture capital firm is relevant to analyze.

Moreover, probability based models are used to determine gender of the founders in the data. This is a further limitation of the data gathering process as it causes uncertainty in the data that is difficult to account for. Conversely, the automated data gathering aspect of the sample data in part two is of likeness to the reality, to such an end that this would be the predominant data approach for VCs. Since, one purpose of our study is to contribute to literature about decision making among venture capitalists, this aspect of the data gathering process still can provide valuable

insights. How potential biases in data because of automated processes influence machine learning classifications are relevant relative to the purpose of this paper.

The relevant variables from the aggregated dataset include information about founder gender, geographic location of organizations and the number of capital rounds raised. The different datasets received from EQT have been merged into one single file with the help of SQL to facilitate the empirical analysis. Furthermore, there are observations where values are missing for certain characteristics, because of the automatic aggregation of different data sources. These observations have been removed to enable an analysis. Consequently, this is a limitation since certain founding teams where information is incomplete will not be taken into account.

The data has been filtered to only show Swedish founders and organizations. Also the time period 2010-2015 was chosen to ensure the data is comparable. The rationale of this decision is the need to ensure enough time has been provided for organizations to raise more than one round of capital. Still we acknowledge that the most recently founded organizations in the chosen time period might have had difficulties to raise numerous rounds. An organization founded 2010 has had more time to raise more than one round of capital relative to an organization founded 2015.

Gender of founders are taken directly from the aggregated dataset. The characteristic used for the EQT part of our paper is the percentage of female founders in each founding team, a continuous variable from 0 to 1. Data for the city of the organizations are available in the dataset. The distance in kilometers from Stockholm for the organizations in the dataset has been determined similarly as in part one of our paper.

Also, data about prior founder experience has been scraped with the help of Linkedin. A dummy variable has been used and prior founder experience of a minimum of one founder in an organization has been equaled to 1, whereas no founder with previous founder experience has been numerized as 0. The final proxy for this is data engineered to be useful on an organizational level. Resulting in the feature that organizations with at least one founder with previous founder experience are set to 1 and otherwise are set to 0.

Data for the number of rounds raised is part of the aggregated dataset. The success metric used is whether an organization has raised more than one round of funding. A dummy variable has been created where the value has been set to 1 for organizations that have raised more than one round of financing. Otherwise the value of the variable has been set to 0. The usage of only one success metric is a limitation of our study.

Naturally, the characteristics investigated in this study are limited and many other factors could influence the number of rounds raised. The investigation of additional success metrics for startups would be ideal. Naturally whether a startup reaches an IPO and continuous measures of valuation would be relevant to complement the number of rounds raised, in determining success. However, based on the available data in the EQT dataset the success metric used in this part of the paper is whether an organization raises more than one round of funding.

A limitation of the analysis in part two is the lack of characteristics about financial metrics of the organizations. Although these important financial variables are omitted in part two, the analysis is still valuable. For instance, the analysis is relevant from the perspective of VC decision making based on a limited number of founder characteristics. What if VC investors place substantial importance on the team in investment decisions as suggested by Gompers et al. (2020).

Table 5. Summary Statistics- Part 2

N = 296	Num_Male_ Founders	Num_Female_ Founders	Number_of_ Rounds	Distance from Stockholm (km)	Number of founders with prior experience	Serial Founder	Female Dominant	Share of female founders
Mean	1.363636	0.061818	1.447273	161.367636	0.654545	0.512727	0.007273	0.025091
Std	0.643996	0.269828	0.739410	240.640692	0.774082	0.500749	0.085125	0.109211
Min	1	0	1	0	0	0	0	0
25%	1	0	1	0	0	0	0	0
50%	1	0	1	0	1	1	0	0
75%	2	0	2	466	1	1	0	0
max	4	2	5	637	4	1	1	0.666667

Summary statistics of part 2, EQT dataset

There are 296 organizations in the sample of the dataset used in part two, after the process of data cleaning has been completed. Interestingly it can be observed that the average number of male founders in an organization in the sample is 1.36. In contrast, the average number of female founders per organization is only 0.06. Additionally, only 0.007 of organizations in the data are female dominated, that is, have a majority of female founders. Consequently, it is of interest to investigate the effects of usage of data reflecting overall male dominated startups for prediction of success. An interesting question is whether male founders consequently will be benefited or not vis-à-vis female founders in predictions of success.

Furthermore, it can be observed that only the fourth quartile of the organizations in the data in part two has raised more than one round of funding. As a consequence, only 25% of the organizations are categorized as successful according to the definition of success in our paper. Again in this paper we define successful organizations as those able to raise more than one funding round. A dummy variable is used for the success metric and raising more than one round of financing is equaled to 1, whereas raising one round of financing is numerized to zero. Further it can be observed that the top two quartiles of organizations have a founder with prior founder experience. Moreover, the bottom two quartiles of organizations are also located in Stockholm, denoted by 0 under Distance from Stockholm (km). For transparency definitions for what is defined as a funding round cannot be provided for part two of the paper.

V. Empirical Implementation

Ideally, we would have a comprehensive, exhaustive and accurate dataset of Swedish startups that seeked first round funding, relating to the priorly mentioned RCT. The data would distinguish different rounds (angel, incubator, VC), enabling an accurate definition of which rounds are venture capital and which are not. Furthermore, the data would include the distance from the start up to the venture capital firm, this given the findings that geographical location is a factor in success. Moreover, founder characteristics are of importance, especially university, given that specific venture capital labs or/and funding is set for specific universities. Conclusively, the dataset would contain an accurate array of financial, organizational and founders data.

From such a dataset an adoption of progressively more advanced statistical models would be utilized to infer whether a company succeeds or fails in securing venture capital funding. The models, in our case, begin with a Logistic regression of

the binary classification of success and failure, and culminate in a Random Forest ML method.

Logit Model

Similarly to Fuster et. al (2022) a random forest model is compared to a logit model. The logit model represents a less sophisticated prediction technology compared to random forest. In logit models the following link function is typical used:

$$\log\left(\frac{g(x)}{1-g(x)}\right) = x'\beta$$

See below for the advantages of a random forest model relative to a logit model for the empirical implementation in our paper.

Random Forest Model

The following is a discourse of the empirical framework used to examine the intersection of machine learning and venture capital firms. Machine learning models are the product of their data.

Utilizing the combined data from Sweden Tech ecosystem and Serrano, progressively more advanced statistical models are utilized to infer whether a company succeeds or fails in securing further venture capital funding rounds. The initial model is a simple logistic regression model[19], which is then superseded by a random forest machine learning model (see Breiman (2001) and Ho (1998)).

Random forest is a machine learning algorithm that works on the premise of aggregated decision trees. In the context of binary classification, random forest is the mode of the outputs of the decision trees. Decision trees are structures based on nodes and branches. Nodes are questions or conditions that split the data, whereas branches are the output of nodes and connect nodes in a downward fashion. The purpose of each tree is to determine a classification through a series of question/conditions (Nodes). Here ordering of nodes and condition choice is of importance in the efficiency of determining a class (Breiman (2001); Hastie, Tibshirani and Friedman. (2009)).[20] To such an end the concept of gini impurity will be proposed. Gini impurity is a measure of the frequency of incorrect classification of a randomly chosen element if labeled randomly according to the distribution of classes, and is defined as (James, Witten, Hastie, Tibshirani, and Taylor (2023); Hastie et al. (2009)):

$$\text{Gini}(S) = 1 - \sum p_i^2 \quad [21]$$

Where Gini(S) is the impurity of set S, p_i is the proportion of class i in S, and n is the number of classes. In relation to the decision tree, we utilize Gini impurity for optimizing node split (condition choices), with the goal to minimize Gini impurity (James et al (2023); Hastie et al. (2009)).[22]

---

[19] By technical nomenclature the model used is actually a logistic classification. This being the case as the target variable (More than one funding round) is categorical (binary). To minimize confusion the paper will stick to logistic regression.

[20] For example if one had to think about trying to guess a chosen number from numbers 1-10 using only yes no questions. Naturally the first question would be larger/smaller than 5 as the set of possible numbers would be halved.

[21] Rewritten from James et al. (2023) equation 8.6

[22] It would be remisive to not acknowledge that gini impurity is only one of many methods to optimize node pruning. See James et al. (2023) for more

Random forest inherently has a version of bootstrap aggregating (bagging). Bootstrapping is the processing where multiple samples are drawn with replacement from the original dataset. These new samples are known as bootstrapped samples. Bagging is a machine learning ensemble technique that utilizes bootstrapped samples to build independent models (decision trees) on the different bootstrapped samples. The aggregation aspect of bagging is then that output of the overall model (random forest) is the aggregated output of all the independent decision trees. For binary classification specifically, this is the mode output of these trees as mentioned.[23] The aggregation also makes the random forest model much more robust, making it usable on both clustered data and unclustered data (James et al. (2023)).

The pay off of using such a method is several fold. Consider first the averaging effect on a set of observations and variance: Given a set of n independent observations $Z_1 \ldots Z_n$, all having a variance of σ^2 then the variance of the mean is $\text{var}(\bar{Z}) = \sigma^2/n$. In other words, averaging a set of observations reduces variance. Applying the same logic to a set of B bootstrapped samples is a great method in decreasing variance and increasing test accuracy. We can define the bagging by denoting the different predictor outputs as

$$\hat{f}^1(x), \hat{f}^2(x), \ldots, \hat{f}^B(x)$$

which gives (James et al (2023); Hastie et al. (2009)):

$$\hat{f}^B_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x)$$ [24]

The final aspect of the random forest is building uncorrelated trees. Thus far each node in each tree has been given the full set of predictors p (independent features used to predict the dependent feature). Unpropitiously, aggregating several correlated quantities (as our current trees are) fails in reducing variance significantly, especially when compared to the aggregation of uncorrelated samples. To achieve trees decorrelation, each node, when it is approaching a split samples a random unique set of predictors m, where $m \subseteq p$. This ensures that the trees grown are less correlated, especially in the case where there are dominant predictors. To integrate this aspect in the above equation, we introduce: Θ_b - representing the characteristics of the bth tree, as such, random forest is defined by: (James et al (2023); Hastie et al. (2009)):

$$\hat{f}^B_{RF}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x; \Theta_b)$$ [25]

The purpose of the random forest model over that of a more simple logistic model is multipronged. The first and most prominent is the gain in flexibility. Random forests are far more versatile in capturing complex data structures and correlations than logistic regressions. The interactions between different features are automatically captured by a random forest model. In contrast, for logit explicit specialization and modeling of these interactions are needed (Hastie et al. (2009)).

---

[23] For continuous data predictions, that is regression, the mean is used instead of the mode
[24] The equation is most closely related to James et al. (2023)
[25] The equation is most closely related to Hastie et al. (2009)

Further, It is especially the case that random forest scales well with higher dimensionality. Logit does not scale as well with more dimensions relative to a random forest model. This is due to an inherent feature selection where a large number of features can be handled by a random forest model. A random subset of features for each split is considered by each decision tree in the forest. Thereby, irrelevant features are better handled by a random forest model (Hastie et al. (2009)).

The flexibility and dimensionality advantages discussed above imply that underfitting is reduced through the usage of a random forest model. Moreover, random forest is relevant for our data because of its possibility to handle variables without dummy encoding, which logit is not capable of (James et al (2023); Hastie et al. (2009)).

Given success in VC being complex to map, simple logit models may not be sufficient. As such, random forest is of interest given its often better capabilities in determining underlying relations in complex data.

Finally, our random forest model is built using the python package Oputna's parameter optimisation for the two dimensions.[26] In short, the optimisation finds the ideal values for a set of random forest parameters to maximize accuracy and cross validation accuracy.The choice of this optimization is later addressed.

As priorly mentioned the ideal data set would come from a natural RCT where the difference between treatment and control is simply the characteristics to be viewed. In addition to this a healthily sized, exhaustive dataset is of importance, for both the model and the significance of its output.

From such a dataset, we would utilize both random forest and logistic regression to make predictions concerning the importance and direction of each feature. From here our method differs from the method in Fuster et. al (2022). While many of the model evaluation metrics are the same, for example brier score, accuracy and ROC AUC it is adapted to our setting both in the construction of the random forest and its explanation.

Firstly, in the construction of the random forest model, we tuned the model parameters to maximize both accuracy and k-fold cross-validation accuracy. Accuracy is defined as (TP+TN)/(TP+TN+FN+FP).[27] Cross-validation accuracy is slightly more complex; dividing the model training set into k sections (in our case 7), one fold is used for testing while the remaining k-1 (6) folds are used to train the model. The overall cross-validation accuracy of the model is determined by averaging the accuracies obtained from each individual round of testing, where each distinct fold serves as the test set one time (M. Stone (1974) ; James et al (2023); López, López and Crossa (2022)). The choice behind maximizing accuracy, is to mimic that of a venture capital maximizing profit. Similarly, this logic holds for cross-validation but serves the concept of minimizing the model's overfitting susceptibility. From these results, we then run a robustness test to determine any difference between the ability of the two models. We utilize bootstrapping 100 samples of the test data to determine accuracy and variance. Furthermore, we utilize a placebo test to verify the ability of the two models.

Further, similar to Fuster et al. (2022) the Brier score is used to evaluate the models. The Brier score is defined as:

[26] See https://optuna.org/?fbclid=IwAR1ykIFE-rdfwX9YQIODd-66t7qMMeiOc81uCpRJJ9pEuwIkY2kz7ZuglA8

[27] TP = true positives, TN = true negatives, FP = false positives, FN = false negatives

$$\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2$$

where N is the number of predictions, f_i represents predicted probability of the event for instance i. f_i takes on values from 0% to 100% based on the probability that i is class 1. The actual outcome of instance i is represented by o_i and takes on a value of 1 for success and 0 for failure. The lower the Brier score the more confident a model is. For example if a model has threshold 50 and for instance i the model predicts 51% that i belongs to, or is a success then it will be classified as success. This would be reflected in the brier score for i as f_i - o_i being equal to 0.49. Whereas if the probability would be 99% then the score would be 0.1.

SHAP Values

The second aspect is the interpretation of the models. Here we utilize Shapley additive (SHAP) values, based on Lundberg and Lee (2017). In relation to ML, SHAP values are implemented as a method to explain any instance (a model prediction given a set of features) by dissecting the output into its features and the baseline prediction. The baseline prediction: base value E[f(X)] is the average prediction of the model across all possible inputs. It represents the prediction that would be made without any specific information about a given instance. The classical Shapley values are produced through the introduction of each feature into a conditional expectation function: $fx(S) = E[f(X) \mid do(XS = xS)]$[28] , here S is the subset of features conditioned on. Shapley values have merit in the uniqueness of fulfilling three properties, namely; local accuracy, consistency, and missingness (Lundberg, Erion, Chen, DeGrave, Prutkin, Nair, Katz, Himmelfarb, Bansal and Lee, 2020). See appendix C for a mathematical derivation of the SHAP formula. Culminating these properties results in the game theory Shapley value formula presented by Lloyd Shapley . To apply Shapley's theory to machine learning Lundberg and Lee (2017) utilized Lloyd's Formula and postulated SHAP (Shapley additive) Values. In brevity SHAP values are the Lloyd Shapley values of a conditional expectation function of the random forest model - in our case. The final SHAP equation is:

$$\phi_i(f,x) = \sum_{S \subseteq M \setminus \{i\}} \frac{|S|!(|M| - |S| - 1)!}{|M|!}[f_x(S \cup \{i\}) - f_x(S)]$$ [29]

The equation can be bisected into two aspects. The fraction represents (seen below) the weighting of each subset S, and ensures fair attribution of each feature's importance. $|M|!$ is the number of ways one can order all features M, or the number of sets of features one can form. This divides the number of ways to form a subset of features S ($|S|!$) multiplied by the number of ways to arrange the remaining features in the set *M* that are not in S (|*M*| - |S| - 1)! The fraction adjusts the weighting or importance of each subset based on the probability of its occurrence out of all possible subsets.

$$\frac{|S|!(|M| - |S| - 1)!}{|M|!}$$

[28] Here the use of do-notation is to indicate that this is not an observation rather a direct implementation. That is, $do(XS = xS)$, should be interpreted as setting Xs to xs. This is the only implication of this.

[29] The formula notation has been slightly altered from (Lundberg and Lee, 2017), to be consistent with the paper's form.

The final aspect of the formula is simply the contribution of the ith feature. recall from property 3 that the below expression would result in 0 in the case of no impact.

$$[f(S \cup \{i\}) - f(S)]$$

In short the final SHAP value $\phi_i(f, x)$ is the product of each feature i's marginal contribution and its occurrence or weight, sum over all possible features i in *M,* where $i \in M$.

SHAP facilitates the interpretation of ML model predictions. SHAP provides a framework in which additive feature importance in a ML context can be discussed. (Lundberg and Lee, 2017). This is relevant for our paper since we aim to investigate which founder and financial characteristics that impact ML predictions. SHAP values contribute to an understanding of the decision making process of the ML model in predicting which organizations will raise more than one round.

VI. Empirical Results And Analysis

Part 1. Serrano and Sweden Tech Ecosystem Part

A tuned random forest model is implemented on the data in part one. The tuned model is primed to maximize training accuracy and seven fold cross validation training accuracy. The choice of such metrics is to minimize overfitting while maximizing generalized accuracy. In this process numerous forests and trees have been generated to determine the tuned random forest model with the best possible parameters. For part 1 excluding the share of female founders data, the parameter optimization determines the ideal number of trees to be 94, with a maximum depth of two. Moreover, the optimisation determines that each node must have at least 18 samples to be eligible to conduct a split while also maintaining that any final node, a leaf node, must have a minimum of three instances. These parameters produce low risk of overfitting due to; the constraint depth of two (two node levels), the requirement of 18 samples for a split- restraining splits on insufficient data-, and finally the need for leaf nodes to have minimum three samples ensure that final decision are also based on sufficiently sized dataset. The minimum leaf node's size of three is understandable given the distance for Stockholm variety. The implication of these parameters being ideal indicates two possible aspects. The first is that the data is easy to predict, thus a more simple model is ideal. The second is that the data has complex patterns, but the model prioritizes minimizing overfitting and maximizing generalizability. The second explanation could be due to the fact that capturing these complex patterns is difficult and does not yield the confidence necessary to be worth incorporating. Finally, a more simple logit model is then implemented on the same data and the results for both models are showcased in table 6.

Table 6- Random Forest Summary Overview- Part 1 Excluding Gender

Model	Accuracy	Cross Validation Test Accuracy	Cross Validation Train Accuracy	Precision	Recall	F1 Score	Brier Score	ROC AUC Score
Logit	0.650794	0.634921	0.653061	0.684211	0.906977	0.78	0.212339	0.627907
Random Forest	0.730159	0.761905	0.666667	0.742143	0.953488	0.828283	0.199096	0.646512

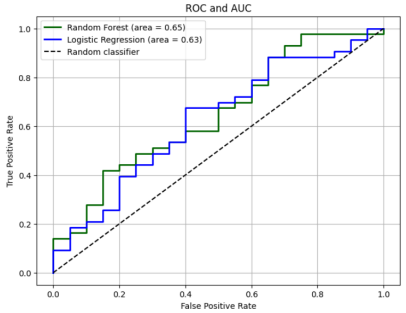
Overview of evaluation metrics for the logit and random forest models for part 1, merged Sweden Tech Ecosystem and Serrano dataset excluding gender.

The accuracy of the logit model for the data in part one is 65.08% and 73.02% for the tuned random forest model. The tuned random forest model also has higher cross validation test and train accuracies than the logit model. Further, it can be observed

that both models display relatively high levels of F1 scores, which is a relevant metric to evaluate in the case of unbalanced data, as in our case. The F1 score is a weighted average of precision and recall.[30] True positives in this context are startups that have raised more than one round and were classified correctly. False positives are startups that did not raise more than one round yet were predicted to have raised more than one round. The F1 score is higher for the random forest model however.

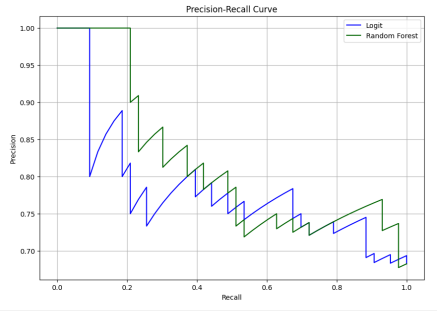
When predicting a class based on inputs, models will assign a probability to the likelihood of that instance being a class. A threshold for these probabilities are also chosen. To exemplify this, consider the situation where an input(s) x is predicted to belong to the class A with probability 60%. This means that for any threshold below 60 will classify this point as class A, however once above 60%, the output will be B (for binary classification). Brier score provides a measure of the accuracy of classifications in relation to their relation to the model's predicted probability of said classification. In short the Brier score is a measure of the model's confidence when predicting, where 0 is highly confident and 1 is the least confident. The Brier score for both models are low (with random forest being the better of the two), highlighting that both models are quite confident in their classifications. In conclusion the Random forest exhibits superior performance to the logit model in predicting whether startups will raise more than one round of funding, with better accuracy, cross validation accuracy, F1 and Brier score.

Figure 1: ROC Curves Part 1 Excluding Gender



Receiver operating characteristics curves for the logit and random forest model for part 1 excluding gender. A larger area under curve means the model is better at distinguishing between organizations that have and have not raised more than one round. It also reduces overfitting.

Figure 2: Precision Recall Curves Part 1 Excluding Gender



Precision-Recall curves for the logit and random forest model for part 1 excluding gender. The random forest model displays the best combinations of precision and recall.

Similar to Fuster et al. (2022) ROC (receiver operating characteristics) curves are also featured in our paper. These plot the true positive rate and the false positive rate at every threshold. The optimal threshold is where the curve is bent the most toward the north west corner. Here the true positive rate is maximized and the false positive rate minimized (James, 2023). For example for the logit function in figure 1 it would be at (0.4, 0.7).

In addition to the ROC the AUC (area under curve) is also of importance. The AUC represents degree or measure of separability. It displays how much the model is capable of distinguishing between classes. In short a larger AUC means the model displays greater effectiveness in distinguishing between successful and unsuccessful startups (James, 2023).

[30] Precision = (TP)/(TP + FP), Recall = (TP)/(TP + FN)

Figure 1 illustrates the ROCs and AUCs of the random forest and logit model. While both models produce relatively modest AUC scores, it is clear that the tuned random forest model is superior to the logit model with a larger AUC of 0.66 relative to 0.63. The random forest model does not outstrip logit at every threshold, but is superior when aggregating across all thresholds (seen by AUC). The low AUC scores in relation to the above tabled scores, implies that while the model performs well at the optimized threshold, it is poor when considered against all thresholds. The reasoning behind this is due to the assumption that the amount of venture capital invested is the same over our sample period. With this assumption the model has not been tuned for cross-threshold performance. The assumption is faulty and further literature should review the performance across thresholds. The model is also impacted as in actuality for years 2010-2012, 2010 had the highest VC investment followed by 2011, 2012 (Härd, 2022). As such companies in 2010 may have had an easier time getting investment, but may also have received more in their first rounds. Ideally this data would be available for the model and is something that should be further researched.

Precision and recall curves are present in Figure 2. The ideal model maximizes both precision and recall simultaneously resulting in a precision recall curve that is bent outward north east in Figure 2. Based on figure 2 the tuned random forest model is closer to the ideal indicating a lower rate of false positives, something we wish to avoid.

Given the above it is fair to assert that the random forest model is the better choice. While the focus of the following analysis will be on SHAP values for random forest, traditional values will also be analyzed.

The Random forest model exhibits 66.67% true positives, 7.93% true negative, 1.59% false negatives and 23.81% false positives. Appendix D Panel I visualize these in a heat map. The Feature importances are ordered as follows: the net sales ratio has the largest reliance at 32.66%. This is followed by the quick ratio at 28.33%, ROE at 25.21%, serial founder at 7.12% and lastly distance from Stockholm at 6.67% weighting. Thus far the presented results have not been evaluated against the data distribution and collection, which have major implications.

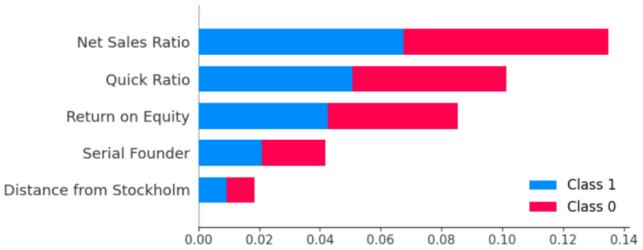
The inclusion of funding rounds pre professional VC investments in part one of our paper likely influence the results. Literature by Howell (2020), highlights how the winning of a startup competition can serve as a certification that signals a high quality of an organization to VC investors. Howell's results underlines that a higher percentage of startups raise capital from angels and VCs combined, compared to VC investments alone. Consequently, the inclusion of early pre VC funding rounds in the data might explain the large share of successes. In table 2 it is shown that 68.1% of observations in the data in part one excluding gender represent organizations that have raised more than one round of funding. It is likely that this number would be smaller if early pre VC rounds would have been excluded in the data gathering process. In appendix D, panel I it can be observed that the vast majority of the observations in the test data are classified as successes by the tuned random forest model in part one. This, for example, leads to quite a large number of false positives, namely 15. False positives in this context are firms that have raised only one round, yet that are classified as having raised more than one round. This reduces the precision of the model totaling at 74.21%. At the same time the recall remains at a higher level of 95.35%. The fact that there is only one false negative in the test dataset, means this false negative does not impact recall to the same extent as the 15 false positives influence the precision number. There might be a number of other explanations why the majority of observations are successes. For example, financial

information in the Serrano database might be missing for firms that have failed to raise multiple rounds. This potential selection bias could have resulted in certain startups that had raised only one round being omitted from the final dataset. However, the inclusion of early angel and seed rounds might be another reason. An early angel round might serve as signaling through certification similarly to winning a startup competition as presented by Howell. Relating to this, Hellman and Thiel (2015) highlight how the angel market works as a screening mechanism for subsequent VC dealflow. In short the high share of successes influences the machine learning prediction with a majority of observations categorized as successes. Although, we should acknowledge that these potential explanations for our results cannot be established with certainty.

In Appendix E partial dependence plots for the different characteristics in the tuned random forest model in part one are displayed. The partial dependence plots only showcase the magnitude of the marginal contribution to success prediction, not the direction. In panel I it can be seen that the marginal contribution to the ML prediction of the ROE variable is highest for values below 0% ROE. This is likely a reflection of the difficulty for most startups to be profitable. Both those that fail and succeed might struggle with profitability which could influence the input data and hence the success prediction, either negatively or positively. In panel II it can be observed that the marginal effect for the success prediction is the strongest at around a net sales ratio of 0. Panel III highlights that for quick ratio the contribution to the prediction of success is the strongest for values above about 1. Interestingly, panel IV suggests the marginal effect for the prediction is the strongest for values of distance from Stockholm at zero. Panel V underlines that prior founder experience has a higher marginal effect for the prediction of success when compared to lack of prior experience.

While the above traditional analysis are of use, SHAP values allow for a better understanding of the direction of magnitudes that each plot may have.

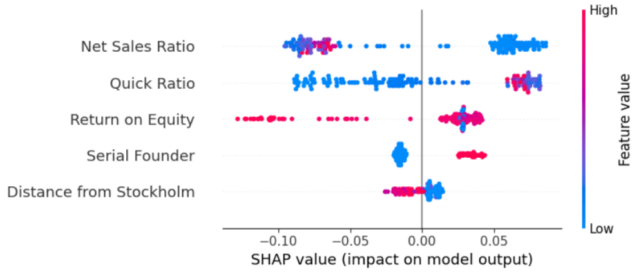
Figure 3. Absolute Mean SHAP Value- RF Part 1 Excluding Gender



Absolute mean SHAP values for random forest part 1, excluding gender. The larger the bar the larger the absolute mean SHAP value of the variable. The larger the absolute mean SHAP value the larger the magnitude of the contribution of a feature to the ML prediction of raising more than one round. Here class 1 represents success whereas class 0 represents failure.

Figure 3 indicates that the net sales ratio has had the largest overall impact on the success prediction. However, it is not possible to establish whether the effect is negative or positive based on Figure 3. Interestingly the financial metrics of the firms in the sample has the highest overall impact on the success prediction. The characteristics of the founding team, namely serial founder and distance from Stockholm has the least overall impact on the success predictions of the tuned random forest model. It should be reiterated that SHAP values relate to correlational effects and not causal effects.

Figure 4. SHAP values- Tuned RF Model Part 1 Excluding Gender



SHAP values for random forest part 1, excluding gender. A blue value indicates a low value of a variable and red indicates a high value of a variable. Blue for distance from Stockholm indicates a 0 distance from Stockholm and red a high distance from Stockholm. Red indicates a value of 1 for the serial founder variable indicates a serial founder in a team and blue represents a value of 0 representing a lack of serial founders in a team. A positive SHAP value means a positive contribution of a variable to the ML prediction of raising more than one round.

Figure 4 outlines that high (red dots) and medium (purple dots) net sales ratios are negatively correlated with success, while low ratios are more ambiguous. Another observation is that the high and medium quick ratios are correlated with success, while low quick ratios are weakly correlated with failure. While low ROEs are correlated with success, high ROEs are more ambiguous. For the serial founder variable it can be observed that organizations with a serial founder contribute positively to the success prediction. This is likely influenced by the low number of organizations with serial founders in the dataset. Lastly, for the distance from Stockholm parameter the blue dots in Figure 4 that represent organizations from Stockholm have slightly positive SHAP values. Most red dots, representing a high distance from Stockholm, have negatively influenced the success prediction. Again it should be noted that the magnitude of the contribution to the ML prediction for the distance from Stockholm and serial founder variables are lower relative to the financial metrics. Often there is a spread of low (net sales ratio and quick ratio) or high (ROE) results in terms of the SHAP values. The spread is indicative of complexity in financial analysis, indicating that the impact of each financial metric on a startup's success can significantly vary between companies.

The contribution of the net sales ratio and ROE variables can be related to literature by Puri and Zarutskie (2012). For example, the authors study VC backed and non VC backed startups. Results suggest scale to be an important firm attribute that venture capitalists focus attention on. Specifically the authors point out the importance of the scale potential rather than short-term profit. While we focus on organizations that have raised a minimum of one round, organizations that have raised multiple rounds are likely similar to the VC backed firms in the study. The positive contribution of a low value of the ROE variable in figure 4 is aligned with the relatively less importance of profitability for startups. However, figure 4 displays that high sales numbers relative to the industry average negatively contributes to the prediction. This hence does not suggest a larger scale, here more revenue relative to the industry, contribute positively to the prediction of raising more than one round. The negative SHAP values for high net sales levels in Figure 4 might be influenced by the inclusion of early angel and seed rounds. This implies that the second round for certain firms in the sample is in a relatively early phase of development. At early stages many firms might receive financing based on the future potential rather than past or current performance as suggested by Puri and Zarutskie (2012).

The contribution of the quick ratio variable to the ML prediction can be related to literature by Davila and Foster (2007). The authors point out that firms with negative

cash flow levels in early stages are usually those in need of venture capital. Further, the authors point out that numerous private funding rounds are usually needed to turn the cash flow of companies backed by VCs positive. Davila and Foster write that financial planning can help organizations in the negotiation for new funding. Financial planning helps with cash management in firms constrained in terms of cash. The authors find that the usage of financial planning is not as substantial for companies not funded by VCs compared to VC financed startups. It is also found that HR and strategic planning are introduced earlier compared to financial planning in non VC backed firms. We do not focus on the distinction between non VC backed and VC backed startups. However, similar to VC backed firms adopting financial planning to a greater extent compared to non VC backed firms, it might be the case that firms that raise multiple funding rounds employ more sophisticated financial planning systems. This might hence have increased the liquidity of these firms and hence the quick ratio. This could reflect itself in the ML prediction where companies with higher quick ratios might be benefitted.

Bernstein et al. (2017) find evidence of the importance of the founding team and human characteristics for the success of early stage ventures in attracting investors. Figure 4 suggests a larger magnitude of the contribution of the financial characteristics of firms to the ML prediction relative to the founder characteristics. The authors conduct an experiment to establish causal effects. Unlike the authors we focus on how the usage of founder and financial characteristics impact ML success predictions. Still the authors acknowledge that the results do not indicate a lack of importance of nonhuman assets. Instead the authors relate their results to Rajan (2012). The importance of human capital in early development and the need for the founder to make him or herself possible to replace is highlighted by Rajan (2012). The rationale behind this need for replacement is the provision of control rights to investors, hence enabling the raising of external financing, for example by a VC. The relatively higher importance of for instance the net sales ratio to the founder team variables is contrary to the results by Bernstein et al. (2017) on the importance of the founding team to get investor interests. However, our results are likely influenced by the focus on second round funding as the success metric. This implies the organizations in our dataset are at a relatively later stage of development compared to very early stage ventures. It could be that firms have moved from differentiation, closer to standardization in the language of Rajan (2012).

The inclusion of prior founder experience relates to literature by Gompers, Kovner, Lerner and Scharfstein (2010). Gompers et al. (2010) highlight that prior successful entrepreneurial experience displays greater likelihood of future success. Our results only indicate whether a founder has prior founder experience and not the success of that entrepreneurial experience. However, the importance of prior founder experience has also been addressed by Gompers et al (2020) that raise entrepreneurial experience as one of the factors representing the team in the decision making process of VCs. The authors specifically underline the importance of a prior relationship with a particular VC for entrepreneurs. The results in figure 4 for the serial founder variable indicate a positive effect for prior founder experience and a slight negative effect for a lack thereof. From the perspective of the direction of the SHAP values, the red dots that represent prior founder experience are aligned with literature that highlights the importance of prior founder experience. In short the ML predictions favor founders with prior founder experience.

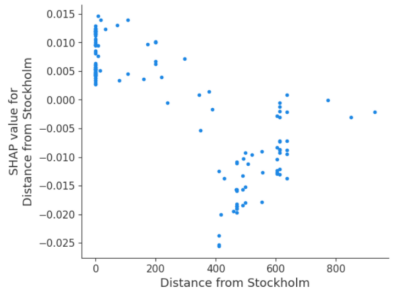
The slightly positive SHAP values for a zero kilometer distance from Stockholm in Figure 4 could reflect the dominance of Stockholm startups in the dataset. However,

simultaneously it might also be indicative of the importance of being near a hub for VC investments to attract capital.

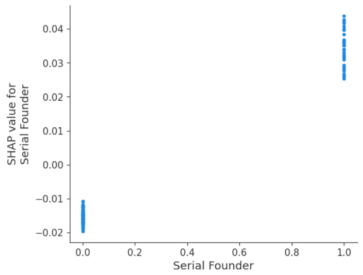
Chen, Gompers and Kovner (2010) find geographical clustering in three US cities among US venture capital firms. Chen et al. (2010) find that the degree of localization is higher within the VC industry compared to the rest of the financial industry. In figure 4 somewhat positive SHAP values for a low distance from Stockholm are shown. Moreover, a high distance from Stockholm, represented by red dots, has negative SHAP values. This implies that a low distance from Stockholm positively contributes to the prediction of raising more than one round. At the same time, a higher distance from Stockholm negatively influence this prediction. This might be a reflection of Stockholm as a VC cluster in Sweden, which can influence the ML prediction through the input data. Our results are aligned with findings by Chen et al. (2010). This could potentially also relate to the importance of networks and referrals in VCs generation of deals highlighted by Gompers et al. (2020). Stockholm based startups might be more networked compared to non Stockholm based organizations. This might explain the positive contribution to the prediction of a low distance from Stockholm. Further, the similarity between Stockholm based founders and VCs could potentially relate to Ewens and Townsend's (2020) discussion about homophily. If VCs are more inclined to invest in those similar to themselves this might also entail geographic location of the startup and consequently the founders. From this perspective one would expect organizations located near Stockholm as a hub for venture capital to attract more attention from VC investors and consequently potentially raise more funding rounds. This in turn might have increased the number of organizations located in Stockholm in the dataset. Again a ML prediction is a product of its input data.

Tian (2011), presents a monitoring hypothesis that implies that monitoring costs are reduced for firms located close geographically to VCs. As a consequence, Tian claims VC investors might conduct less funding rounds. On the contrary organizations located far from VCs might receive staged investments because of the higher monitoring costs. The result presented by the author supports this hypothesis. Tian reports a regression with the number of funding rounds as the dependent variable and different independent variables for distance measures. Unlike Tian, our paper does not include the distance between individual VCs and organizations. However, we use a variable that represents the distance from Stockholm of organizations. This can be valuable in a Swedish context where the majority of venture capital is raised by firms in Stockholm. Naturally other VC clusters might exist in Sweden which highlights a limitation of our study. The results presented in part 1 of our paper only takes into account whether a startup has raised more than one round of funding. In reality a VC might stage investments in more than two rounds. Also the inclusion of early seed and angel rounds negatively influence the possibility to draw conclusions about VC staging in our paper. Our results indicate that a high distance from Stockholm negatively influences the success prediction. This is contrary to Tian's monitoring hypothesis where one would expect a large distance from Stockholm to positively influence the prediction of raising more than one round.

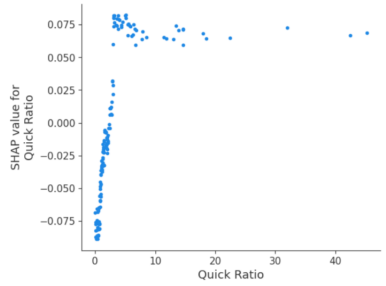
Figure 5 - Dependence Plots- Part 1 Excluding Gender



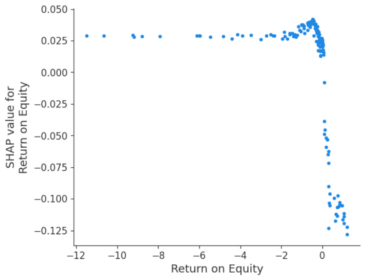
Panel I. Distance from Stockholm



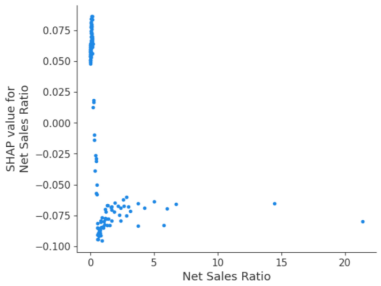
Panel II. Serial Founder



Panel III. Quick Ratio



Panel IV. Return on equity



Panel V. Net Sales Ratio

Figure 5 displays dependence plots for the different variables used in the random forest model for part 1 excluding gender. Positive SHAP values imply a positive contribution of a variable to the ML prediction of raising more than one round.

Figure 5 showcases dependence plots for the different independent variables used in the random forest model in part one, excluding gender. In panel I it can mainly be observed that most of the SHAP values are positive for a distance of 0 kilometers from Stockholm. For most of the distances above 0 distance from Stockholm the SHAP values are negative. Somewhat of a downward trend in the SHAP values as the distance from Stockholm increases can be observed. Thus, for larger distances, the distance from Stockholm variable negatively impacts the success prediction of the tuned random forest model in part 2. Panel II depicts SHAP values in relation to the serial founder feature. The binary feature exhibits a positive correlation; the serial founder attributed is correlated with success. Moreover, data in Panel III expounds the relation between quick ratio and SHAP values. While the concentration around 0 is difficult to classify, it appears that the SHAP values increase as the quick ratio increases. Panel IV indicates higher ROE negatively contributes to the ML prediction of raising more than one round. Panel V suggests that the net sales ratio for the

organizations with the lowest sales ratio close to zero contribute positively to the success prediction.

The percentage of female founders has been added to the random forest model in part one. Table 7 displays an overview for the model in part one including gender. It should be noted that certain observations have been dropped for the model including gender as described in section IV. The estimators (in this case weak learner trees) in the cross validation process for the model in part one including gender are 245, with a maximum depth of 14 levels or questions, used for splitting.

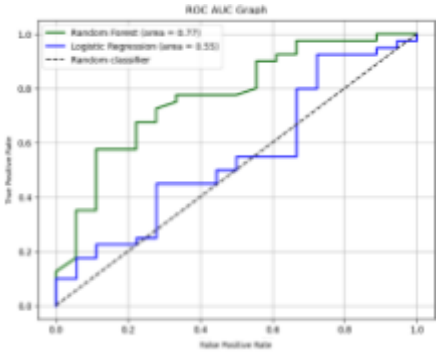
Table 7. Random Forest Summary Overview- Part 1 Including Gender

Model	Accuracy	Cross Validation Test Accuracy	Cross Validation Train Accuracy	Precision	Recall	F1 Score	Brier Score	ROC AUC Score
Logit	0.689655	0.640873	0.684211	0.703704	0.95	0.808511	0.216152	0.554167
Random Forest	0.775862	0.640873	0.729323	0.775510	0.95	0.853933	0.173163	0.772917

Overview of evaluation metrics for the logit and random forest models for the part 1, merged Sweden Tech Ecosystem and Serrano dataset including gender.

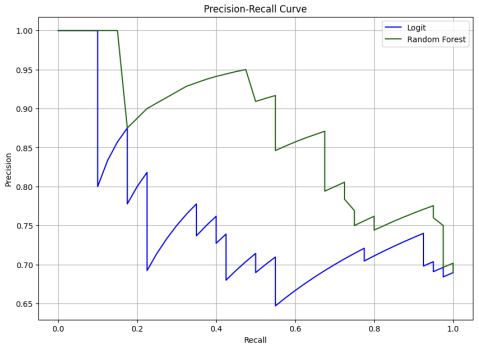
The accuracy and cross validation train accuracy increases for both the tuned random forest and logit model. The tuned random forest model has a train cross validation accuracy of 72.93%, and hence appears superior to the logit model. The F1 score is also higher for the tuned random forest model at a value of 85.39%. Also the F1 score is higher for both the tuned random forest model and the logit model that includes gender compared to the models that exclude gender. This is mainly due to a reduced precision of the models that negatively impact the F1 score. The Brier score is also lower for the random forest model which is positive for the confidence of the model's prediction.

Figure 6. ROC Curves RF Part 1 Including Gender



Receiver operating characteristics curves for the logit and random forest model for part 1 including gender. A larger area under curve means the model is better at distinguishing between organizations that have and have not raised more than one round. It also reduces overfitting.

Figure 7. Precision Recall Curves Part 1 Including Gender



Precision-Recall curves for the logit and random forest model for part 1 including gender. The random forest model displays the best combinations of precision and recall.

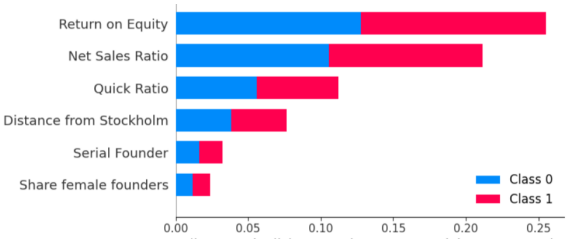
As showcased in Figure 6 the tuned random forest model including gender has a notably larger AUC than the logit model. Further, the precision recall curve for the tuned model indicates that the tuned model has superior combinations of precision and recall.

See Appendix F for the confusion matrix together with the explanatory weights for the tuned random forest model in part one including gender. Similarly to the model excluding gender the majority of the observations in the test set have been classified as successes for the model including gender (see panel I). This is likely influenced by the fact that the majority of the observations in the data are represented by organizations that have raised more than one round. Panel II interestingly shows that the financial characteristics all have the highest explanatory weight for the tuned random forest model including gender. Distance from Stockholm is the founding team variable with the highest explanatory weight. The share of female founders and the serial founder variables both have the lowest explanatory weight of the model.

Interestingly the team is the most important factor for decision making for many VCs as highlighted by Gompers et al. (2020). At the same time our results for part one indicate that founder team characteristics have the least explanatory weight for ML success predictions. Literature by Kaplan et al. (2009) that highlights the importance of the business relative to the team for investment decisions is relevant in this context. Over time the authors show a tendency for founder replacement over time. Our results indicate that the financial characteristics of firms influence the ML predictions the most. This might be influenced by the somewhat later stage focus in our paper, with a focus on second round funding. As a firm matures the original founders might be less important relative to the financials of the business. However, still the limitation in terms of not distinguishing between early angel and seed rounds and later stage VC is hereby noted. This aspect of our data gathering process makes these interpretations more difficult since for certain firms the second round might represent a seed or the first early VC round.

Appendix G displays partial dependence plots for the model in part one including gender. The interpretations of panel I, II and IV are relatively similar to the interpretations of the partial dependence plots for the tuned random forest model in part one excluding gender. However, the contribution to the ML prediction is the strongest for values slightly above 0 and the magnitude of the contribution then declines as the quick ratio increases in panel III. For the distance from Stockholm variable the magnitude of the effect of the variable on the prediction is strongest for distances above 600 kilometers, followed by distances between 0 and about 400 kilometers. The effect has the lowest magnitude for values between 400 and about 600 kilometers. The partial dependence plot for the share of female founder variable indicates the magnitude of the effect is the strongest for 0 to 0.2 share of female founders. Then the magnitude of the contribution to the prediction declines as the share of female founder variable increases. This is likely influenced by the low number of startups in the sample with a high share of female founders.

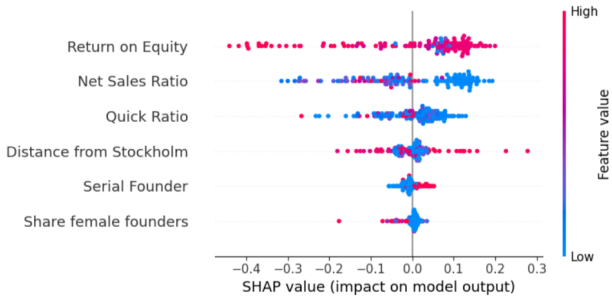
Figure 8. Absolut Mean SHAP Value- RF Part 1 Including Gender



Absolute mean SHAP values for random forest part 1, including gender. The larger the bar the larger the absolute mean SHAP value of the variable. The larger the absolute mean SHAP value the larger the magnitude of the contribution of a feature to the ML prediction of raising more than one round. Here class 1 represents success whereas class 0 represents failure.

In Figure 8 it can be observed that ROE has the largest impact on the ML predictions for the model in part 1 including gender. This is followed by the net sales ratio and quick ratio for the financial metrics. The change is likely due to dropped values that might have altered the results slightly. However, when gender is added to the model it can be seen that the share of female founder variable has the lowest impact on the predictions of the model. Again Figure 8 can only be used to draw conclusions about the magnitude of the impact of the different variables. In order to discuss the direction of these impacts see Figure 9

Figure 9. SHAP values- Tuned RF Model Part 1 Including Gender



SHAP values for random forest part 1, including gender. A blue value indicates a low value of a variable and red indicates a high value of a variable. Blue for distance from Stockholm indicates a 0 distance from Stockholm and red a high distance from Stockholm. Red indicates a value of 1 for the serial founder variable indicates a serial founder in a team and blue represents a value of 0 representing a lack of serial founders in a team. A blue value also represents a 0% share of female founders in a founding team and red a 100% share of female founders. A positive SHAP value means a positive contribution of a variable to the ML prediction of raising more than one round.

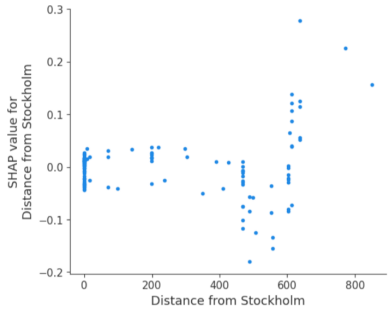
The directions for the SHAP values for ROE and net sales are relatively similar to the model excluding gender. For quick ratio it can be observed that high quick ratios negatively contribute to the ML prediction, which is the opposite to the model excluding gender. The effect of the distance from Stockholm variable is more difficult to interpret in figure 9 compared to in figure 4. The direction for the serial founder variable is similar to the model without gender. However, it appears as though the magnitudes are greater for the model including gender. Figure 9 shows that a high share of female founders, represented by red dots, negatively influence the success prediction of raising more than one round. At the same time the low share of female founders, represented by blue dots has a SHAP value of 0 or at slightly above 0.

The results on the effect of the share of female founder variable in figure 9 relate to prior literature that highlights how female founders are disadvantaged in the context of investor interest. In the figure, fully female teams contribute negatively to the success prediction in the tuned random forest model. Fully female teams in figure 9 are represented by red dots, whereas fully male teams are represented by blue dots. Although the magnitude of the contribution is low the direction is still negative for female teams. That is it contributes negatively to the ML prediction that organizations raise more than one funding round. Ewens and Townsend (2020) for instance proxy interest among venture capitalists as whether a startup's profile is shared on AngelList. In contrast, to the authors our results cannot provide evidence of individual investors and their potential gender biases. However, our results can be interpreted from the perspective of how historic data of founder characteristics influences machine learning success predictions of organizations. Considering that the adoption of machine learning in the venture capital context is expanding (Bonelli, 2022), these are interesting empirical results. Ewens and Townsend (2020) find that

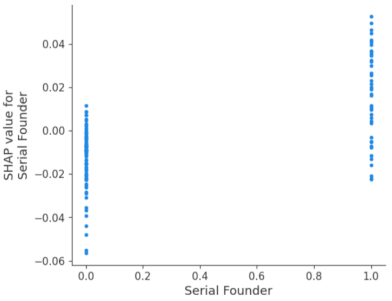
female founders experience less success in attracting interest from male investors compared to male founders. If a venture capitalist would rely on machine learning predictions similar to those in our paper there is a risk that gender biases could become institutionalized. The insignificance of gender for the ML prediction could lead investors to deprioritize diversity in capital allocation. This could have negative effects on equality in a VC context. This highlights the importance for critical thinking in the context of usage of ML algorithms in a VC context. It should be acknowledged that the ML model in our paper is likely less sophisticated than the models used by established VCs. Naturally it is difficult to retrieve information about the exact algorithms used by VCs since these are often proprietary.

Figure 10 showcases dependence plots for the different characteristics in the part one model including gender. Here focus is placed on analysis of the SHAP values for the share of female founder variable. It appears as though the SHAP values are slightly positive for a 0% share of female founders. Then it seems the SHAP values decrease as the share of female founder variable increases. Thus, the results in figure 10 suggest a higher share of female founders contribute negatively to the ML prediction of raising more than one funding round. This also relates to the above discussion related to literature by for instance Ewens and Townsend (2020). The authors also mention homophily as a potential explanation for the discrimination of female founders. Relating to the institutional background, there are few VC firms with female partners in Sweden. If the majority of VC partners that allocate capital are male this can lead to inequality in who raises numerous capital rounds. From the perspective of homophily it could be the case that male investors relate more to male founders. This might have influenced the number of female founders that received venture funding. This in turn might influence the ML prediction since the dataset is male dominated. Again ML algorithms are a product of their input data.

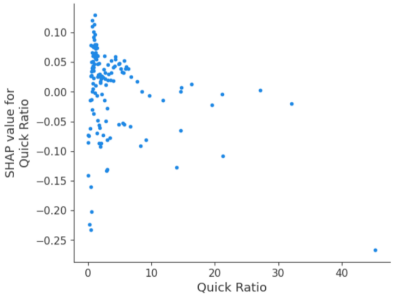
Figure 10. Dependence Plots- Part 1 Including Gender



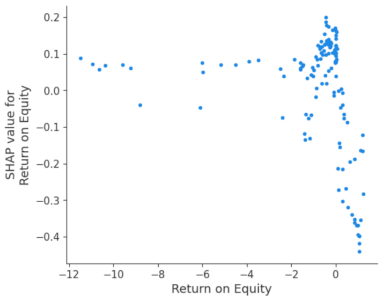
Panel I. Distance from Stockholm



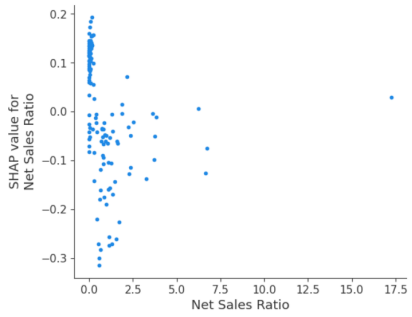
Panel II. Serial Founder



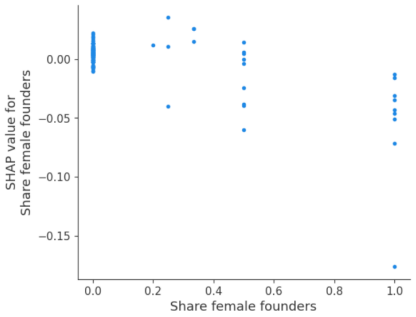
Panel III. Quick Ratio



Panel IV. Return on equity



Panel V. Net Sales Ratio



Panel VI. Share of Female Founders

Figure 10 displays dependence plots for the different variables used in the random forest model for part 1 including gender. Positive SHAP values imply a positive contribution of a variable to the ML prediction of raising more than one round.

Although our method is based on Fuster et al. (2022) some fundamental differences naturally influence the results. The difference in quality in terms of our substantially smaller dataset has already been noted. Essentially the authors use data more representative of a broader population. Additionally, Fuster et al. (2022) differ since they analyze mortgage data in contrast to the startup and founder data analyzed in our paper. While the authors for instance use default of mortgage as the target variable our target variable is whether organizations raise more than one round. Furthermore,the inclusion of more non-financial factors relative to Fuster et al. (2022) adds more complexity to the model in our paper.

Then Fuster et al. (2022) differ in their usage of triangulation to investigate potential inequalities in predictions of default. Triangulation in Fuster et al. (2022) is where ML models might indirectly infer effects of race when not explicitly included in the model. For example, zip codes tied with income, or even frequented stores/purchases could have different correlations with different ethnic groups. The implications of this is that a form of unintentional inferred discrimination could become present in models. The absence of this method for our paper is due to the limited possibility of triangulation. To elaborate, distance from Stockholm is a one dimensional metric, were it to include direction and distance, that is a specific location, the risk of triangulation would increase, however this is not the case. The Serial founder factor bears risk of triangulation, to the extent that discrimination exists in the very aspect of having already founded a company. As such this factor could cause model triangulation, but only if there is significant startup discrimination to begin with. The serial founder argument can also be extended to the financial metrics.

Here, our approach differs since we unlike Fuster et al. (2022) use SHAP values to investigate how different characteristics impact ML predictions. Then we attempt to analyze which founder profiles are benefited by the ML algorithm, from the perspective of geographical location, prior founder experience and gender.

Given these differences it should still be noted that Fuster et al. (2022) similar to the results in part 1 of this paper find that the random forest outperforms the logit model. Similar to the authors we find higher accuracy for the random forest model compared to the less sophisticated logistic regression. Fuster et al. (2022) implements two random forest models, one with a race variable and one without. Here our approach differs since it implements the model with and without gender. Both models have an AUC of 0.86, when rounded in Fuster et al. (2022). This is

around 0.1-0.2 higher than the AUC for our default and random forest models in part one. The author's higher AUC is in part related to a markedly larger dataset. The data includes millions of observations. Additionally, we only use datasets with 210 and 191 observations in part 1. The authors analyze the credit market and US mortgages where data is much more structured and available compared to the private venture capital scene analyzed in our paper.

The inclusion of SHAP values can be related to Griffin et al. (2023). Utilizing a tree based algorithm (which random forest is) called Gradient Boosting Decision Trees (GBDT), the authors analyze dealer markups in the municipal bond market. The machine learning method is implemented to understand and predict markup behavior based on factors such as dealer characteristics and practices. The authors employ SHAP values to quantify contribution of different variables to predict markup for a specific trade. In such a manner, we adopt and adapt the method of the authors to the venture capital industry. In our paper the raising of more than one round is predicted based on founder and financial characteristics.

Part 2. EQT Part

Based on the data in part two, a logit and tuned random forest model were constructed. The tuned model is primed to out perform on the 7 fold cross validation average accuracy similar to in part one. The number of estimators, that is decision trees, used in the process amount to 465 trees, with a maximum depth of four levels or questions, used for splitting, as in part one. Depth of four suggests that the model is designed to capture some level of interaction but is restrained to prevent overfitting. As such a lower depth secure is good against overfitting.

Table 8. Random Forest Summary Overview - Part 2

Model	Accuracy	Cross Validation Test Accuracy	Cross Validation Train Accuracy	Precision	Recall	F1 Score	Brier Score	ROC AUC Score
Logit	0.662921	0.652015	0.671921	0	0	0	0.196698	0.668966
Random Forest	0.719101	0.684982	0.729557	0.833333	0.172414	0.172414	0.190487	0.738218

Overview of evaluation metrics for the logit and random forest models for the part 2, EQT data

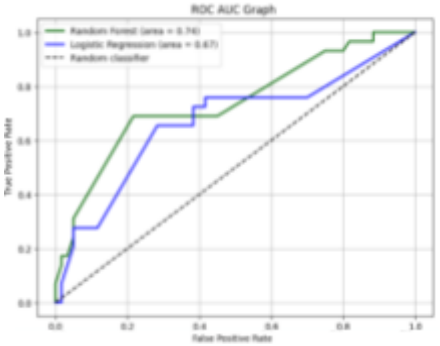
The accuracy of the tuned random forest model, that treats the variable of female dominant startup as a continuous measure is 71.19% as displayed in table 8. Taking into account cross validation similar to Fuster et al. (2022), the cross validation test accuracy is 65.2% for the logit model. As observed the tuned model has a higher cross validation test accuracy at 68.49%. From this perspective the tuned random forest model appears to be better at prediction relative to the default model from the perspective of model accuracy. The Brier score is also slightly lower for the random forest model, indicating it has somewhat more confidence in its predictions.

Interesting observations can be made from the perspective of precision and recall of the logit and tuned random forest models. In short the logit model in table 8 correctly predicts 0% of successful observations, of the observations classified as successes. The recall of 0% of the logit model indicates that none of the actual successes were predicted to be successes. This could likely be explained by the skewness in the data as explained in section IV. The vast majority of observations in the data in part two are categorized as not successful, that is represented by organizations with only one round of funding. It is interesting to note how the model incorrectly classifies observations as not successful from this perspective. For the

tuned random forest model the precision is high at 83.33%, with a low recall of only 17.24%. This can likely be explained by the skewed nature of the data. It is difficult to attempt to compare the results in section one with those in section two because of the differences in the datasets. Further, because of the limited information about the data aggregation process and variable definitions in part two this also makes a potential comparison difficult.

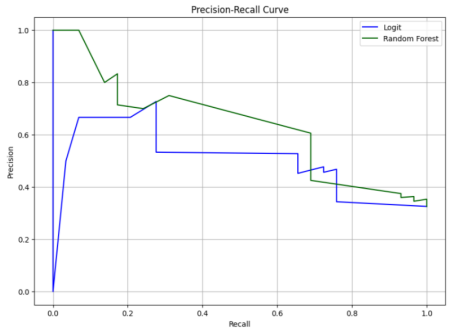
The tuned random forest model at first appears somewhat better than the default model. However, further analysis highlights that neither of the models perform very well at the binary classification from the perspective of the F1 score. The F1 score is relevant in this context for example, because of the skewness in the data. The F1 score of 0% for the logit model and 28.57% for the tuned random forest model can be explained by the low recall of both models. This underlines the difficulty to create a highly accurate binary classification model for skewed data.

Figure 11. ROC Curves- Part 2



Receiver operating characteristics curves for the logit and random forest model for part 2. A larger area under curve means the model is better at distinguishing between organizations that have and have not raised more than one round. It also reduces overfitting.

Figure 12. Precision Recall Curves- Part 2



Precision-Recall curves for the logit and random forest model for part 2. The random forest model displays the best combinations of precision and recall. Note that the line 0-1 for recall 0 is due to logit model having a 0 recall, 0 precision and 0 fi score

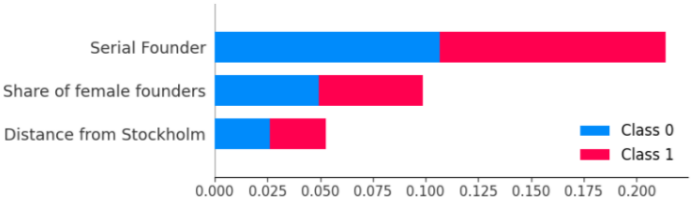
ROC curves are also reported for part two in figure 11. Here the tuned random forest model appears to be the preferred relative to the logit model with a larger AUC. The interesting tradeoff between precision and recall can further be visualized with the help of precision and recall curves (see Figure 12). The curve for the tuned random forest model is closer to the north east section of the graph. That is the tuned model appears better in this sense than the logit model. However, at the same time it is clear that neither of the logit or tuned random forest models displays a great combination of both precision and recall.

As depicted in Appendix H, panel I the number of true positives amount to 5.62% and the number of true negatives are 66.29%. The number of false negatives are 26.97% and the number of false positives are 1.12%. Further panel II in Appendix H showcases that the percentage of female founders is the characteristic with the most weight in terms of predicting success or failure. Specifically, the percentage of female founders weighs 36.74%. The distance from Stockholm is the characteristic with the second most weight, namely at 33.09%. This is followed by founders with prior experience at a weight of 30.17%.

The three graphs in appendix I show the average partial dependence for different decision trees generated in the random forest method. The average partial

dependence is plotted against the three different characteristics used in part 2. The plots show only the magnitude of the reliance on the feature, not whether the feature is of positive or negative bearing. For example, in panel I of appendix I, three especially notable peaks in partial dependence can be observed. These peaks are at zero distance from Stockholm, at about 400 kilometers and at around 600 kilometers, distance from Stockholm. The partial dependence of percentage of female founders is the highest for organizations with mixed teams. This is reflected in panel II, appendix I, where the partial dependence value is the highest at around 50% female founders. Panel III outlines that when one or more founders with prior experience is present in the figure, the weight of this feature is nearly 12% higher, showing a more assured correlation between the feature and the success metric.

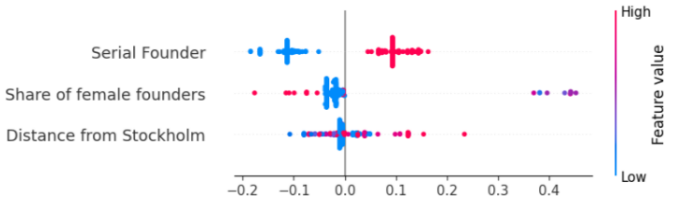
Figure 13. Mean(|SHAP value|) (average impact on model output magnitude) - Tuned RF - Part 2



Absolute mean SHAP values for random forest part 2. The larger the bar the larger the absolute mean SHAP value of the variable. The larger the absolute mean SHAP value the larger the magnitude of the contribution of a feature to the ML prediction of raising more than one round. Here class 1 represents success whereas class 0 represents failure.

SHAP values can help facilitate the understanding of different characteristics impact on the prediction of the random forest model. Investigating the SHAP values in the model is relevant to investigate the characteristics of most importance for the results generated by the model. As observed in Figure 13, founders with prior experience have the highest impact on the prediction of the random forest model, based on the mean of the absolute SHAP value. This is followed by percentage Female founders with the second highest mean SHAP value and lastly distance from Stockholm. In short this implies that on average over all observations the magnitude of the contribution is the largest for founders with prior experience and percentage of female founders.

Figure 14. SHAP value - Tuned RF- Part 2



SHAP values for random forest part 2. A blue value indicates a low value of a variable and red indicates a high value of a variable. Blue for distance from Stockholm indicates a 0 distance from Stockholm and red a high distance from Stockholm. Red indicates a value of 1 for the serial founder variable indicates a serial founder in a team and blue represents a value of 0 representing a lack of serial founders in a team. A blue value also represents a 0% share of female founders in a founding team and red a 100% share of female founders. A positive SHAP value means a positive contribution of a variable to the ML prediction of raising more than one round.

From Figure 14 it can be discerned that prior founder experience, represented by the red color on the first row, is nearly always useful for predicting success in part 2. This

is reflected by positive SHAP values for the red dots that contribute positively to the success metric on the first row. It can also be seen that in contrast lack of prior founder experience, represented by the blue dots on the same row in the figure, negatively impacts the success prediction. Further, organizations with only male teams are represented by the blue color and fully female teams are represented by the red color on the second row in Figure 14. Mixed teams are represented by purple dots. It can be seen that neither fully male nor fully female teams seem to benefit in terms of success prediction, represented by the negative SHAP values. However, the negative effect appears to be less strong for fully male teams compared to fully female teams. The positive effect for success predictions is strongest for mixed teams, represented by a positive SHAP value for the purple colored dots on the second row near the value 0.4. Mixed teams are strongly tilted positively by near 0.4 in SHAP value. Based on the third row for distance from Stockholm in Figure 14 it is difficult to draw any clear conclusions.

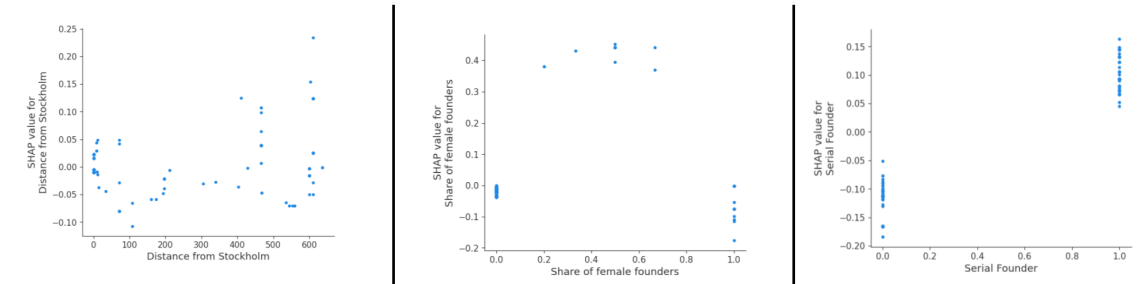
The literature by Ewens and Townsend (2020) can also be discussed relative to the results in Figure 14. In Figure 14 fully female teams contribute negatively to the success prediction in the tuned random forest model. That is it contributes negatively to the prediction that organizations raise more than one round of financing. However, the results in part two indicate that both fully female and fully male teams contribute negatively to the success prediction of the organization. Based on Figure 14 this negative effect is stronger for fully female teams compared to male teams. Although, one might try to explain this from the perspective of Ewens and Townsend's results, this should be done with caution. There are positive SHAP values for what appears to be primarily mixed teams, represented by purple dots in Figure 14 for the share of female founders characteristic.

A similar discussion as in part 1 related to literature by Gompers et al. (2010) on the importance of prior entrepreneurial experience could also be conducted here. The SHAP values for the prior founder experience variable in Figure 14 is aligned with the notion that prior entrepreneurial experience is important for future success.

A comment about the quality of the data used in part two is important in relation to the analysis of the results. The substantially smaller dataset used in our paper, compared to for example Fuster et al. (2022) is a limitation of our study. Naturally the difficulty to clearly establish conclusions about the impact of different characteristics will be influenced by the quality of the data used. It should be noted again that for instance important financial characteristics of firms are not included in the analysis in part two. Consequently, the analysis of the results in part two should be viewed as a discussion about potential explanations of the mechanisms underpinning the results. Further, for example the skewed nature of the data will influence the results. A vast majority of the observations are represented by organizations categorized as not successful. Also, as previously mentioned the dataset is highly male dominated. Therefore, this might explain why both fully male and fully female teams as a characteristic negatively influence the machine learning success predictions of the organizations. As depicted in Appendix H, panel I a substantial portion of the predictions are categorized as not successful and relatively many false negatives in the prediction negatively influences the recall of the tuned random forest model. Our results highlight the importance of thoroughly examining the input data used for predictions since it naturally impacts the success predictions. For instance, if a venture capital investor would use a ML model with skewed data without awareness of the associated risks this could negatively influence decision making. It is difficult to discuss the potential reasons behind the low number of successes in the dataset in part two because of the limited information of the data

aggregation process in part two. For instance, information about the definition of what constitutes a funding round in part 2 is missing.

Figure 15. Dependence Plots Part 2 Tuned Random Forest



Panel I Panel II Panel III

Figure 15 displays dependence plots for the different variables used in the random forest model for part 2 including gender. Positive SHAP values imply a positive contribution of a variable to the ML prediction of raising more than one round.

Figure 15, panel I further highlights the difficulty to draw conclusions about geographical clustering in part 2. The SHAP values are relatively dispersed and it is not evident that a low distance from Stockholm is associated with positive SHAP values. This is what one would expect from the perspective of for instance literature on VC clustering by Chen et al. (2010). Panel II and III confirm the above discussions related to literature by Ewens and Townsend (2020) and Kovner et al. (2010).

VII. Robustness

The Dimension of robustness has already been touched upon earlier, however will be more intensely viewed in this section. Robustness will be investigated in two manners. First in model creation and second in model and result evaluation.

Building a robust model is defined by how well the model performs when tasked with an unseen set of data. To achieve a robust model we predominantly focused on three methods, stratification, k-fold cross-validation, and hyperparameter optimization.

Stratification sampling is a technique utilized to ensure the diversity and characteristics of a dataset are well represented in any sample of said dataset (Géron, 2019). This is utilized twice in our models, initially in our split of training and test data, to make sure the proportion of successes to failures is the same. Additionally, this was conducted for the k-fold cross-validation of the training data, meaning that each fold shares the same or similar characteristics. For the cross-validation folds of the test data, stratification was omitted to mimic real life data variance.

When optimizing the hyperparameters of the model, certain restrictions were set to minimize overfitting. Namely, the number of trees has to be between 50 and 550, and have max depth between 1-15.

The model evaluations underwent several robustness checks. The first is similar to Fuster et al. (2022), where 100 bootstrapped samples are constructed from the test set and the models (logit and random forest) were applied to predict upon the same 100 different samples. Additionally, a placebo test, where the feature values are entirely random provides a method to determine whether the models “learn”. Here one wishes to see a drop in accuracy indicative of the fact that the model has learnt and is not only good at guessing. The placebo in addition to the 100 bootstrapped

samples provide ground for evaluating the stability of the model, if there is an aspect of over-fitting or underfitting specifically, and comparison of the models.

Table 9. *Random Forest Robustness Summary Overview- Part 1 Excl. Gender*

Model	Bootstrap Average Score	Placebo Accuracy	Bootstrap Variance Score
Logit	0.650476	0.587302	0.004071
Random Forest	0.73	0.460317	0.003243

Robustness summary overview for the random forest model in part 1 excluding gender

Recall that in part 1 excluding gender, the random forest model outperforms the logistic regression model across all metrics. This is evidenced by the random forest achieving higher accuracy (0.730159) and cross-validation accuracy (0.761905) over that of the logit model 's (0.650794; 0.634921). Both models exhibit learning capabilities, their accuracy differing minorly to their bootstrapped accuracy (random forest: 0.73, Logit: 0.650476) and their significantly lower placebo accuracy (Random Forest: 0.460317, Logit: 0.587302). The random forest displays even lower placebo accuracy which is indicative of a robustness to over-fitting, and reliability in capturing true signals. Additionally, both for the random forest and the logit model, the placebo score accuracies are lower than that of the naive model that only votes the majority class yield as 68.34% accuracy (the target variable is skewed: 1: 149, 0: 69). This is indicative that the models are not just enhanced guessing machines, but have, to some extent, absorbed underlying pattern data. Pursuant to the random forest's performance dominance, it also outstrips the logistic regression in terms of recall (Random Forest: 0.953488, Logit: 0.906977), precision (Random Forest: 0.742143, Logit: 0.684211), and F1 score (Random Forest: 0.828283, Logit: 0.78). These aspects accentuate random forest's eminence in part 1 excluding gender.

Table 10. Random Forest Robustness Overview- Part 1 Including Gender

Model	Bootstrap Average Score	Placebo Accuracy	Bootstrap Variance Score
Logit	0.693793	0.620690	0.003110
Random Forest	0.772069	0.482756	0.002905

Robustness summary overview for the random forest model in part 1 including gender

In part 1 including gender, the random forest model is superior to the logistic regression. The random forest has a higher accuracy (0.775862) and the same cross-validation accuracy (0.640873) compared to the logit model's (0.689655; 0.640873). Both models exhibit learning capabilities, their accuracy with only some differences to their bootstrapped accuracy (random forest :0.772069, Logit: 0.693793 and their lower placebo accuracy (Random Forest: 0.482756, Logit: 0.620690). However, it should be noted that the placebo accuracy for the logit model is relatively similar to the normal accuracy, which is not positive. The placebo accuracy for the random forest model is significantly lower. The lower placebo accuracy for the random forest indicates a robustness to over-fitting, and that the model reliably captures true signals. Additionally, both for the random forest and the logit model, the placebo score accuracies are lower than that of the naive model that only votes the majority class yield as 68.34% accuracy (the target variable is skewed: 1: 149, 0: 69). This is indicative that the models to some extent absorb underlying pattern data. The recall is the same for random forest and logistic regression here.

Random forest outperforms in terms of precision (Random Forest: 0.775510, Logit: 0.703704), and F1 score (Random Forest: 0.853933, Logit: 0.808511).

Table 11. Random Forest Robustness Overview- Part 2

Model	Bootstrap Average Score	Placebo Accuracy	Bootstrap Variance Score
Logit	0.663371	0.471910	0.002262
Random Forest	0.713146	0.685393	0.002157

Robustness summary overview for the tuned random forest model in part 2

Lastly in part 2, the random forest model is better than the logistic regression for all evaluation metrics. The random forest has a higher accuracy (0.719101) and cross-validation accuracy (0.684982) compared to the logit model 's (0.662921; 0.652015). Both models' accuracy only differ somewhat to their bootstrapped accuracy (random forest : 0.712146, Logit: 0.663371). The random forest model outperforms both in terms of precision, recall, and F1 score. The logit model has a recall, precision and F1 score equal to 0% implying that the model is only predicting failures. The fact that the data for this section is skewed in favor of failure, makes this result not entirely alien, given the simplicity of the logit model. Finally, there is the placebo score. The placebo accuracy is relatively high for the random forest model at 0.685393, although low for the logit model at 0.471910. The relatively similar placebo accuracy to the normal accuracy for the random forest model in part two accentuates the inability of the model to capture the true underlying features. This highlights the fact that more complex models are not always the better choice. Here it seems the logit model, while producing lower accuracy, has better captured the true signals, while the random forest model has picked up nonsensical links between data and predicting success. This highlights the fact that more complex model are not always the more correct models.

VIII. Conclusion

In this paper, the implementation of machine learning models in predicting the success in raising more than one funding round for Swedish startups is examined. The analysis is relevant considering the increasing usage of machine learning algorithms in venture capital screening. In part 1 financial data from Serrano is merged with data over founding teams and funding rounds from Sweden Tech Ecosystem. In part 2 data, provided by EQT, over founder characteristics and the raising of second round funding is used.

In part one excluding gender the tuned random forest model performs better than the logit model, with a higher cross validation test accuracy of 76.19% relative to logit's 65.35%. Further, the overall tradeoff between precision and recall is better for the tuned random forest model. The financial metrics have the strongest explanatory weight for the predictions of the tuned random forest model both with and without gender included. This is related to literature by for instance Kaplan et al. (2009) on founder replacement over time and the importance of the business relative to the founders for investments. Prior founder experience positively influences ML success predictions of organizations in the dataset in part one. This relates to literature by Gompers et al. (2010) and Gompers et al. (2020) on the importance of prior entrepreneurial experience for future success. For part 1, excluding gender a low distance from Stockholm slightly positively contributes to the prediction of raising more than one round. This relates to findings by Gomers et al. (2010) on geographical clustering in VC. For part 1 including gender a high share of female

founders negatively contributes to the prediction of raising more than one round. This relates to literature by Ewens and Townsend (2020) that highlights difficulties for female founders to get investor interest. A venture capitalist's reliance on the ML prediction in this paper could lead to an institutionalization of biases. The low weight of the founder characteristics and the negative contribution for a high value of the share of female founder variable could lead to a deprioritization of diversity in investment decisions and capital allocation.

In part two of the paper we find that it is difficult to accurately predict success with skewed data. The placebo accuracy in part 2 is relatively large, indicative of little significant learning in part two. There are signs of advantages in terms of success prediction for teams with serial founders on the team are found. This aligns with the importance of prior entrepreneurial experience for future success highlighted by Gompers et al. (2010). Mixed teams also contribute positively to the ML prediction in part 2. The analysis in part 2 highlights the importance of a thorough examination of the quality of data used for ML predictions.

The datasets used in this study are relatively small. Analyzing data over founder characteristics and startup success and evaluating random forest models, while insightful, are of little statistical power in determining causality. However, our paper provides contributions to the literature on decision making and the usage of machine learning algorithms in a VC context. For future research potential biases in investment selection for first round funding in a Swedish context could be investigated through an experimental approach. Furthermore, additional founder characteristics such as educational background could also be included. Moreover, it would be interesting to investigate the effects of the removal of early angel and seed rounds. Future studies can also consider the size of the rounds raised, not only the number of rounds. It is also advisable for future research to control for the volume of venture capital available during the years analyzed. Finally, in the future to contribute to the robustness of the results of a similar study additional ML models such as Xgboost could be employed similar to Fuster et al. (2022).

REFERENCES

Ahmadi, Jan, *Dagens Industri,* "Nya siffror: Kvinnor fick 0,5 procent av riskkapitalet 2021" (26 October 2022), <https://www.di.se/hallbart-naringsliv/nya-siffror-kvinnor-fick-0-5-procent-av-riskkapitalet-2021/> , retrieved 26 September 2023.

Antonio, Osval, López, Montesinos, Montesinos López, Abelardo and Crossa, José, "Multivariate Statistical Machine Learning Methods for Genomic Prediction", 2022, 111-121.

Arash Sangari, *Tillväxtverket,* "Sweden Tech Ecosystem", <https://tillvaxtverket.se/tillvaxtverket/seminarierochnatverk/natverkochsamverkan/nationellsamverkan/swedentechecosystem.1687.html>, retrieved 5 November 2023.

Bengtsson, Ola and Hsu, David H., "Ethnic matching in the US venture capital market," (March 2015), Journal of Business Venturing 30, 340.

Bernstein, Shai, Korteweg, Arthur, and Laws, Kevin, "Attracting Early-Stage Investors: Evidence from a Randomized Field Experiment," (April 2017), The Journal of Finance Vol. LXXII, No. 2, 509.

Bonelli, Maxime, "The Adoption of Artificial Intelligence by Venture Capitalists" (November 13, 2022), 1, 4, 8, 34.

Bradshaw, Tim, Financial Times, "Jawbone reaches end of the road as it goes into liquidation" (July 7, 2017), <https://www.ft.com/content/c146f144-62ad-11e7-8814-0ac7eb84e5f1>, retrieved 25 November 2023.

Breiman, Leo, 2001, Random forests, Machine Learning 45, 5–3.
Boström, Towe, *Dagens Nyheter,* "Kvinnorna osynliga för manliga kapitalister" (2022-01-15), <https://www.dn.se/ekonomi/kvinnorna-osynliga-for-manliga-kapitalister/>, retrieved 26 September 2023.

Bubna, Amit, R Das, Sanjiv and Prabhala, Nagpurnanand, "Venture Capital Communities" (March 2020), Journal of Financial and Quantitative Analysis Vol. 55, No. 2, 621-622, 632.

Caesar, Julia and Leijonhufvud, Jonas, *Dagens Industri,* "Snabb ökning av andelen kvinnliga partners i riskkapitalbolag" (31 August 2019), <https://www.di.se/digital/snabb-okning-av-andelen-kvinnliga-partners-i-riskkapitalbolag/> , retrieved 19 November 2023.

Cao, Sean, Jiang, Wei, Wang, Junbo and Yang, Baozhong, "From Man vs. Machine to Man + Machine: The Art and AI of Stock Analyses" (June 2022), Columbia Business School Research Paper.

Chen, Henry, Gompers, Paul, Kovner, Anna, Lerner, Josh, "Buy local? The geography of venture capital" (2010), Journal of Urban Economics 67, 90-91.

Cremades, Alejandro, *Forbes,* "The Pros And Cons Of Bootstrapping Startups" (Jan 13, 2019), <https://www.forbes.com/sites/alejandrocremades/2019/01/13/the-pros-and-cons-of-bootstrapping-startups/?sh=24a51688273d>, retrieved 3 December 2023.

Currarini, Sergio, Jackson, Matthew O. and Pin, Paolo, "An economic model of friendship: Homophily, minorities, and segregation" (July 2009), Econometrica Vol 77, No.4, 1003-1004.

Davila, Antonio and Foster George, "Management Control Systems in Early-Stage Startup Companies", The Accounting Review , Vol. 82, No. 4 (Jul., 2007), 910, 919-921.

Ewens, Michael and Townsend, Richard R., "Are early stage investors biased against women?" (March 2020), Journal of Financial Economics 135, 653-654, 656, 658-659, 661, 665, 672, 676.

Fuster, Andreas, Goldsmith-Pinkham, Paul, Ramadorai, Tarun and Walther, Ansgar, "Predictably Unequal? The Effects of Machine Learning on Credit Markets" (February 2022), The Journal of Finance Vol. LXXVII, No.1, 5-10, 17-26, 32, 42-43.

Géron, Aurélien, "Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow - Concepts, Tools, and Techniques to Build Intelligent Systems", 2019, 56.

Gompers, Paul, Kovner, Anna, Lerner, Josh, Scharfstein, David, "Performance persistence in entrepreneurship" (April 2010), Journal of Financial Economics 96, 18.

Gompers, Paul A., "Optimal Investment, Monitoring, and the Staging of Venture Capital" (1995), The Journal of Finance Vol 50, No. 5, 1463-1467.

Gompers, Paul A., Mukharlyamov, Vladimir and Xuan, Yuhai, "The cost of friendship" (March 2016), Journal of Financial Economics 119, 627, 634.

Gompers, Paul A., Gornall, Will, Kaplan, Steven N. and Strebulaev, Ilya A., "How do venture capitalists make decisions?" (January 2020), Journal of Financial Economics 135, 170, 177.

Government Offices of Sweden, "Nytt uppdrag till Tillväxtverket ska stärka förutsättningarna för kvinnors företagande" (17 March 2023), <https://www.regeringen.se/pressmeddelanden/2023/03/nytt-uppdrag-till-tillvaxtverket-ska-starka-forutsattningarna-for-kvinnors-foretagande>, retrieved 16 September 2023.

Griffin, John M., Hirschey, Nicholas and Kruger, Samuel, "Do Municipal Bond Dealers Give Their Customers "Fair and Reasonable" Pricing?" (April 2023), The Journal of Finance, Vol. LXXVII, No.2, 887, 919-920.

Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome, The Elements of Statistical Learning Data Mining, Inference and Prediction (2009), 119-128, 309, 587-593.

Hellmann, Thomas and Thiele, Veikko, "Friends or foes? The interrelationship between angel and venture capital markets" (2015), Journal of Financial Economics 115, 640, 642.

Ho, Tin Kam, 1998, The random subspace method for constructing decision forests, IEEE Transactions on Pattern Analysis and Machine Intelligence 20, 832–844.

Hodgson, Leah, *PitchBook,* "Sweden leads, totals slip: Nordic VC trends for 2022 in 6 charts" (March 8 2023), <https://pitchbook.com/news/articles/nordic-vc-deals-exits-fundraising>, retrieved 1 October 2023.

Howell, Sabrina T., "Reducing information frictions in venture capital: The role of new venture competitions" (2020), *Journal of Financial Economics 136*, 676 -679, 694.

Härd, Sverker, *Tillväxtanalys,* "Riskkapitalstatistik 2020 - venture capital" (January, 2022), <https://www.tillvaxtanalys.se/download/18.3e9519917ddd5d304b70e78/1643310005214/Statistik_2022_01_Riskkapitalstatistik_2020_venture_capital.pdf>, 11.

James, Gareth, Witten, Daniela, Hastie, Trevor, Tibshirani, Robert and Taylor, Jonathan, An Introduction to Statistical Learning with Applications in Python (2023), 154-155, 331, 337-339, 343, 345-346, 382-390, 492.

Johansson, Ida, *Göteborgs-Posten,* "Fem procent av riskkapitalet hamnar i Göteborgsbolag – "skrämmande siffror"" (16 Febuary 2020), <https://www.gp.se/ekonomi/fem-procent-av-riskkapitalet-hamnar-i-g%C3%B6teborgsbolag-skr%C3%A4mmande-siffror-1.24089400>, retrieved 26 September 2023.

Kaplan, Steven N., Sensoy, Berk A. and Strömberg, Per, "Should Investors Bet on the Jockey or the Horse? Evidence from the Evolution of Firms from Early Business Plans to Public Companies" (23 January 2009), The Journal of Finance Vol. LXIV, No 1, 75, 112-113.

Kaplan, Steven N. and Sorensen, Morten, "Are CEOs different?" (March 9 2021), The Journal of Finance Vol. LXXVI, No.4, 1773.

Kessler, Judd B., Low, Corinne, and Sullivan, Colin D., "Incentivized Resume Rating: Eliciting Employer Preferences without Deception", (November 2019), The American Economic Review 109, 3739-3740.

Kruppa, Miles and Lee, Dave, *Financial Times,* "Airbnb raises $1bn from new investors", <https://www.ft.com/content/0dff2d66-ca0f-4257-8a3d-867e6aa6544b>, retrieved 24 November 2023.

Lundberg, Scott M., Erion, Gabriel, Chen, Hugh, DeGrave, Alex, Prutkin, Jordan M., Nair, Bala, Katz, Ronit, Himmelfarb, Jonathan, Bansal, Nisha, and Lee, Su-In, "From local explanations to global understanding with explainable AI for trees" (Jan 2, 2020), *Nature Machine Intelligence*
Scott M. Lundberg and Su-In Lee, "A Unified Approach to Interpreting Model Predictions", 2017.

M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions.", Journal of the Royal Statistical Society. Series B (Methodological), vol. 36, no. 2, 1974.

Olsson Jeffery, Miriam, *Di Digital,* "Miljarderna fortsätter rulla in till män – dödläge för kvinnors bolag" (24 July 2021), <https://www.di.se/digital/miljarderna-fortsatter-rulla-in-till-man-dodlage-for-kvinnors-bolag/>, retrieved 26 September 2023.

Omni, "Majoriteten riskkapital till Stockholm: "Skrämmande" " (16 Febuary 2020), <https://omni.se/majoriteten-riskkapital-till-stockholm-skrammande/a/y3rgBx>, retrieved 26 September 2023.

Optuna, “Optimize your Optimization”, <https://optuna.org/?fbclid=IwAR1ykIFE-rdfwX9YQIODd-66t7qMMeiOc81uCpRJJ9pEuwIkY2kz7ZuglA8>, retrieved 29 October.

Phelps, Edmund S., “The Statistical Theory of Racism and Sexism” (Sep 1972), The American Economic Review Vol. 62, No.4, 659-661.

Puri, Manju and Zarutskie, Rebecca, “On the Life Cycle Dynamics of Venture-Capital and Non-Venture-Capital-Financed Firms” (December, 2012), The Journal of Finance Vol. LXVII, No 6., 2248, 2267.

Rajan, Raghuram G, "Presidential Address: The Corporation in Finance” (August 2012), The Journal of Finance Vol. LXVII, No. 4, 1173-1176.

Stockholm Chamber of Commerce, “Därför är riskkapitalbolagen avgörande för Stockholms startupscen” (3 April 2023), <https://stockholmshandelskammare.se/nyheter/darfor-ar-riskkapitalbolagen-avgorande-stockholms-startupscen>, retrived 16 September 2023.

SVT, “Entreprenören: Här är fyra hinder för kvinnligt företagande” (16 August 2023), <https://www.svt.se/nyheter/inrikes/entreprenoren-darfor-startar-farre-kvinnor-an-man-foretag>, retreived 16 September 2023.

Sweden Tech Ecosystem, “Dashboard”, *(2023),* <https://techecosystem.startupsweden.com/dashboard/f/geo/anyof_V%C3%A4stra%20G%C3%B6taland>, retrieved 2-5 November 2023.

Sweden Tech Ecosystem, “Funding Rounds”, *(2023),* <https://sweden.dealroom.co/transactions.rounds/f/growth_stages/anyof_late%20growth_early%20growth_seed_not_mature/launch_year_max/anyof_2012/launch_year_min/anyof_2010/rounds/not_GRANT_SPAC%20PRIVATE%20PLACEMENT/slug_locations/allof_sweden/tags/not_outside%20tech>, retrieved 2-5 November 2023.

Sweden Tech Ecosystem, “Funding Rounds”, *(2023),* <https://sweden.dealroom.co/transactions.rounds/f/founders_is_serial_founder/anyof_yes/growth_stages/anyof_late%20growth_early%20growth_seed_not_mature/launch_year_max/anyof_2012/launch_year_min/anyof_2010/rounds/not_GRANT_SPAC%20PRIVATE%20PLACEMENT/slug_locations/allof_sweden/tags/not_outside%20tech>, retrieved 2-5 November 2023.

Swedish Private Equity & Venture Capital Association (SVCA), “The Economic Footprint of Swedish Venture Capital and Private Equity” (November 2022), <https://www.svca.se/wp-content/uploads/2022/11/The_Economic_Footprint_of_Swedish_VC_and_PE_Final-1.pdf>, retrieved 1 October 2023.

The Nobel Prize, “The Prize in Economic Sciences 2012”, <https://www.nobelprize.org/prizes/economic-sciences/2012/popular-information/>, retrieved 5 November 2023.

Tian, Xuan, “The causes and consequences of venture capital stage financing” (2011), Journal of Financial Economics 101, 132-135, 140.

Vinnova, “New digital platform will strengthen Swedish tech companies” (12 October, 2021), <https://www.vinnova.se/en/news/2021/10/new-digital-platform-will-strengthen-swedish-tech-companies/>, retrieved 24 November 2023.

Wallenberg, Björn and Caesar, Julia, *Dagens Industri,* “Kartläggning: Stockholm slukar tillväxtpengarna – resten halkar efter: ”Skrämmande” “ (17 Febuary 2020), <https://www.di.se/nyheter/kartlaggning-stockholm-slukar-tillvaxtpengarna-resten-halkar-efter-skrammande/>, retrieved 26 September 2023.

Weidenman, Per, “The Serrano Database for Analysis and Register-Based Statistics.” at Swedish House of Finance Research Data Center. Accessed 11 03, 2023. https://www.hhs.se/en/houseoffinance/data-center/

Zhang, Ye, “Discrimination in the Venture Capital Industry: Evidence from Field Experiments” (April 2, 2023), Stockholm School of Economics, 28-30.

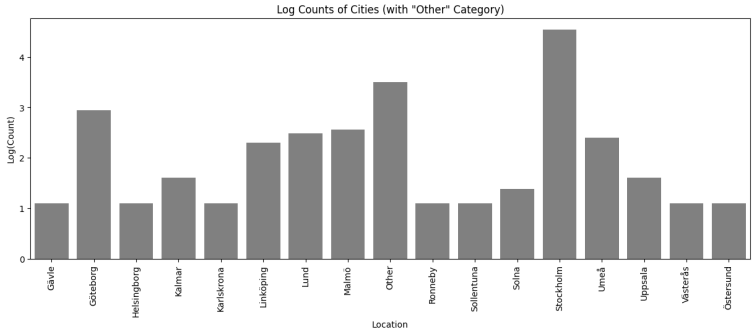
Appendix

Appendix A. Location Data - Part 1 Excluding Gender

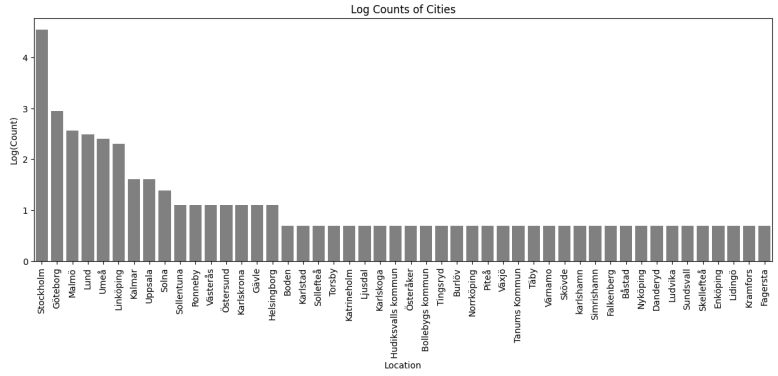
Location	Count	Percentage
Stockholm	93	44.285714
Göteborg	18	8.571429
Malmö	12	5.714286
Lund	11	5.238095
Umeå	10	4.761905
Linköping	9	4.285714
Kalmar	4	1.904762
Uppsala	4	1.904762
Solna	3	1.428571
Sollentuna	2	0.952381
Ronneby	2	0.952381
Västerås	2	0.952381
Östersund	2	0.952381
Karlskrona	2	0.952381
Gävle	2	0.952381
Helsingborg	2	0.952381
Boden	1	0.47619
Karlstad	1	0.47619
Sollefteå	1	0.47619
Torsby	1	0.47619
Katrineholm	1	0.47619
Ljusdal	1	0.47619
Karlskoga	1	0.47619
Hudiksvalls kommun	1	0.47619
Osteråker	1	0.47619
Bollebygds kommun	1	0.47619
Tingsryd	1	0.47619
Burlöv	1	0.47619
Norrköping	1	0.47619
Piteå	1	0.47619
Växjö	1	0.47619
Tanums Kommun	1	0.47619

Täby	1	0.47619
Värnamo	1	0.47619
Skövde	1	0.47619
Karlshamn	1	0.47619
Simrishamn	1	0.47619
Falkenberg	1	0.47619
Båstad	1	0.47619
Nyköping	1	0.47619
Danderyd	1	0.47619
Ludvika	1	0.47619
Sundsvall	1	0.47619
Skellefteå	1	0.47619
Enköping	1	0.47619
Lidingö	1	0.47619
Kramfors	1	0.47619
Fagersta	1	0.47619

Panel I- Counts of Cities With "Other" Category- Part 1 Excluding Gender



Panel II- Log(Count) For Cities With "Other" Category- Part 1 Excluding Gender



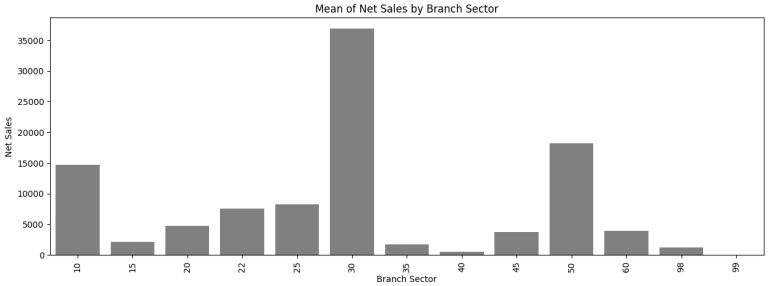
Panel III- Log(Count) For Cities- Part 1 Excluding Gender

Panel I-III shows the count for different cities for part 1 data excluding gender. The data is dominated by organizations located in Stockholm.

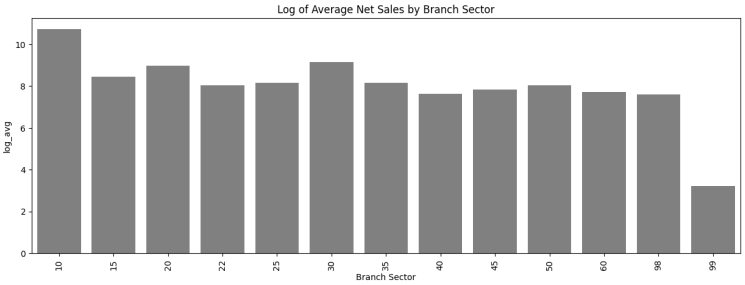
Appendix B. Sales Data - Part 1 Excluding Gender

10 = Energy & Environment
15 = Materials
20 = Industrial goods
22 = Construction industry
25 = Shopping goods
30 = Convenience goods
35 = Health & Education
40 = Finance & Real estate
45 = IT & Electronics
50 = Telecom & Media
60 = Corporate Services
98 = Other
99 = Missing

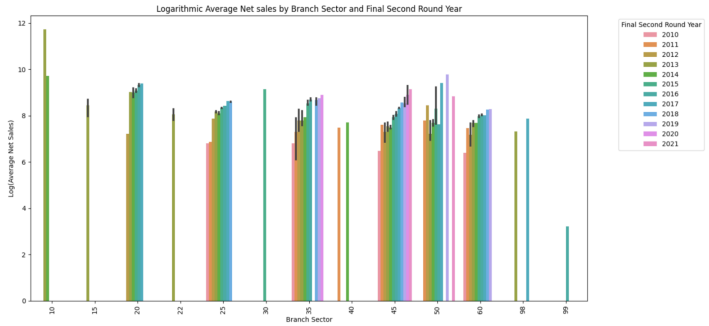
Panel I: Definition Branch Sectors



Panel II: Mean of Net Sales by Branch Sector



Panel III: Log of Average Net Sales by Branch Sector



Panel IV: Logarithmic Average Net Sales by Branch Sector And Final Second Round Year

Panel I-IV displays sales measures for the organizations in the dataset in part 1, excluding gender, by branch sector.

Appendix C- Derivation SHAP Formula

Property 1: Local accuracy- or additivity in game theory- assert that the sum of individual contributions of all features to the prediction should equal the actual prediction of the model for that specific instance (Lundberg et al, 2020):

$$f(x) = \phi_0(f) + \sum_{i=1}^{M} \phi_i(f, x)$$

Where f is the machine learning model that maps input data to an output. x is the input or instance. f(x) is the output of the machine learning model for instance x. Furthermore, $\phi_o(f) = E[f(z)]$, is the models prediction when without input denoted z. $\phi_i(f,x)$ is the attribution of the ith feature in the input x towards the model's output. M is the total number of features.

This ensures that SHAP values provide a consistent and complete decomposition of the prediction, as well as an accurate and fair contribution of each feature to the prediction.

Property 2: consistency- monotonicity in game theory- explicates that if the contribution of a feature (the difference in models output with and without the feature) to the prediction increase or remains the same then its attribution (how much feature i is responsible for the difference between the actual and expected prediction = $\phi_i(f, x)$) should not decrease. Mathematically, for any differing models f' and f this is represented as (Lundberg et al, 2020):

$$f'_x(S) - f'_x(S \setminus i) \geq f_x(S) - f_x(S \setminus i)$$

Here S is a subset of all features M, with S\i being the subset not including feature i.

Property 3: Missingness, which is referred to as null effects in game theory implies that if a feature i has no change on the function output ($\phi_i(f, x) = 0$), then that feature is considered to have impact score of 0. Mathematically this is (Lundberg et al, 2020):

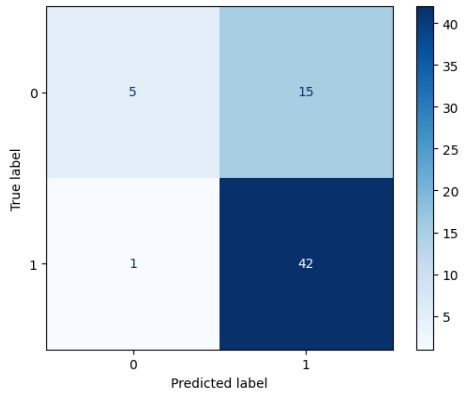
$$f_x(S \cup i) = f_x(S)$$

Culminating these properties results in the Game theory Shapley value formula presented by Lloyd Shapley . To apply Shapley's theory to machine learning Lundberg and Lee (2017) utilized Lloyd's Formula and postulated SHAP (Shapley additive) Values. In brevity SHAP values are the Lloyd Shapley values of a conditional expectation function of the random forest model - in our case. The final SHAP equation is:

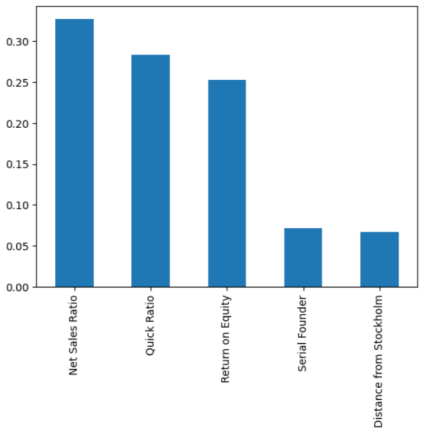
$$\phi_i(f, x) = \sum_{S \subseteq M \setminus \{i\}} \frac{|S|!(|M| - |S| - 1)!}{|M|!} [f_x(S \cup \{i\}) - f_x(S)]$$ [31]

[31] The formula notation has been slightly altered from (Lundberg and Lee, 2017), to be consistent with the paper's form.

Appendix D- Tuned RF Model- Part 1 Excl Gender



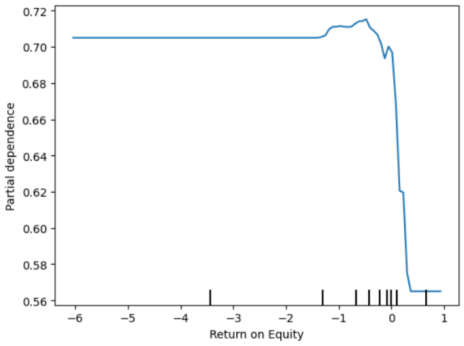
Panel I



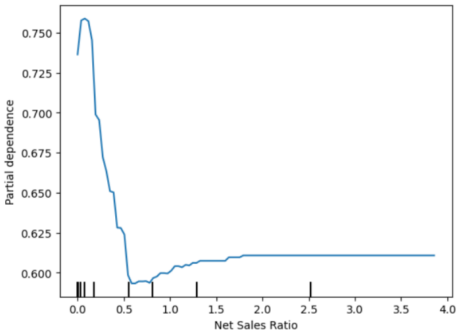
Panel II

Panel I showcases the confusion matrix for the ML prediction on the test dataset for the random forest model in part 1, excluding gender. Panel I showcases that the vast majority of observations in the test set are classified as successes. Panel II shows that net sales has the highest explanatory weight for the ML prediction for part 1 excluding gender. Distance from Stockholm in contrast has the lowest explanatory weight.

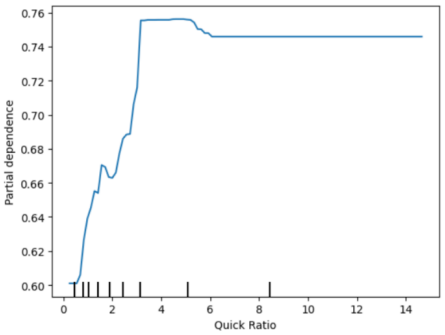
Appendix E- Partial Dependence Plots Part 1 Excl. Gender



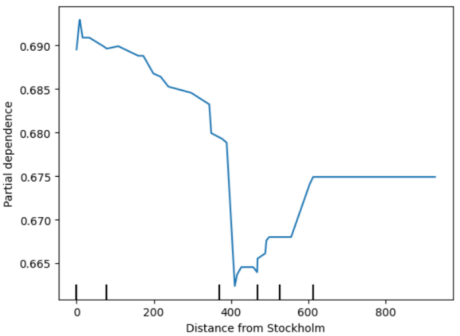
Panel I



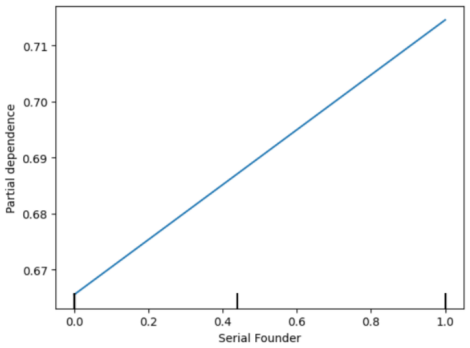
Panel II



Panel III



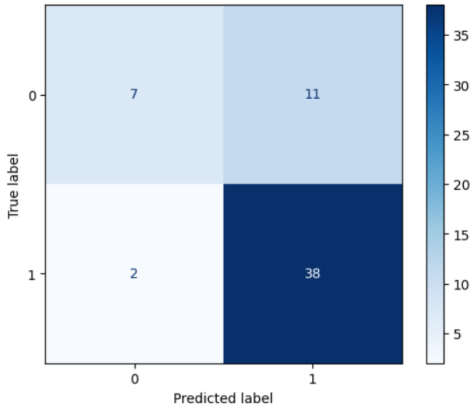
Panel IV



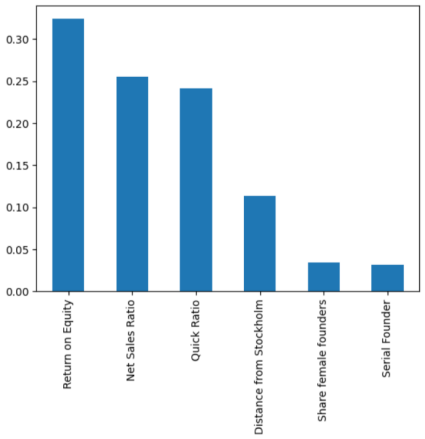
Panel V

Partial dependence plots for random forest model part 1, excluding gender. The partial dependence represents a value of the degree to which the random forest model depends on a variable for the ML prediction of raising more than one round. Panel I showcase the partial dependence for ROE for example.

Appendix F- Tuned Random Forest Model Part 1 Including Gender



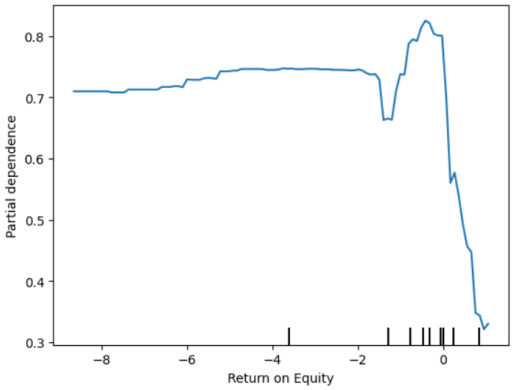
Panel I.



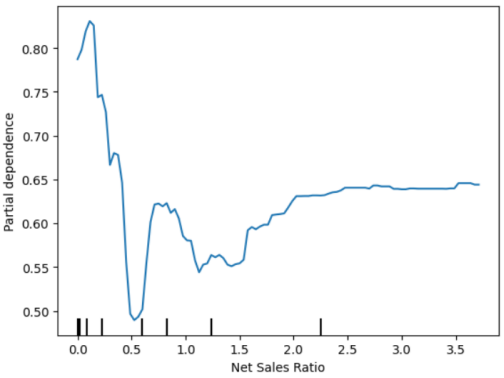
Panel II.

Panel I showcases the confusion matrix for the ML prediction on the test dataset for the random forest model in part 1, including gender. Panel I showcases that the vast majority of observations in the test set are classified as successes. Panel II shows that ROE has the highest explanatory weight for the ML prediction for part 1 including gender. Serial founder in contrast has the lowest explanatory weight.

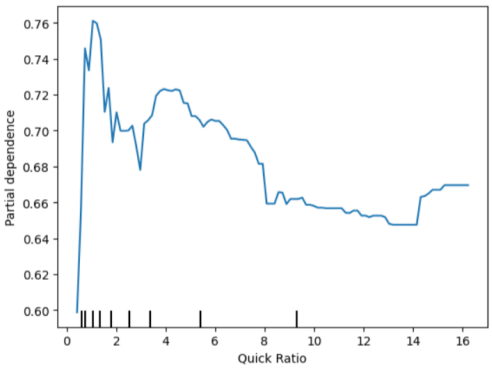
Appendix G- Partial Dependence Plots Part 1 Including Gender



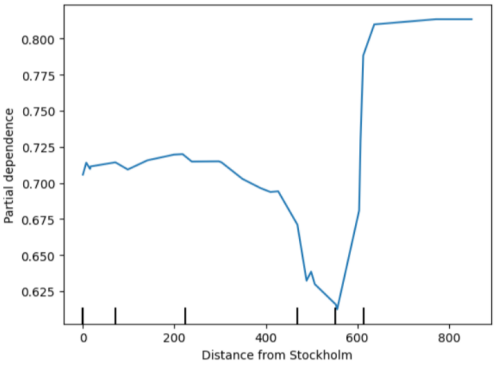
Panel I: ROE



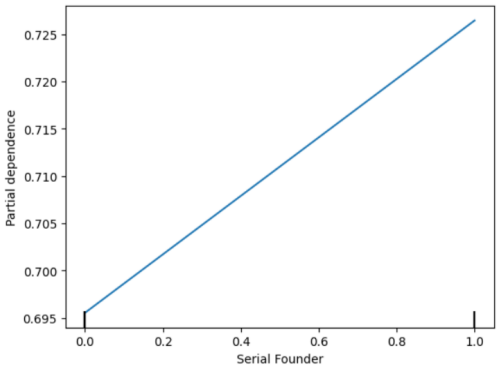
Panel II: Net Sales Ratio



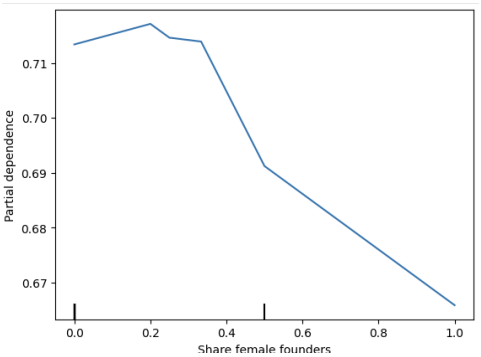
Panel III: Quick Ratio



Panel IV: Distance From Stockholm



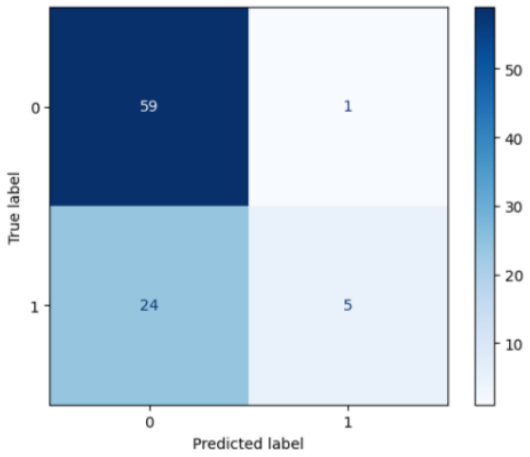
Panel V: Prior Founder Experience



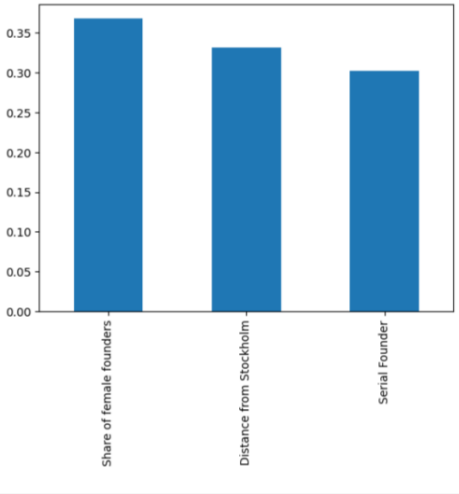
Panel VI: Share of Female Founders

Partial dependence plots for random forest model part 1, including gender. The partial dependence represents a value of the degree to which the random forest model depends on a variable for the ML prediction of raising more than one round. Panel I showcase the partial dependence for ROE for example.

Appendix H. Results Tuned Random Forest Model- Part 2



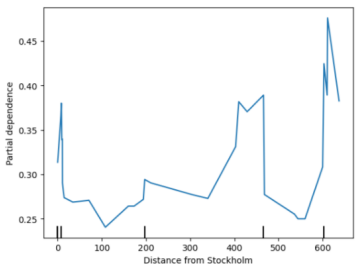
Panel I: Confusion Matrix, of true success and predicted success of tuned rf model.



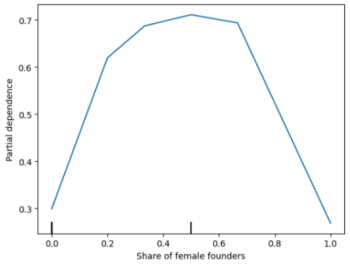
Panel II: Explanatory Weights - Tuned Random Forest

Panel I showcases the confusion matrix for the ML prediction on the test dataset for the random forest model in part 2. Panel I showcases that the vast majority of observations in the test set are classified as not successes. Panel II shows that Share of female founders has the highest explanatory weight for the ML prediction for part 2. Serial founder in contrast has the lowest explanatory weight.

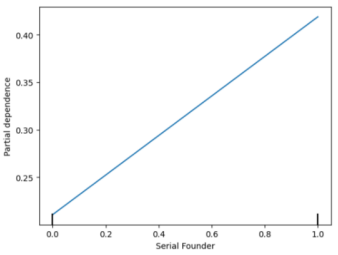
Appendix I. *Partial Dependence Plots Part 2*



Panel I



Panel II



Panel III

Partial dependence plots for random forest model part 2. The partial dependence represents a value of the degree to which the random forest model depends on a variable for the ML prediction of raising more than one round. Panel I showcase the partial dependence for distance from Stockholm for example.