UNIVERSITY OF COPENHAGEN
FACULTY OF HUMANITIES, DEPARTMENT OF NORDIC STUDIES AND LINGUISTICS
MSC IT AND COGNITION

ENTREPRENEURIAL FINANCING AND MACHINE LEARNING: CAN WE MITIGATE GENDER BIAS?

Master Thesis

Author: Marlene Dahle
Student ID: zfk742

Supervisor: Nora Hollenstein

Submitted on: 31 May 2022

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 ABSTRACT

Machine learning models are making their way into decision-making processes at a fast pace in a range of fields. Although machines are inherently thought of as unbiased and objective when making decisions, there is an increasing awareness that this is not the case. Machine learning models learn on historic data, which is prone to a range of human decision-making biases. One type of decision-making process that is highly prone to gender bias, is the field of early-stage investment decision-making. Less than 1% of venture capital are awarded to women, investors are mostly men, and the entrepreneurial stereotype is largely male. There is also extensive research conducted showing that there is gender bias in the investment decision-making process. When historic data with these limitations are used to build machine learning models, the models are likely to be as biased as their human decision-making counterparts and could potentially amplify this bias.

To the best of the author's knowledge, no work has been conducted to investigate fairness in machine learning models in the context of entrepreneurial financing. This paper seeks to make a first contribution to this field, by investigating gender bias mitigation strategies in machine learning models made for early-stage decision making assistance. As no research has been done in this exact domain, this paper will give a theoretical background from three angles. First, investigating the current attempts at creating machine learning models in early-stage investment decision-making, second understanding in depth the gender bias that exist when making investment-decisions and third investigating broadly the causes and occurrences of gender bias in machine learning.

The most important findings and contributions of this paper are:

- Machine learning models built on data from the commonly used Crunchbase data show gender bias on all measured fairness metrics.
- Fairness through unawareness (excluding information about gender to the model) does not make a difference to the model's gender bias.
- The original data is fair by two metrics, while the model using no gender bias mitigation techniques reduce fairness on these two metrics.
- Techniques such as over- and under-sampling of either protected attribute or of outcome class can significantly increase gender bias measured by fairness metrics.
- Machine learning mitigation tools cannot be viewed or applied in isolation. Gender bias in this field should be addressed from a range of different angles.

2 INTRODUCTION

2.1 MOTIVATION

The creation of new ventures is a core part of society by solving problems, increasing quality of life, contributing to economic progress, and creating jobs. Funding at an early stage is imperative for the survival of most start-up ventures. By controlling funding resources, investors hold the power to decide which ventures gets to succeed and which does not.

Within early-stage investment, such as VC investment, only 0.7% of funding was in the Nordic countries in 2020 awarded to startups with only female founders. For founders with mixed genders the number is 7.3%, and the remaining 92% was awarded to all-male founders (Bavey et al., 2021). The reasons for this are complex, and some of them will be investigated in this paper, but the fact of the outcome remains. This has been a focus in both research and for many investors, including governmental investment institutions, but continues to be a problem.

The financial industry, as many other industries, are increasingly adopting machine learning methods to help decision making. Investors and VC companies are no exception, and it is expected that the use of machine learning models in early-stage investment decisions will increase significantly in the coming years (Gartner, 2019). There is optimism that the introduction of algorithms in the investment domain can help make better decisions about investments, identifying the very best start-ups based on historic data, at the same time as making decision making more neutral and freer from human bias. A machine has no prejudice or reasons to be sexist, is the underlying assumption. This optimism stems in a common misunderstanding about machine learning models being objective. Any machine learning model is based on data, and that data is usually prone to different types of human biases. Machine learning models will learn from data, find patterns, and produce outcomes based on what it finds. When machine learning models are created on data where the outcomes are so heavily skewed, the machine learning models will reproduce these results, possibly keeping the proportion of funding to female founders low. Looking at some of the research trying to create machine learning models for investment decisions, suggested models even use gender of the founder(s) as a variable.

The risk of female founders being even more disadvantaged in start-up funding in the future thus increases with the introduction of machine learning to the field. To keep this from

happening, there is much work that needs to be done, some being around awareness and knowledge about both gender bias and machine learning, as well as technical tools for mitigating the bias stemming from the data, and tools to teach human decision makers about the reasoning behind the decisions made by algorithms. This paper seeks to contribute to this work, with a special focus on machine learning tools for bias mitigation.

2.2 PURPOSE AND RESEARCH QUESTION

Funding at an early stage is important for the survival of a new venture (Croce et al., 2012; Kortum & Lerner, 2000). Gender bias is found in early-stage investment decisions (see section 3.2). From a report made by Unconventional Ventures (Bavey et al., 2021), less than 1% of VC funding in Scandinavia went to all female founder teams.

Gender bias within entrepreneurial financing decisions does not only keep female founders from getting funded, but also keep investors from investing in the best possible business ventures. Evaluating startups, and especially early-stage startups, can be largely subjective and unpredictable (Bai & Zhao, 2021). The strive to identify and attract the best startups is increasingly competitive as the number of VC companies increase. The earlier a VC company can identify which startups to invest money in, the less they need to pay for equity, and the more precise they are in the process of choosing successful startups, the more return on their investment they can expect. However, identifying which companies are successful at such an early state is no easy feat. There is little information available about the company itself, making the investors reliant on information about the founders and their own gut feel.

We can expect that AI will be making its way into investment decision making at a fast pace (Gartner, 2021; Ferrati & Muffatto, 2021). Although AI can seem like something that could prevent gender bias in early-stage investments – after all, a machine has no reason to be sexist – there are some major pitfalls to be aware of. A ML model is only as good as the data it is trained on – biased data will result in a biased model. Further, the risk of gender bias might be even larger when investors use ML models, precisely because they believe the machine to not have any biases, but to be 'objective'. However, ML models are not objective, and this is important to acknowledge and work to find ways to mitigate the occurring biases (Hooker, 2021; Smith & Rustagi, 2021). ML models can amplify existing biases from data, and even if the data is unbiased, design choices can lead to an algorithm displaying biased behavior. Additionally, when biased models are used by humans, it can

lead to a feedback loop of even more bias (Mehrabi et al., 2019). Leavy (2018) warns that "It would be unfortunate to have to wait until gender biased machine learning algorithms repeat the injustices of the past before action preventing gender bias is taken".

An important question emerges: Is it not enough to use the gender bias mitigation methods used in other types of ML algorithms, and apply them to ML algorithms for early-stage investment decision-making? Motivations and priorities will differ from industry to industry. It is important to find customize tools and strategies to fit each context, as well as being designing for real and not just imaginary situations (Holstein et al., 2019; Veale & Binns, 2017). Additionally, the appropriate definitions of fairness are contextual, and must be agreed on in the specific domain. Testing and creating guidelines for this specific field can assist developers and decision-makers and make it more likely that these types of mitigation tools will be utilized. Lastly, there is a need to understand the whole picture of gender bias, where it is coming from and the various ways to combat it. The machine learning tools should not be viewed in isolation.

The purpose of this research is to find solutions to mitigate gender bias in machine learning models for investment-decision making.
The research question (RQ) is

> *Which gender bias mitigation strategies are most effective for machine learning models created to assist in early-stage investment decision-making?*

2.3 SCOPE AND LIMITATIONS

The aim in this paper is not to create the best possible success predictions of startups, but rather to test how different bias mitigation strategies work. The data collection and model building are nonetheless aimed at being as similar as possible to other recent models that are described in section 3.1.2, however, not to be considered 'state-of-the-art' in terms of being a decision-making model.
Machine learning algorithms are commonly grouped in three main categories: supervised learning, unsupervised learning, and semi-supervised learning. The focus in this paper will be on supervised learning, where the target variable is specified and known for training data.

The focus will be on venture capital funding at an early stage, with the research focusing on the first investment round, except crowdfunding. While crowdfunding is an increasingly used source of fundraising, it is not included as part of this paper. The reason is that crowdfunding investors are often more of a customer or donor, rather than an investor (Ewens & Townsend, 2020), and amounts are low, in a study by Johnson et al. (2018) averaging at $50 per contribution, and often including some goods or services.

The paper will only consider metrics and methods for data where the gender of the founders is known. In many cases, gender will be an unknown variable, in which case neither gender bias (by fairness metrics) or mitigation techniques can be used.

Further, the paper will focus on data where the protected attribute of gender is known. The gender bias will in the machine learning context be measured by fairness metrics presented in section 4.

2.4 TERMS

All mentions of "investment decisions" or "entrepreneurial financing" will refer to early-stage investment decisions, primarily angel investments and venture capital investments.

The definition of gender bias from Ntoutsi et al. (2019) will be used: "the inclination or prejudice of a decision made by an AI system which is for or against one person or group, especially in a way considered to be unfair". Feldman and Peake (2021) state that a machine learning model is considered biased if it produces discriminatory results based on sensitive attributes such as gender. This 'occurs when a selection process has widely different outcomes for different groups, even as it appears to be neutral' (Feldman et al., 2015). This notion fits with the definitions of disparate impact as a fairness metric, described further in section 4.2.1.

When referring to fairness, a broad definition will be used: "The absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making" (Saxena et al., 2019). Bellamy et al. (2019) gives the following definition to connect bias and fairness notions: "*Bias is a systematic error. In the context of fairness, we are concerned with unwanted bias that places privileged groups at a systematic advantage and unprivileged groups at a systematic disadvantage".*

3 THEORETICAL BACKGROUND

While gender bias is not a new topic within machine learning, to the best of the author's knowledge, no research is found in the specific domain of mitigation of gender bias in machine learning models for early-stage decision making. There are commercial applications/companies that work on this problem and sell solutions to this, such as VenturePole[1]. This paper will thus triangulate the topic.

Relevant theoretical work comes from three different domains: (1) use of ML models in investment decision-making, (2) gender bias in early-stage investments, and (3) gender bias and fairness in general in ML models. These three domains will be investigated to triangulate the research questions. The first part will investigate how investment decisions are made and some attempt from researchers to create machine learning models for investment decision-making. The second part will give a deeper understanding of gender bias in investment decision-making, and its consequences. The third will an overview of how gender bias in human decision-making processes is transferred into machine learning models.

The three main assumptions that stem from the theoretical background and that is essential for this paper to be relevant is: 1) that machine learning models increasingly be used in making investment decisions at an early stage, 2) that there is gender bias found in the (human) decision-making process when making funding decisions for early-stage investments, and 3) that the machine learning models created for this purpose will exhibit and possibly amplify the gender bias.

3.1 MACHINE LEARNING IN EARLY-STAGE INVESTMENT

The first domain to be investigated, is machine learning in early-stage investments. According to Obschonka and Audretsch (2019), AI and big data represent a disruptive potential in the field of entrepreneurship research but has not received much attention yet. They suggest entrepreneurial finance as one of the potential research areas. Ferrati and Muffato (2021) gives an overview of machine learning research that have used Crunchbase data to make prediction on investment decisions. They note that machine learning techniques in the entrepreneurial financing context is still new, and not many studies are

[1] https://venturepole.com

available. They argue that it will have a “game-changing impact on the traditional venture capital industry”, as it has had in many other areas of the financial sector such as stock exchanges and insurance (Ferrati & Muffato, 2021).

3.1.1 HOW ARE INVESTMENT DECISIONS MADE?

According to Krishna et al. (2016) early-stage investors focus on the founding team. Very little information is known about the company, as the companies often have just started operating and need capital to grow. The decision-making process from the investor perspective is based on a combination of assessment criteria and their own experience. It is associated with high risk as the number of successes are low; most startups fail. Additionally, decisions are “characterized by a high level of uncertainty and information asymmetry” such as a lack of data about each company. This leads to decisions being made with a high reliance on human judgment and experience, with little consensus as to which specific variables are the most relevant (Ferrati & Muffato, 2021). Huang and Zhan (2015) find that founders that have previously been employed by what they define as prominent companies are more likely to obtain funding at an early stage from high quality investors. The effect is larger if the startup is in a similar type of business as the one in which they were previously employed.

Gompers et al. (2020) investigate how VC investors make investment decisions through a survey of 885 VCs at 681 firms. According to the survey results, VC investors see the management team as the most important aspect, followed by business-related aspects such as the product the startup is developing or the technology they are using. The survey also show that VC investors, when evaluating the cause of the success or failure of a startup, attribute this to the team more than the business itself.

Bao & Zhao (2021) created six different machine learning models to see if they could predict the investment decisions of a VC firm. They found that the two most prominent feature categories for predictions were qualitative categories of criteria, namely “team management” and “planning strategy”.

3.1.2 PREVIOUS ATTEMPTS: MACHINE LEARNING MODELS FOR EARLY-STAGE INVESTMENT DECISION-MAKING

In a review, Ferrati & Muffato (2021) find a total of twenty papers investigating machine learning models within entrepreneurial financing. Six of the studies have made predictions of an exit event of a company (being acquired or gone public – ie 'acquired' or 'ipo' as the success variable) using Crunchbase data, that will be briefly described below. They report that the most used models in the papers investigated are Random Forest (9 papers), Logistic Regression Classifier (7 papers) and Support Vector Machine (7 papers).

Arroyo et al. (2019) uses a dataset from Crunchbase of more than 120,000 companies and builds a model to predict their progress in a 3-year period. Five models are tested: Decision Tree (DT) Random Forest (RF), Gradient Tree Boosting (GBT), Extremely Randomized Trees (ERT), and Support Vector Machines (SVM). The first three are all tree-based models that are white-box classifiers making it easier to understand the decisions and success drivers. The latter two are black-box classifiers. The metrics used to calculate performance are precision, recall and F1 score. They create 105 variables for each company, in the following categories: company information, funding information, and founders' information. Within the latter category, they use sensitive variables such as number of male and female founders in the training data. The count of male founders is listed as the 4th and 7th most important variable. The best performer was the GBT model with an accuracy of 82.2%.

In a conference paper, Xiang et al. (2012) uses data from Crunchbase combined with news articles related to the startups from Techcrunch to predict startup success, using Bayesian Network. They find that funding-related variables are important in predicting startup success, here measured by whether it has been acquired (not including IPO). In another conference paper Huang (2016) use network analysis with SVM, Adaptive Boosting (AB) and RF, from Crunchbase data. Investigating feature importance, they find that four out of the five most important features are related to geography, in addition to the third most important feature being "Total Funding". In a master thesis, Bento (2016) use Crunchbase data with classifiers Logistic Regression (LR), RF and SVM to build models predicting startup success, including SMOTE as an oversampling technique on the minority class of 'closed'. This provides them with a dataset size of 143,348 companies. They find the most important features to be whether the startup has a VC, whether it has an investor listed as a

top 500 investor in Crunchbase, and whether it has completed a round A of funding. In another master thesis, Ünal (2019) use data from 44,522 companies, including synthetically created instances from a technique called ADASYN, which is an oversampling technique based on SMOTE described in section 5.1.3. They create models with LR and RF, and four other models. The most important features are found to be last funding to date, first funding lag and company age. In a conference paper Krishna et al. (2016) tests 30 different machine learning techniques on 20 features using data from 11,000 companies. RF and a LR model called SimpleLogistic gave the best results, with "amount of funds raised" as one of the most important features. In a report by Stanford University researchers Pan et al. (2018) created a K-Nearest Neighbours (KNN), LR and RF to create their model, using data from 32,700 companies, gathered from Crunchbase. The best accuracy is achieved with the RF model at 84%, but with a higher recall with the KNN model.

Another relevant paper to consider is by Zbikowski and Antosiuk (2021). While all the above-mentioned papers include data that creates a look-ahead bias, they attempt to create a model that does not, so that it will be more useful for real life decision-making. From an investors' perspective, when making a prediction about a startup's future success, a model should not include information that would not be known to the investor at the moment of investing. Information about funding is not possible to know before the company has been funded. They created a model using a training set of 213,171 companies, using Crunchbase data and three different models: LR, SVM, and Gradient Boosting Classifier (GBC). They note that surviving over time is not adequate for a company to be labelled as successful, and they choose 'completing series B' as a label for success. Undertaking an IPO and acquisition (exit) are mentioned as other popular success metrics. They received success rates measured by precision, recall and F1 scores of 57%, 34% and 43%, respectively, using a GBC classifier. The model they created is high in accuracy and precision, but very low on recall. The models missed on 70% of the successful companies, classifying them as unsuccessful. They note that accuracy is not an appropriate metric when classes are imbalanced, in this case referring to an imbalance between successful/unsuccessful companies. Their model does not address gender bias, but rather uses gender as an input variable, noting that it is "one of the most important features". Further, they suggest that their general approach as outlined in their paper could be used by VC funds to build their own models.

3.2 GENDER BIAS IN EARLY-STAGE INVESTMENT DECISIONS

Start-up ventures that are led and/or owned by women tend to receive less funding than ventures led or owned by males, but the reasons why this is the case are unclear and debated amongst researchers in the field (Kanze et al., 2018). When developing and/or applying mitigation tools and strategies it is important to have thorough domain knowledge of the field in which we are applying them (Feuerriegel & Schwabe, 2020), here in the context of early-stage investment decisions. We need to understand where the bias is coming from, before we can know what path to take to prevent them (Mehrabi et al., 2022). This section will look at some of the research on gender bias in different areas of investment decision-making to get a better understanding of the problem. The focus will be on VC investors. The main point of this section of presenting relevant literature on gender bias in early-stage investment decisions is to show how gender bias affect investment decision-making, that is not founded in a rational evaluation of a startup, but rather part of implicit human bias that favors the male entrepreneur. This section has been awarded much space in this paper, as an agreement that there is a problem is an important first step in gender bias mitigation work. The statistics of funding on their own is not sufficient to state that there is gender bias. Neither is fairness metrics on its own sufficient. The area of gender bias in entrepreneurial financing has been well researched, and some of this research will be presented below.

3.2.1 GENDER ROLE CONGRUENCE THEORY AND REGULATORY FOCUS

The first set of studies presented here are within gender role congruence theory and regulatory focus theory. The field of entrepreneurship is, and has historically been, dominated by men, and consequently is mostly viewed as a 'masculine' field with 'entrepreneurial attitude' described as aggressive, risk taking, innovative, autonomous, and proactive (Alsos and Ljunggren, 2016). The female stereotype is not automatically linked to these types of attitudes, and gender role congruence theory state that women are penalized if they deviate from their gendered stereotype. Being gender role congruent means acting according to the gendered stereotype, while being gender role incongruent means acting in a way that does not match the gendered stereotype (Eagly & Karau, 2002). Role congruence theory has received significant support in gender research (Malmström et al., 2019), such as in leadership (Carli, 2010; Eagly & Karau, 2002) as well as in politics (Meeks, 2012). However, support is not found in law, meaning that female lawyers do not

seem to be “penalized” for displaying stereotypical male characteristics (Schneider et. al, 2010).

A second theory from the research on gender bias in entrepreneurial financing is regulatory focus. Regulatory focus theory offers two different categories of regulatory focus made by investors when evaluating entrepreneurs: prevention focus and promotion focus (Higgins, 1997). If investors are prevention focused, questions will emphasize “maintaining non-losses and not losing capital”, while being promotion-focused will elicit questions that “emphasize attaining growth-oriented gains that are facilitated by capital” (Kanze et al., 2018). The most important consequence of these two different types of considerations, is that promotion considerations from an investor will lead to less investment (Malmström et al., 2019; Kanze et al., 2018).

Signaling theory explains how signals are sent from one communicating party who chooses how to communicate (the sender) to another who interprets the signal (the receiver) (Spence, 2002). One study investigating gender congruence theory, regulatory focus and signaling theory combined is by Malmström et al. (2020). Their hypothesis is that signaling an ‘entrepreneurial attitude’ will have different impacts on investors depending on whether it is displayed from female or male entrepreneurs. Signaling an ‘entrepreneurial attitude’ is viewed by investors as positive, but when female entrepreneurs signal this, it is not in line with their gendered stereotype. Thus, it will elicit different reactions from investors, and the study looks at regulatory focus (prevention and promotion considerations) as a measure of these reactions. The study found that when female entrepreneurs signal an ‘entrepreneurial attitude’, they are more likely to elicit prevention considerations. When male entrepreneurs signal ‘entrepreneurial attitude’, they are more likely to elicit promotion considerations. Further, it is found that promotion consideration increases the amount of funding, while the opposite is true for prevention considerations.

Another study within the same environment (Swedish governmental VC firm) by Malmström et al. (2017) investigates whether the governmental venture capitalists describe male and female entrepreneurs differently, and whether this affects funding decisions. This is done by analyzing the words used to describe the female and male entrepreneurs, respectively. The characteristics ascribed to the female entrepreneurs differed considerably from the male entrepreneurs “with women’s potential *undermined,* but men’s potential *underpinned*”. Men and women are evaluated by different standards, likely because of the focus on masculine traits that is expected in entrepreneurship, the study find. Examples mentioned are describing who seemingly obtain financial resources as “She seems to have

expensive habits; who knows what she will do with the money if we approve her application" and a male entrepreneur with examples such as "Owning such an exclusive car tells me that he is financially solid". Another example found was that of being cautious were a female entrepreneur would be described in terms of "She is cautious as women often are, and she does not dare…" or "She is very cautious in what she does, and she does not have the guts to…", while a male described as cautious could be "He is cautious and that is good. He makes level-headed decisions". Overall, similar characteristics could be positive or negative depending on whether it was describing a male or a female entrepreneur.

The study shows how the governmental financiers socially construct gender stereotypes with their language and rhetoric. The study also finds that despite female entrepreneurs on average applying for lower amounts of funding, they are only awarded 25% of their applied funding amount, while male entrepreneurs received an average of 52% of their applied funding amount. Additionally, female entrepreneurs'' applications had a higher dismissal rate of around 53%, compared with the male entrepreneurs' dismissal rate of 38%.

Kanze et al. (2018) conducts a study to investigate whether male and female entrepreneurs receive different questions from VC investors, and whether such differences might explain the funding disparity between the genders. The study contains a total of 189 companies in pitching competitions (TechCrunch Disrupt Startup Battlefield) from 2010 to 2016. They measure the types of questions by regulatory focus. The study find that male entrepreneurs are being asked questions that are promotion focused, while female entrepreneurs are being asked questions that are prevention focused. The gender of the investor does not matter, both genders display this bias towards the female entrepreneurs. The study confirm that the regulatory focus significantly impacts the chance of and the amount of funding; being asked prevention focused questions (and responding accordingly) will affect funding negatively. Thus, the study confirms that female entrepreneurs raise significantly less funding than male entrepreneurs with comparable funding needs, and that the regulatory focus of the investors are the mechanism for the funding disparity. They also notice that entrepreneurs tend to respond with a matching regulatory focus, ie when they are asked promotion focused questions they will respond with a promotion focus. On the positive side, they find that when the prevention questions are asked, all entrepreneurs regardless of gender benefit from a switching intervention by replying in a promotion focused way.

Edelman et al. (2018) looked at gender stereotypes in angel investment processes from 2007 to 2016 in a northeast US angel group. The study investigates the comments (strengths

and weaknesses) from investors after a company presentation and Q&A with entrepreneurs. The comments are made without the presence of the entrepreneurs, and the written summaries are part of the study analysis. They separate comments into positive and negative, as well as personal and business characteristics. 358 ventures are included in the sample, where 15% are female led. The study did not find that female entrepreneurs received a higher number of negative comments from investors (compared to male entrepreneurs), but female entrepreneurs received a significantly higher number of comments on the management team. Additionally, it found some support surrounding legitimacy signals from female entrepreneurs. When women communicated legitimacy signals, such as referring to endorsements from strategic partners, angel investors were somewhat less likely to register it when it was from a female entrepreneur.

Johansson et al. (2021) find similar results as Edelman et al. (2018) when they investigate cognitive processes of financiers when evaluating investment proposals. This is another study conducted in the context of a Swedish government VC firm/institution, evaluating 77 investment proposals (40 female-owned). The study finds that significantly more weight is put on the personal characteristics of female entrepreneur rather than business characteristics, when compared male entrepreneurs. The authors conclude that there is a role incongruency with female entrepreneurs, in that they deviate from the stereotype of an entrepreneur, and as a result becomes more cognitively complex for financiers to evaluate as a business case.

3.2.2 GENDER EFFECTS ON VALUATION

Veer and Bringmann (2021) conducts a study with a sample of 166 startups from 2013-2017 to see whether it affects valuation of the company that an entrepreneur self-initiates a negotiation for funding by a European accelerator, and whether the gender of the entrepreneur matters in this context. If they self-initiate it means that they contact the accelerator, and not the other way around. The startups that self-initiate all receive lower valuations by the accelerator compared to those who do not self-initiate, and especially startups with female founders are disadvantage – the valuation gap is 219% between self-initiating startups with female founders and self-initiating startups with male founders. Female-led startups that do not self-initiate still receive worse valuations than male-led startups that self-initiate. If a female-led venture does not self-initiate, they receive a 69% higher valuation compared to if they do. The similar effect for male-led startups is a 28%

higher valuation when not self-initiating. The study finds a strong and significant effect of the CEO gender on startups' pre-money valuation in the favor of male-led startups. Industry, country, and venture-period fixed effects are controlled for, as well as weighting for startup quality. Overall, male CEOs receive a 72% higher valuation than female CEOs.

3.2.3 BIG DATA ANALYSIS ON EFFECT OF FOUNDER GENDER

Ewens and Townsend (2020) examines whether early-stage investors are biased against women using a large sample of 17.780 startups from the platform AngelList, a social media for fundraising. Even when female-led startups are comparable to male-led startups, the data analysis show that it is significantly more difficult for the female founders to obtain both interest and capital from male investors. The exception is when they are seeking low amounts of capital or if they are in female-centric industries, in this case the disadvantage is slightly smaller. On the contrary, when a female-led venture is affiliated with an incubator, the gap in male investor interest is even larger compared with a similar male-led venture. The pattern is not found with female investors. Ewens and Townsend (2020) note: "This suggests that male investors may pigeonhole female entrepreneurs to some extent, believing them only able to succeed in relatively unambitious businesses or businesses oriented toward women".

3.2.4 EFFECT OF FEMININE BEHAVIOR

Balachandra et al. (2017) examine how feminine or masculine behavior in addition to gender may elicit different reactions from VC investors. The study used a sample of 185 one-minute video pitches (20% from female entrepreneurs) from competitions in 2007 and 2008 that were shown to VC investors to evaluate, although not make real investment decisions. The study found that any display of feminine behavior from either male or female entrepreneurs creates a funding disparity in the disfavor of those displaying feminine behavior. They did not find evidence for gender incongruence penalties, meaning that women who acted inconsistently with their gender stereotype in the pitching situation were not penalized. The study adds that a display of masculine-stereotyped behaviors is not an advantage, but female-stereotyped behavior leads to a disadvantage.

3.2.5 GENDER AND COMMUNICATION

Balachandra et al. (2021) investigates whether male and female entrepreneurs use different language styles, using the same data as from Balachandra et al. (2019). They did not find any differences from the entrepreneurs, but they did find that investors overall preferred a neutral language – neither feminine or masculine linguistic style. The exception was that it was an advantage to use 'inspirational language' which was classified as a masculine linguistic style. They did not find evidence for role congruence theory, meaning that women were not disadvantaged for using a masculine linguistic style more than men.

Huang et al. (2020) also investigate the communication styles of entrepreneurs of different gender and whether this can be a cause of the funding disparity. Specifically, they look at differences in abstract and concrete communication styles through three different studies. They note that entrepreneurs communicating with a concrete language can signal to investors that they are focused on immediate problems and goals. Entrepreneurs using abstract language signal that the focus is on the bigger picture. The study finds that the use of abstract language is more common from male entrepreneurs, and that using this type of language increase chances of funding.

Brooks et al. (2014) conducted three separate studies on the effect of entrepreneur gender and attractiveness and communication on investor evaluation. The first study looked at whether the rated attractiveness of an entrepreneur would affect outcome, where the attractiveness was rated separately from the investors evaluating the pitch. They found that an entrepreneur's rated attractiveness influences the pitch success for males, but not for females. In the second study, participants watched two videos of a pitch that was manipulated to seem like it was presented by a male or female entrepreneur where voice-over was manipulated but was otherwise identical. 68% chose that they would fund the ventures pitched by a male voice, 32% chose the ventures pitched by a female voice, where an equal outcome with no gender bias would have yielded 50% for each gender. Participants were recruited from Amazon Mechanical Turk and were not real investors, neither did they make any actual investments. The same setup was used in the third study with 194 participants (57.7% female), except only one video was presented, and participants asked how likely they would hypothetically be to invest, and whether they thought the pitch was 1) persuasive, 2) fact-based, and 3) logical. Pitches were rated as significantly more persuasive, fact based and logical when they were presented by men, although the pitches were identical to the ones presented by a female voice and image.

Participants were significantly more likely to invest after watching the high-attractiveness male pitchers vs the low-attractiveness male pitchers, while there was no attractiveness effect with female pitchers.

3.3 GENDER BIAS IN MACHINE LEARNING MODELS

"When [machine learning models] produce discriminatory results based on sensitive traits such as gender, we consider them to be 'biased' or 'unfair'" (Feldman & Peake, 2021).
Machine learning is often portrayed as an objective and automated tool free from human bias, although both researchers and practitioners are increasingly seeing to what extent human bias is being built into machine learning models (Ajunwa, 2020). The level of publication activity on gender bias in AI models has rapidly increased in recent years, starting in 2017, with a focus on finding the contributing factors and methods to approach gender bias in AI. The research is still in its early phases but growing fast (Fu et al., 2021; Nadeem et al., 2020).
Crawford (2016) warns that "we risk constructing machine intelligence that mirrors a narrow and privileged vision of society, with its old, familiar biases and stereotypes" if we do not take action to prevent it. We risk that existing gender bias is magnified in machine learning models, and even that new types of biases can be created (Ntoutsi et al., 2019). Leavy et al. (2020) suggest incorporating both gender theory and ensure gender balance in data to minimize the risk of gender imbalances. As demonstrated in section 2.1, gender bias is difficult to identify in real-life settings even without using any machine learning. One of the big challenges when it comes to identifying bias in algorithms, is that they will often be proprietary information and not openly available, as well as possibly being too complex for humans to understand (Fu et al., 2020). This is important to keep in mind when developing mitigation tools and strategies. This section will first look at the issue of black-box and proprietary algorithms, causes and occurrences of gender bias in ML discuss the issue of protected attributes, investigate industry needs (including both developers and users).

3.3.1 BLACK BOX ALGORITHMS

When discussing fairness and interpretability of a machine learning model, there is usually two main categories: interpretable algorithms and black box algorithms. An interpretable machine learning model is possible for a human to interpret. A black box algorithm is not possible for a human to interpret. According to Zhou et al. (2021), "The most successful

Machine Learning (ML) systems remain complex black boxes to users, and even experts are often unable to understand the rationale behind their decisions".

Rudin (2019) gives to possible paths to a black box model. First, it could be 'a function that is too complicated for any human to comprehend', or it can be a proprietary function/algorithm. In the context of entrepreneurial finance, it is likely that both types will be found. Rudin (2019) argues that instead of trying to come up with methods or new models that can explain the black box models, developers should rather focus on designing transparent and interpretable models, especially in high-stake decisions. They point at the fact that models such as deep learning that create a black box, might not give much better results than simpler models that are transparent. Omidi (2021) use different models to test ML techniques on a dataset predicting heart failure, finding that neural networks performed worst in terms of accuracy, while Logistic Regression and Decision Trees had the best accuracy. Additionally, these simpler and better interpretable models showed better results on various fairness metrics.

The second type of black box algorithms can occur when a company makes it proprietary. If a VC develop its own machine learning model to pick winners, this is likely to be a proprietary algorithm as it could give the VC a competitive advantage in the market.

The ability for a VC company to pick the most prominent startup is at the core of their business. To put it simply: If they pick non-successful startups, they are unlikely to make money, and the more successful their chosen startups are, the more revenue a VC will make. Having a proven successful track record is also likely to attract more successful startups and investors to the VC.

3.3.2 CAUSES AND OCCURRENCES

Although there is no work investigating gender bias in machine learning models made within the entrepreneurial finance domain, it can be useful to look at closely related domains, such as gender bias in recruitment algorithms or financial decision-making algorithms such as credit assignment or loan approval algorithms.

From the media and literature there are some well-known examples of algorithmic gender bias (or even gender discrimination) that has been covered extensively. Ishita Rustagi and Genevieve Smith from Center for Equity, Gender, and Leadership at Berkeley Haas School

of Business has made an AI bias examples tracker, an open Google spreadsheet[2]. Some examples include the Apple Card algorithm which seemed to assign lower credit to women (O'Sullivan, 2021), using Google News articles as training data using word embeddings which exhibited female/male gender stereotypes (Bolukbasi et al., 2016) and Amazon's hiring AI that was biased against female candidates due to the historical male dominance in the tech industry (Dastin, 2018). Smith and Rustagi (2021) found in an analysis of 133 systems across industries from 1988-2021 that 44.2% demonstrated gender bias. They found that 70% of these systems led to women and non-binary people receiving services of lower quality, 61.5% led to an unfair allocation of resources, information, and opportunities and 28.8% reinforced existing harmful stereotypes and prejudices. Zhao et al. (2017) showed how existing gender bias was amplified by a machine learning model; in the training set of the dataset imSitu, there were images of people cooking where 33% were men and the rest women. The model amplified this so that after training only 16% of the cooking images had men in them.

One commonly used open dataset is the Adult Census Income dataset, and according to Feldman and Peake (2021) this is one of the most used datasets to explore gender fairness in machine learning. They use it to explore gender bias with deep learning models. Without any fairness improving techniques are applied, their deep learning model has the highest accuracy, as well as the highest level of unfairness based on the following fairness metrics: Statistical Parity Difference, Equal Opportunity Difference and Average Odds Difference. These are all described in section 4.2. This means that using this dataset in creating machine learning models will likely create a gender bias, that would be unknown to researchers or practitioners simply measuring by accuracy and not fairness scores.

3.3.3 THE ISSUE OF PROTECTED ATTRIBUTES

It is not recommended to use protected attributes such as gender or race in the training of a machine learning model, however they should not be ignored either, as this could lead to less fairness (Haeri & Zweig, 2020). Bellamy et al. (2019) describes a protected attribute as *"an attribute that partitions a population into groups that have parity in terms of benefit*

[2] Google Spreadsheet of AI bias examples https://docs.google.com/spreadsheets/d/1eyZZW7eZAfzlUMD8kSU30IPwshHS4ZBOyZXfEBiZum4/edit#gid=1838901553

received. Protected attributes are not universal but are application specific". In the context of this paper the protected attribute we are focused on is gender. Some models that are currently found in investment decision making use gender as a predictive variable, as can be seen in section 3.1.2, which is generally not recommended. Removing them might not be enough on its own, although they should not be part of any training data. This is because other variables might correlate with gender, and thus serve as proxy attributes (Feuerriegel & Schwabe, 2020; Ntoutsi et al., 2019; Hardt et al., 2016; Calmon etal., 2017; Besse et al., 2020). Besse et al. (2020) shows that on when gender is removed from the UCI Adult Census Dataset, both disparate impact and accuracy are almost unchanged. Calmon et al. (2017) finds a significant reduction in discrimination when removing the protected attribute (gender) from the Adult UCI dataset, while this is not the case when removing protected attributes from the COMPAS dataset. When protected attributes are removed but there is still discrimination found in the predictions, it is likely that there are strong dependencies of the protected attribute(s) to other attributes in the data.

If the attribute is available, it can be used to measure causal influences between gender and other attributes, in addition to being used to measure equality of outcomes (Ntoutsi et al., 2019). Many of the existing fairness tools require access to the sensitive attributes (Veale & Binns, 2017). Haeri & Zweig (2020) argue for the important role of awareness protected attributes in datasets and show that the complete removal or avoidance of them might decrease fairness. Their conclusion is that it is important to evaluate the exclusion or inclusion of protected attributes on a case-by-case basis – including in different datasets and in different models applied to the datasets. This is because, they argue, there are cases in which using the protected attributes can result in both a fairer and more accurate model. This can mean to include these attributes in any phase of the model development to remove bias, such as applying techniques to remove bias in the pre-processing or to measure result in post-processing.

3.3.4 ML INDUSTRY NEED FOR TOOLS AND STRATEGIES

Holstein et al. (2019) state that even though there has been an increased focus on improving fairness in ML models in recent years, the design of the tools developed are not driven by real-world needs, but rather by the availability of algorithmic methods. They conduct 3 interviews and survey 267 ML practitioners to understand real world needs to enable more

fair ML models. One important feedback point from practitioners is that many of the algorithmic methods to mitigate biases view the dataset as fixed, while real life datasets change continuously. Further, the tools are too focused on mitigating bias on the ML development pipeline rather than in the data collection. Collecting high quality and un-biased data from the beginning should receive more focus.

Practitioners often describe their fairness attempts as reactive rather than proactive. Often, fairness issues are not discovered before the machine learning models are put to use and they receive complaints either from customers directly or through negative media coverage. Most of the teams of the interviewees in Holstein et al. (2019) state that they currently do not have any fairness metrics to monitor performance and progress. This was not necessarily because they have not tried at this, but because in the search for fairness metrics they could not find any that were directly applicable to the specific domain in which they were creating a ML model. Further, Holstein et al. (2019) collected feedback from the ML practitioners stating that most of the fairness literature assume that the data set contains sensitive variables such as gender and race, while real world data does not always have this at an individual level.

4 FAIRNESS IN MACHINE LEARNING MODELS

In the computer science literature alone, there are more than twenty definitions of fairness. Many academics have tried to develop formal and mathematical definitions, but without success; the definition of fairness is still an open question (Verma & Rubin, 2018; Ntoutsi, 2019). Since there are so many different, and at times conflicting, definitions of fairness, it is important to reflect on the different definitions and choose which ones to use when developing tools to increase fairness. It is possible to say that one data set or machine learning model is fair (or unfair) depending on one definition, but not another (Verma & Rubin, 2018). The differences between the definitions are not purely theoretical, they also result in different outcomes, and it is not possible to use them all at the same time (Bellamy et al., 2018; Kleinberg et al., 2016). Mehrabi et al. (2019) notes that many fairness definitions in the literature revolve around equality, rather than equity. Equality ensures an equal number of resources, attention, or outcome while equity is more focused on each group or individual being given the necessary resources needed to succeed. There is often a tradeoff between accuracy and fairness in machine learning algorithms (Kamishima et al. 2012; Kamishima et al. 2011). Verma and Rubin (2018) concludes their overview of algorithmic fairness definitions with the statement "more work is needed to clarify which definitions are appropriate to each particular situation". Fu et al. (2020) concludes that choosing the most important fairness notions must be made on a case-by-case basis.
The next sections will describe fairness categories and fairness metrics.

4.1 FAIRNESS CATEGORIES

Fairness notions are commonly described in four major categories: fairness through unawareness, individual fairness, group fairness and counterfactual fairness (Fu et al. 2020; Verma & Rubin, 2018). Fairness through unawareness is a fairness notion that is consistent with what legislation refers to as Disparate Treatment (DT), which will be explained in section 4.2.1. In machine learning this is simple to achieve by excluding any protected attributes. In the context of this paper, to achieve fairness through unawareness any feature involving gender must be excluded from the machine learning model. The idea is that if the protected attribute is not known, it is not possible to discriminate based on it. However, as described in section 3.3.3, excluding a protected attribute might not matter in terms of other notions of fairness in a machine learning context, such as Disparate Impact (DI). Simply

ignoring the protected features is not enough to conclude with a model being fair (Haeri & Zweig, 2020). Until now, the most common strategy in trying to create unbiased or fair models have been fairness through unawareness – not including protected attributes such as gender, race or age in the model development and thus letting the model be unaware of these with the assumption that the model cannot discriminate if it is unaware of these protected attributes. However, as has been shown in previous works, this is not enough.
Individual fairness is by Dwork et al. (2011) the "principle that any two individuals who are similar *with respect to a particular task* should be classified similarly". As an example, this means that if two founders are equal on all aspects other than gender, they should be treated similarly, ie have the same probability of a certain outcome. Group fairness is described by Bellamy et al. (2019) is *"*the goal of groups defined by protected attributes receiving similar treatments or outcomes". Counterfactual fairness "reflects the idea that a decision regarding an individual should not change if the individual were in a counterfactual world with a different sensitive attribute value" (Fu et al., 2020).

4.2 FAIRNESS METRICS

In the next sections, some of the most popular fairness notions and their metrics in machine learning models will be presented. Following the papers in which the AIF360 toolkit (described in section 5.5.1) is based on, the following five fairness metrics will be investigated: Disparate Impact (DI), Average Odds (AO), Statistical Parity (SP) and Equal Opportunity (EO). Note that all four assume that the protected class, in this case gender, is known (Bellamy et al., 2019). In a real life setting it might be the exception that the protected attributes are known, and thus some methods that does not require knowledge of protected attributes are explored (Zhao et al., 2021).

4.2.1 DISPARATE IMPACT (DI)

Disparate Impact occurs when (positive) outcomes from a selection process varies significantly between groups, even if the selection process seems to be similar for all groups (Barocas and Selbst, 2016). Feldman et al. (2015) define DI in algorithms to occur "when a selection process has widely different outcomes for different groups even as it appears to be neutral". DI is in the category of group fairness. It is important to know, as Feldman et al. (2015) point out, that DI is usually not illegal – in contrast to Disparate Treatment (DT) which refers to intentional discrimination and could occur if the protected attributes are

inputs to a machine learning model (Zafar et al., 2017). DI refers to the unintentional rather than intentional or direct bias and discrimination and might occur even when a model does not have access to information about protected characteristics. This is especially a risk if there is unfair treatment of certain groups in the historical data that the machine is trained on.

The Disparate Impact is measured by:

$$\frac{Pr(\hat{Y} = 1 | D = unprivileged)}{Pr(\hat{Y} = 1 | D = privileged)}$$

The optimal value is 1 – both groups receive an equal number of positive outcomes. A value of [0.8, 1.25] will be considered fair. This is related to a guideline used in legal settings, where disparate impact is said to be identified when the value as calculated by the above formula is =<0.8, called 'the 80% rule' or the 'four-fifths-rule' (Foulds et al., 2019)

4.2.2 STATISTICAL PARITY (SP)

Statistical parity also goes by the term mean difference or demographic parity (Zhao & Gordon, 2022). As with the disparate impact, SP is based on predicted labels, and thus it is possible to calculate it both from the entire input dataset or from a model's predicted outcomes (Bellamy et al., 2019). It requires independence between the protected attributes and outcomes (Fu et al., 2020).

The measurement of SP is called Statistical Parity Difference and is measured in the following way between two groups (unprivileged and privileged):

$$Pr(\hat{Y} = 1 \mid D = unprivileged) - Pr(\hat{Y} = 1 | D = privileged)$$

A value of 0 means that there is no difference between the groups. If the SPD value is negative, the male founders as a group has a higher probability of a favorable outcome. If the SPD is positive, the female founder as a group has a higher probability of a favorable outcome. A model is considered fair by this definition if the value is in the range [-0.1,0.1] (Feldman & Peake, 2021), but this is dependent on the use case. Statistical Parity is one of the most popular fairness notions in computer science. Zhao and Gordon (2022) note that

a tradeoff between SP and accuracy has been observed in many experiments. Aiming for exact SP is found not to be compatible with a perfect predictor (Hardt et al., 2016).

4.2.3 EQUAL OPPORTUNITY (EO)

Equal Opportunity, proposed by Hardt et al. (2016) and is also amongst the most popular fairness notions (Fu et al., 2020). Where SPD might propose a 'qualification problem' as presented in the previous section, EO seeks to repair this by stating that it is similarly *qualified individuals* that should have an equal probability of a favorable outcome (Fu et al., 2020; Hardt et al., 2016). The metric for EO in machine learning applications is Equal Opportunity Difference (EOD), developed by Hardt et al. (2016). The creators point out that in the field of machine learning there are few good fairness aimed to avoid discrimination against protected attributes. They look at the limitations of statistical parity and develop EOD as a new metric that also have a higher utility, as it better serves the purpose of achieving a higher accuracy.
The definition of equal opportunity from Hardt et al. (2016) reads that:
A binary predictor $\widehat{Y}$ satisfies equal opportunity with respect to A and Y if

$$P(\widehat{Y} = 1 \mid \mathrm{A} = 0, \mathrm{Y} = 1) = P(\widehat{Y} = 1 | \mathrm{A} = 1, \mathrm{Y} = 1)$$

EOD is possible to measure through a post-processing step and does not require any alteration of the model. It is measured in differences between unprivileged and privileged groups. The optimal score is 0. A model is considered fair by this definition if the value is in the range [-0.1,0.1] (Feldman & Peake, 2021), but this is dependent on the use case. The EOD is defined through the true positive rate difference as:

$$TPR_{D=unprivileged} - TPR_{D=unprivileged}$$

This means that the true positives rates for both groups should be equal for the model to be considered fair according to this metric. It can also be viewed as a difference in the recall scores between two groups. A good EOD score should not be taken as a proof of fairness on its own, and neither should a low EOD be considered as a proof of unfairness. And as the researchers note: the focus should always primarily be on finding better features and more data, rather than simply checking the EOD.

4.2.4 AVERAGE ODDS (AO)

The measurement of AO in a machine learning model is called Average Odds Difference (AOD) and is based on metrics from the confusion matrix of true labels and predicted labels. It uses the average of false positive rate (FPR) with the number of false positives divided by the number of false negatives, and the average of true positive rate (TPR) where the number of true positives is divided by the number of true negatives for each group. If the AOD equals 0, the groups are equal. A negative number implies that the privileged group has a benefit (Bellamy et al., 2019). Fairness is said to be achieved between if values are between -0.1 and 0.1 (Feldman & Peake, 2021), but this is dependent on the use case. In AIF360 the AOD is calculated as follows:

$$\frac{\left(FPR_{unprivileged} - FPR_{privileged}\right) + \left(TPR_{unprivileged} - TPR_{privileged}\right)}{2}$$

5 STATE OF THE ART METHODS AND TOOLS: BIAS MITIGATION TOOLS

This section will look at the methods and tools for mitigating gender bias in machine learning models. Gender bias in a machine learning model can appear at different steps of the development process (Feuerriegel & Schwabe, 2020). Mitigation tools for machine learning models are commonly split in three main categories: 1) preprocessing, 2) in-processing and 3) postprocessing (Bellamy et al., 2019; d'Allesandro, 2017; Feldman and Peake, 2021; Mehrabi et al., 2019). These three categories will be investigated in the section below, together with some of the open-source tools that are available and work across the categories. Feldman and Peake (2021) found that making an end-to-end bias mitigation framework, meaning a combination of pre-, in-, and post-processing tools, was the most effective way to mitigate gender bias.

5.1 PRE-PROCESSING TECHNIQUES

Preprocessing techniques seeks to remove the underlying bias from the data before it is used in the model (d'Alessandro et al., 2019). One of the major benefits of pre-processing algorithms is that they are regardless of the model, meaning that they can be used in a black box setting. It also seeks to remove the root cause of the bias in the data. Unfortunately, it sometimes may result in loss of information in the data, which can cause a decline in accuracy metrics (Feldman and Peake, 2021). The first section on pre-processing techniques will discuss the importance of choosing the right data. In many cases we can assume that the data is biased or can run some tests if we have the information on gender. The second section will look at feature extraction and handling. The third section will discuss how protected attributes should be handled. The fourth section will look at how to handle imbalanced classes, and different techniques for sampling data when classes are imbalanced. The final four sections on the topic of pre-processing techniques will describe various techniques to improve fairness, all of which are available in the AIF360 toolbox described in section 5.5.1. All four are suggested in a paper by Zhang and Zhou (2019) in techniques for bias mitigation in the financial industry. Examples of how the methods affect fairness will be discussed where this is relevant.

5.1.1 DATA SELECTION

The selection of the dataset is the first and major factor to consider in building a machine learning model. Responsible selection of training data shapes the outcomes of the ML model (Nadeem et al., 2020), and most fairness issues seem to arise from the data (Veale & Binns, 2017). Interviews with industry practitioners on ML fairness show that there is a need to focus on data collection (rather than model development) when it comes to mitigating bias (Holstein et al., 2019). Transparency in data collection and design choices is important, and even more so if protected attributes such as gender is not readily available in the data to measure and mitigate bias. Holland et al. (2018) proposes a framework that entails labelling datasets like how we use nutrition labels on food, called "The Dataset Nutrition Label", which could be considered by any company seeking to create a machine learning model within entrepreneurial financing.

5.1.2 FEATURE EXTRACTION

Feature extraction entails extracting information from data. There are normally numerous possibilities of combining information from a dataset to create features. The features we collect from a dataset, is the structured information that is given to a machine learning model. This is where, even in a dataset which might not be biased initially, could become so from the features that are chosen by a human. When determining which data and which input and output variables to use, it is important to keep in mind that the historical success of previously funded ventures is to a large extent dependent on the fact that the ventures were indeed funded. Without funding it is significantly harder to grow a business. Let's also consider that there might be constraints on women to even consider starting an entrepreneurial venture, which also keeps the number of women seeking funding low. There is no easy solution to this, but nonetheless important to keep in mind when working with the ML models in entrepreneurial financing. Some features are clearly more sensitive or prone to bias than others, and both domain knowledge of investment but also gender theory could be important to understand this.

5.1.3 CLASS IMBALANCES AND SAMPLING

As we have seen, there are a massive underrepresentation of women entrepreneurs being funded measured against the 50% demographic women make up. Zhang & Zhou (2019) explain that when classes are imbalanced, a machine learning model can be biased against

the minority class simply because it lacks enough data. According to Chakraborty et al. (2021), this is a widely studied topic in machine learning. Normally, imbalance in data refers to imbalances in the classification classes, but it can also refer to an imbalance in classes of protected attributes. The bias from the imbalance itself can be amplified by a machine learning model (Chakraborty et al. (2021).

To help balance the classes so that they are equal in size, different methods for over- and under-sampling can be applied. Over-sampling seeks to increase the number of the minority group, while under-sampling decreases the majority group so that it is closer to the minority group size.

Chawla et al. (2002), create a method for over-sampling called Synthetic Minority Over-sampling Technique (SMOTE). According to the creators imbalanced classes often lead to a high false negative rate in a model. They suggest resampling as one of the methods to account for imbalanced classes, with a combination of under-sampling and over-sampling. They note that a model's cost of misclassifying examples that belong to a minority group tend to be much higher than the cost of misclassifying examples from the majority class. Their solution is to create synthetic examples from samples of the minority group. Batista et al. (2004) found that SMOTE outperformed other sampling methods. Zhang and Zhou (2019) explored the effect of SMOTE in credit card default predictions. They applied SMOTE on the classification outcomes and not gender variables, as the number of each gender was not imbalanced. In the original dataset, males had 3% more positive outcomes. The predictions were measured in terms of Disparate Impact, Statistical Parity Difference and Equality of Opportunity Difference. They found that SMOTE only decreased accuracy by 0.01 percentage points, but it did improve false negative rates from 0.62 to 0.23. In terms of SPD, EOD and DI, they found that the data after the SMOTE application were more biased, suggesting that SMOTE amplifies bias in the data.

Chakraborty et al. (2021) tested four different class balancing techniques and its effect on fairness on a well-known dataset called the UCI Adult Census Dataset. In this dataset, the goal is to predict income based on features including gender and race, with predictions either as "high income" or "low income". Balancing techniques were applied to the protected attributes of "sex" (male/female). 86% of the instances were male and 14% were female. They measured the effect on fairness metrics AOD, EOD, SPD and DI (fairness metrics described in section 4.2), and found that all four class balancing techniques impaired fairness, ie made the model more biased towards the minority group. In this case

accuracy is also impaired, but they note that normally accuracy is improved while fairness is impaired when using class balancing techniques. The explanation provided for this is that "these techniques randomly generate/discard samples just to equalize two classes and completely ignore the attributes and hence damage the protected attribute balance even more". The researchers develop a new technique to balance classes that in their experiments with it does not damage fairness metrics based on protected attribute, called Fair-SMOTE. They apply this to the UCI Adult Census Dataset (Kohavi & Becker, 1996) and compared to the first methods of class balancing and find that Fair-SMOTE significantly improves fairness metrics. It does not improve accuracy from the baseline (no class balancing), but it gives a better accuracy using Logistic Regression than the original class balancing techniques. It gives a significantly higher F1 score than the baseline model.

5.1.4 REWEIGHING

Calders et al. (2009) developed a new technique for pre-processing of data for machine learning applications, where the aim is to let the model "know" that predictions should not consider protected attributes, such as gender. Their method also looks at features that are correlating with the protected attribute. The method does not change class labels, but re-samples it in a certain way to remove dependencies by assigning different weights to them but maintain the overall probability of the positive class. More specifically Calders et al. (2009) explain this reweighing process in the following way:
Objects with $B = b$ and $Class = +$ will get higher weights than objects with $B = b$ and $Class = -$ and objects with $B \neq b$ and $Class = +$ will get lower weights than objects with $B \neq b$ and $Class = -$. The researchers note that there is an accuracy-bias trade off when applying reweighing to data.

5.1.5 OPTIMIZED PRE-PROCESSING

Calmon et al. (2017) developed a method called Optimized Pre-Processing to prevent discrimination in supervised learning, with the aim of improving both group fairness and individual fairness. The technique transforms both labels and features. An important advantage of this method is that it does not need to know the sensitive attributes, as it seeks to reduce the SPD of all possible combinations of groups found in the data. However, to measure the impact of the technique on a dataset, the protected attributes must be known. The researchers claim that the method has little impact on accuracy, which they confirm in

their tests. The method is tested on the UCI Adult Dataset (Kohavi & Becker, 1996), using Logistic Regression and Random Forest, and it is found to significantly reduce discrimination, both group and individual.

5.1.6 LEARNING FAIR REPRESENTATIONS (LFR)

Zemel et al. (2013) proposes a technique called Learning Fair Representations to achieve group fairness and individual fairness. They formulate fairness as an "optimization problem of finding a good representation of the data with two competing goals: to encode the data as well as possible, while simultaneously obfuscating any information about membership in the protected group".

5.1.7 DISPARATE IMPACT REMOVER

Feldman (2015) created a method to remove disparate impact that does not need access to the machine learning model itself but changes the input data while class values are unchanged. The protected attributes are removed from the data before applying the Disparate Impact Remover. The method is tested on the UCI Adult Census Dataset, using Logistic Regression, Support Vector Machine and Gaussian Naïve Bayes. The method improves fairness to a DI >= 0.8, but decreases accuracy some, demonstrating a fairness/utility tradeoff. The method can also be used on multiple protected attributes in the same data; however, it can only be used on continuous variables.

5.2 IN-PROCESSING TECHNIQUES

In-processing techniques are adjustments to the traditional algorithms in the training phase, that seek to address discrimination/bias, such as cost or loss function, regularizers and other modifications (d'Alessandro et al., 2019). This paper will investigate two such techniques for bias reduction: Adversial Debiasing and Prejudice Remover, both suggested by Zhang and Zhou (2019) for bias mitigation in AI models in the financial industry.

5.2.1 ADVERSARIAL DEBIASING

Zhang et al. (2018) created a technique called Adversarial Learning (AL) with the aim of maximizing a model's ability to predict some target variable, while at the same time minimizing the model's ability to predict a protected attribute. This information is then

used to decorrelate the protected attributes from potential biases, resulting in a reduced impact of the protected attribute on the target variable. The technique is tested on the UCI Adult Census Dataset (Kohavi & Becker, 1996). In their experiment, there is a small fairness utility trade-off, where the fairness metric used is Equality of Odds. The accuracy is decreased from 0.86 to 0.845. Omidi (2021) tested it on various data sets with a range of different ML Models and found that adversarial de-biasing was the overall best technique when working with deep learning classifiers.

According to Feldman and Peake (2021), this technique has the advantage that it 'maintains the integrity of the data', and thus could give a higher accuracy compared with other methods. Additionally, it does not require any assumptions about the distribution of the dataset. It is however, not possible to apply without access to model parameters, and thus cannot be applied in a black-box setting.

5.2.2 PREJUDICE REMOVER

Kamishima et al. (2012) create a regularization approach called Prejudice Remover (PR) that can be applied to probabilistic discriminative models. The PR aims at data where unfairness is from prejudice, where prejudice refers to a dependence between information relating to protected attributes and other data. They use the term "indirect prejudice" which is a synonym to disparate impact, and this is what the PR attempts to remove, specifically the dependence of protected attributes with other attributes in the data. When using this technique, the most relevant metric to measure results is Disparate Impact (DI). They test the PR on a model using Logistic Regression and Naïve Bayes classifiers on the UCI Adult Census Dataset (Kohavi & Becker, 1996). In their experiment, there is a trade-off between accuracy and DI, demonstrating a fairness/utility trade-off.

5.3 POST-PROCESSING TECHNIQUES

Post-processing techniques are applied after the training phase and can be applied even in a black box setting (d'Alessandro, 2019). A major advantage is that they do not need access to the model pipeline itself, and some do not need access to raw data. Thus, it can be applied without access to model parameters, including in a black box setting. This paper will investigate one such technique called Equalized Odds Post-processing and Calibrated Equalized Odds post-processing, both suggested by Zhang and Zhou (2021) for use in bias mitigation in AI models in the financial industry In post-processing techniques, a model's

predictions are manipulated according to a pre-defined fairness constraint (Feldman & Peake, 2021).

5.3.1 EQUALIZED ODDS POST-PROCESSING

Hardt et al. (2016) proposes a technique to "shift the cost of poor classification from disadvantaged groups to the decision maker, who can respond by improving the classification accuracy". Importantly, this technique can be applied even with a proprietary algorithm, as it only requires aggregate information about the data. It does not need access to raw data and does not require any changes to the model pipeline. However, it does require that the true values of protected attributes are known. Based on this it changes output labels with the aim of optimizing Equalized Odds. The creators argue that Equalized Odds is a better fairness metric than Demographic Parity.

5.3.2 CALIBRATED EQUALIZED ODDS POST-PROCESSING

Pleiss et al. (2017) proposes Calibrated Equalized odds as another similar technique as Equalized Odds post-processing. As with Equalized Odds post-processing technique, it needs to have access to protected attributes. When tested on the UCI Adult Census dataset by Pleiss et al. (2017), it helps achieve SP, but accuracy is decreased with around 10%.

5.4 MODEL SELECTION

Different models can affect fairness metrics in different ways. When fairness is discussed in the context of machine learning, the context of transparency often comes up, where it is emphasized that the decision-making process of a model should be as transparent as possible, so that it can be understood by humans who are acting on the information provided by the model. The choice of a model is important for how easy it is for humans to understand and evaluate how the machine learning model came to its predictions. Haeri and Zweig (2020) stress the importance of using models that are transparent and explainable to end users, especially when working with sensitive attributes. Because of these concerns, simpler models might be preferred over other more complicated models that are opaquer and more difficult to understand.

According to Ferrati and Muffato (2021) the most used model in research on entrepreneurial finance is random forest followed by logistic regression and support vector machine.

Logistic regression has the advantage that it is efficient and quite easy to interpret, and commonly used in machine learning applications, especially with binary classification. As can be seen in previous sections, it is also commonly used to test fairness mitigation strategies. Logistic Regression handles both categorical and continuous input variables. It can also handle 'problems of collinearity, missing data, redundant attributes and nonlinear separability' (Maalouf, 2011).

5.5 END-TO-END BIAS MITIGATION

Feldman & Peake (2021) propose a combination of multiple algorithms aimed at improving fairness in deep learning models. Bias mitigation algorithms are applied in each phase of the process: pre-processing, in-processing, and post-processing. In their paper they compare this approach to using single methods. They use Disparate Impact Remover in the pre-processing phase, Adversial Debiasing in in-processing and Calibrated Equalized Odds in post-processing. They train one model without any bias mitigation techniques, then combinations of the three techniques. They find that the unbiased model performs best in terms of classification accuracy but is unfair measured by all three fairness metrics. The worst model in terms of fairness, measured across all three metrics, was the model only using the Disparate Impact Remover as a pre-processing technique – this makes the model, compared to the unmitigated baseline, more biased. When all three techniques are applied together, the model cannot be deemed fair according to the metrics, and accuracy is decreased from 0.83 to 0.81.

5.5.1 OPEN-SOURCE TOOLS FOR DEVELOPERS

Some open-source tools and metrics that are developed for practitioners include: IBM's AI Fairness 360 (AIF360) toolkit (Bellamy et al., 2019), Microsoft's Fairlearn Python Package (Bird et al., 2020), Google's What-If Tool[3], Googles Fairness Indicators[4], FairSight (Ahn & Lin, 2019) and Aequitas (Saleiro et al., 2019). According to Fu et al. (2020) the open-

[3] Google's What-If Tool https://pair-code.github.io/what-if-tool/

[4] Google's Fairness Indicators https://www.tensorflow.org/tfx/guide/fairness_indicators

source tools are complex and challenging and Chouldechova (2016) and Kleinberg et al. (2016) has proven that the tools cannot handle different technical definitions of fairness simultaneously. There are also commercial bias mitigating tools available, such as Amazon SageMaker Clarify[5], which, according to their own website, "*...provides machine learning developers with greater visibility into their training data and models so they can identify and limit bias and explain predictions".*

AIF360 includes techniques from eight different published papers within the algorithm fairness domain, with Python-based tools for bias detection and mitigation including 71 metrics for detection and 9 algorithms as well as explanation to help understand the results. Building on research, have created three approaches to bias mitigation; fair pre-processing, fair in-processing, and fair post-processing (Bellamy et al., 2019).

[5] Amazon SageMaker Clarify. https://aws.amazon.com/sagemaker/clarify/

6 METHODS

Three methods have been combined to investigate the research question of this paper. For the first part, interviews with investors working in VC companies were conducted. For the second part, a machine learning model was created. For the third part, different machine learning techniques and their effect on gender bias as measured by fairness metrics described in section 4.2.

6.1 METHOD DESCRIPTION: INDUSTRY INTERVIEWS

As part of the work for this paper, five interviews with VC investors have been made. E-mails were sent to ten different VC companies in total. The interviews were informal and unstructured with open-ended questions around the three main topics outlined below, and each lasted around 30-60 minutes, all except one through video/phone. Three of the companies were mostly facilitators between founders and investors, but also made or have previously made their own pre-seed and seed investments.

The interviews had the following goals:

- To better understand how investment decisions are made.
- To investigate the use of machine learning in investment decision-making.
- To understand the role of gender in the startup environment

All VC companies are in Northern Europe, some with multiple locations, and three of them with a main business location in Copenhagen. The focus has been on companies/founders that are in the earliest stages of funding, mainly pre-seed and seed stage.

6.1.1 HOW ARE INVESTMENT DECISIONS MADE?

All VC investors interviewed for the purpose of this paper report having a substantial focus on the founders themselves, using words such as having ‘enthusiasm’, ‘inner drive’, ‘grit’ and ‘spike’, as well as some sort of personal chemistry. Because they have little information on the company itself, and perhaps also the market if the product or service represent something new, which is in the nature of startups. Investments at this stage are made often before there are any customers, or even a finished version of a product or service. This means, as the interviewees all point out, that a lot of the decision rests on whether they ‘believe’ in the entrepreneurs. This confirms the research findings that a part of the decision of investing is based on a “gut feel”, not just facts and numbers. The older a company gets,

the more it is possible to base the decisions on facts and numbers, as the company has then been able to build up this sort of data to showcase.

Three VCs point out the importance of having a team of cofounders; if a startup does not have this initially (ie there is only a solo founder), they will either be rejected or told to find cofounders and then return. Up to four founders was said to be ideal.

The experience of the founders is also a major factor. This refers to experience in the industry their startup is in, as well as entrepreneurial experiences. Having previously succeeded (but also failed) is seen as a big advantage – one VC paraphrased a statistic saying that a founder on average succeeds on their 3.6th startup. Industry experience is also essential for investors.

All investors report having different tools and checklists available for the assessment of founder teams and the startups, but report not always using them and rather going with an overall gut feel together with discussions with colleagues.

6.1.2 ROLE OF GENDER IN ENTREPRENEURIAL FINANCING

All interviewees were clear on the need to have more female founders, and that they were actively working towards this. Two of the VCs had this as the primary business goal. However, when it came to discussing why the share of female founders is low (especially within VC funding) and what it takes to change it, opinions differed, and a few were hesitant to give their opinion. One investor's opinion was that there are already enough female founders, that female founders perform better, and that this only needs to be showcased to investors and society at large to increase the share of female founders getting VC funding. The reasons why this needs to be showcased, according to this investor, is that most investors are not used to female founders, as the stereotype is the male founder.

Another said that there simply are not enough female founders with ventures that are within the tech-industry and scalable. Additionally, that because of deep social structures, women and men are raised and behave differently, possibly making males better founders than females. According to this investor, this makes it unlikely that we will see equality in both number of founders and funding in the near future.

Another investor said one possible explanation of the low share of female founders, specifically in Denmark, could be structural as the rights for maternal leave are virtually absent, and a founder on leave cannot work while receiving financial support, meaning they must either fully abandon the startup for their leave, or finance the leave themselves.

Another similar theory presented for the low share of female founders was that the startup world is often presented as an "all or nothing" activity with work 24/7, which might appeal less to women do to the gender roles and responsibilities that still are prominent in our society today.

One investor who had themselves been a founder, reported being met with significant gender bias from investors when raising funds, where the claim was that they needed twice as many meetings to raise a quarter of the funds measured against an average male founder.

6.1.3 USE OF MACHINE LEARNING IN THE DECISION-MAKING PROCESS

Only one of the VC companies use machine learning techniques in their decision-making process. They have based their model on three larger datasets, two of them publicly available and one retrieved through a larger European VC environment. Their aim was to replace the work today done by junior analysts in VC companies of screening companies, that can take up to two weeks per company. In the models that they have created, the approach to fairness is mainly fairness through unawareness as they have not used protected attributes in their models. They had not performed any fairness metrics on the model, mainly because they do not have the information on the protected attributes to do so but acknowledge that the models were certainly biased as the input data is known to be biased. They are also collecting their own data to be able to create machine learning models on novel data. In this data they include collection of protected attributes (gender, ethnicity, age, disability status), so that they can better measure both performance and metrics, with the hypothesis that diversity in founding teams will lead to better performance and success of the startup. This data consists of 40 questions that the startups fill out, both qualitative and quantitative data.

One of the VCs used a technique for valuation which is based on five different questions that is answered by the entrepreneurs to give an estimated value, and the VC explained that they use this together with a human evaluation, and that the calculator is often a good indicator, although should never be used on its own. However, it is not based on systematic historical data, or use any form of machine learning. Another tool they reported using is the Teamstowork[6], a Danish software company developed by psychologist Peter Neville. This is not based on machine learning techniques, but it is a tool said to be based on historical

[6] https://www.teamstowork.com

data that will give an evaluation of the teams who use it, presenting a performance of the team, based on 81 questions. The remaining VCs used primarily own experience and intuition in their decision-making process.

6.2 METHOD DESCRIPTION: MACHINE LEARNING MODEL

To the author's best knowledge, there are no open-source machine learning models for early-stage decision-making. Thus, a machine learning model needs to be created to perform experiments to view the effect of gender bias mitigation strategies. The goal of this model is not to create the best possible decision-making model for early-stage investing, but to create a model that is similar to previous works in the domain, that is as close to real-life decision-making that is obtainable with this data, and that can be used to measure the effect of various machine learning techniques on fairness measures reflecting gender bias. The process will be described in the following sections. The code and the database created will be made available through Github[7] to enable further work in the same domain.
After extracting data, cleaning, and creating the database, AIF360[8] is used as the main tool for building the models as well as testing the techniques (experiments described in section 6.3).

6.2.1 DATA SELECTION

Following advice from Ferrati and Muffatto (2021) the most relevant open database is Crunchbase as it provides free researcher access. The data used in this paper stems from Crunchbase, collected in 2019. Crunchbase is a web-based platform for both founders and investors where information about companies, founders and investors is found[9]. Previous research on machine learning in early-stage investment decision making has mostly used Crunchbase data (Arroyo et al., 2019; Bento, 2016; Huang, 2016, Xiang et al., 2012). See section 3.1.2 for details of these works. Additionally, Crunchbase data is used by one of the interviewees from the VC field, the only company interviewed that was using machine learning in their investment decision-making process.

---

[7] Available from June 15, 2022, through
https://github.com/marlenedah/genderbias_entrepreneurialfinancing

[8] https://github.com/Trusted-AI/AIF360

[9] https://www.crunchbase.com

6.2.2 DATA INFORMATION AND FILTERING

Only companies founded between 1995 and 2015 are included. Companies founded after 2015 could still be in the initial phase and are not included. Only those companies with a primary role of 'company' were included, where the other class is 'investors'. All data that does not contain binary information about gender has been excluded as most of the fairness metrics require knowledge of the protected attribute. Most companies have founders that are listed with unknown gender. These are excluded unless there are other founders where the gender is known.

There are a total of 668,123 organizations in the dataset before pre-processing and filtering. After preprocessing and filtering on company type (excluding investment companies) and founding year (1995-2015), there are 18,467 companies that contains information including founder's gender (male/female), with less than 10 founders. It is a skewed dataset with female founders in the minority class, as well as being imbalanced in terms of the outcome class. Of the 18.467 companies, 10,5% (1.948 companies) are with female founders. The remaining 16.519 companies have a majority of male founders. There are 5.038 in the negative class marked 'closed' (label 0) and the remaining 13.429 are companies marked with either 'acquired' or 'ipo'. This means the dataset is heavily skewed towards the positive class. The two graphs below illustrated the skewed distribution in both the founder of the gender, but also in the target variables.

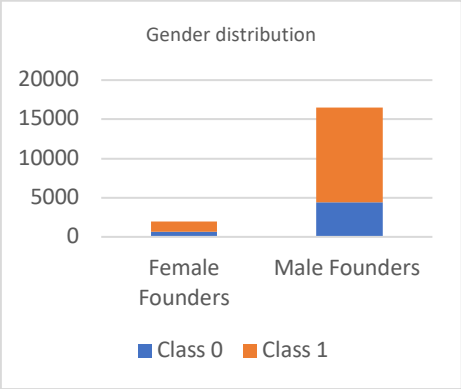

Figure 1: Gender distribution and target classes

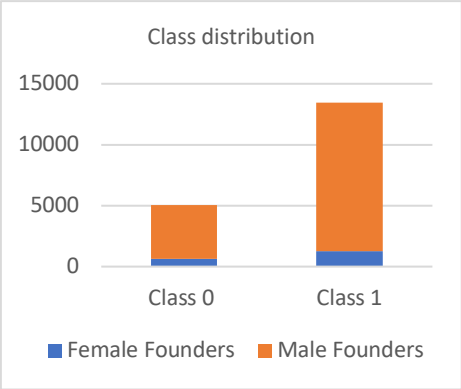

Figure 2: Target class distribution and gender

6.2.2.1 FEATURE EXTRACTION

Most of the Crunchbase data includes information that would not have been available to investors at an early stage, but rather only at a time when it is known that the company has been successful. Previous research from Bento (2016), Huang (2016), Krishna (2016) and Pan et al. (2018) explored in section 3.1.2, includes such features. They contain information on VC companies, funding rounds, funding amounts, number of customers, number of jobs and so on. This creates a look-ahead bias if used (Zbikowski & Antousiuk, 2021).

As Zbikowski & Antosiouk (2021) use a similar dataset from Crunchbase and have in their work used only variables that will not give a look-ahead bias, and this paper will use similar methods for feature extraction. This means only using information about the companies that would have been available to the investors at the time of the investment. This excludes most of the information that can be gathered from the complete database.

This leaves the following data features:

Feature name	Type	Short description
Companies		
company_name	Text	Name of the company
category_list	Text, comma separated	A list of categories describing the company
category_group_list	Text, comma separated	A list of categories describing the company
region	Text	Which region the company was founded in
city	Text	Which city the company was founded in
founded_on	Date	The date on which the company was founded
status	Categorical: closed, operating, acquired, ipo	The most recent status of the company.
Founders		
gender	Binary: male/female	The gender of the founder
is_completed	Binary: true/false	Whether the founder has completed a certain education
started_on	Date	The starting date of the education. Multiple educations possible
completed_on	Date	The completion date of the education. Multiple educations possible
multiple_degrees	Binary: true/false	Information on whether the founder has multiple degrees

Table 1: Features from original data

The region and city could create a geographic bias where the major start-up hubs are favorized. To adjust for this, a ranking of both city and region is created, where the share of the successful companies from each city and region replaces the names. From the founders table, new features are created. The first is work experience before founding the company, calculated in number of days. This is calculated by deducting the founding date with the date the education was completed. Both the total and the average of this is calculated and appended to the company data, in the case where there are multiple founders. The next feature created is the sum of founders with multiple degrees for each company, as well as the average of founders with multiple degree per company. The same is done for the feature is_completed. Next, the total number of male and female founders per company is created, and a new feature called mostly_male_founders is created, which is a true/false variable. The category features are split by comma to have one category per column, where only the three first in both 'category' features is used.

The final data as the following features:

Feature number	Feature name	Short description
1	city_success_ranking	The ranking of the city in terms of number of successes divided with total number of companies
2	region_success_rank	The ranking of the region in terms of number of successes divided with total number of companies
3	multiple_degrees_sum	The sum of founders with multiple degrees per company
4	multiple_degrees_average	The sum of founders with multiple degrees per company
5	Is_completed_sum	The average of founders with completed degrees per company
6	Is_completed_avg	The average of founders with completed degrees per company
7	work_experience_avg	The average number of days of work experience of the founders
8	work_experience_sum	The total number of days of work experience of the founders
9	education_time_avg	The average education time of the founders
10	education_time_sum	The sum of education time of the founders in number of days
11	male_founders	Number of male founders per company
12	female_founders	Number of female founders per company
13	unknown_founders	Number of founders of unknown gender
14	total_num_founders	Total number of founders per company
15	category_list1	First category word describing company
16	category_list2	Second category word describing company
17	category_list3	Third category word describing company
18	category_group_list1	First category word describing company group/sector
19	category_group_list2	Second category word describing company group/sector
20	category_group_list3	Third category word describing company group/sector
21	mostly_male_founders	True if there are strictly more male founders. Used as protected attribute, not in model.
Target	status	Target variable. The status of the company, 0 if it is closed, 1 if acquired/ipo

Table 2: Features used in model

Finally, all variables are one-hot encoded. The organization_uuid kept as index. MinMaxScaler is applied to scale the features, which scale each feature to a range on the training set of (0,1).

6.2.3 CHOICE OF TARGET VARIABLE

The target variable is a binary variable of 0 or 1, where 0 is encoded from the companies marked as ‘closed’ in the dataset. Label 1 is encoded from the companies marked as either ‘acquired’ or ‘ipo’. The largest class in the original dataset is ‘operating’, which has been removed from the dataset. It is removed mainly due to the uncertainty of whether these companies have ended up being acquired, gone public or closed. This distinction is similar to Pan et al. (2018). With the three other classes the result is known.

6.2.4 CLASSIFIER

Logistic Regression classifier is used as the main classifier for the model. Logistic Regression is a probabilistic classifier and uses a nonlinear separator. It is often used for classification problems, computing a weighted sum of the input features and outputs the logistic of the result as a number between 0 and 1 (Han et al., 2012). As seen in section 3.1.2, this is one of the most common classifiers used when creating machine learning models for entrepreneurial decision-making, in addition to Random Forest and Support Vector Machine (Ferrati & Muffato, 2021).

6.2.5 VALIDATION AND EVALUATION METRICS

There will be two types of evaluation metrics used when using this base model for further experiments and evaluation: fairness metrics and accuracy metrics. The fairness metrics are used as an indicator to show whether there is any gender bias found in the output of the machine learning model, as well as in the original data. Accuracy metrics are used to keep track of the accuracy/fairness trade-off.

6.2.5.1 FAIRNESS METRICS

The fairness metrics are described in detail in previous section 4.2: Average Odds Difference, Disparate Impact, Statistical Parity Difference (mean difference), and Equal Opportunity Difference. Two metrics are also possible to use on the original test and training dataset: Statistical Parity Difference and Disparate Impact.

6.2.5.2 ACCURACY METRICS

This paper will report on two types of accuracies: overall accuracy and balanced accuracy. Let's start with the confusion matrix showing some useful relationships between actual values in a test set vs the predicted classes of a test set.

		Predicted class	
Actual class		Predicted positive	Predicted Negative
	Actually positive	True Positives (TP)	False Negative (FN)
	Actually negative	False Positive (FP)	True Negative (TN)

Table 3: Confusion matrix

The **overall accuracy** is straightforward: the number of accurate predictions divided by the total number of predictions made. From the confusion matrix of true positives (TP), true negatives (TN), false positives (FP) and true negatives (TN) it can be calculated by:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{number\ of\ correct\ predictions}{total\ number\ of\ predictions}$$

When classes of target variables are imbalanced, as it in this context, the accuracy is not necessarily a good metric. If the model wants to optimize accuracy, it could just predict the negative class for all instances in the test set and achieve a high accuracy. Nonetheless, it is a commonly used metrics as can be seen from previous works on both machine learning in entrepreneurial finance and in fairness mitigation strategies for machine learning such as to measure a accuracy/fairness trade-off. It should not be used on its own, as it cannot give a full picture of how good the model is.

The balanced accuracy is a better metric to use when target classes are imbalanced. Balanced accuracy is also called AUC (area under the ROC Curve). It measures the model's ability to avoid false classifications (FN and FP). In the context of this paper, this means seeking to avoid classifications of startups as non-successful if they are indeed successful and avoid classifying startups as successful when they turn out not to be successful.

It is calculated by:

$$Balanced\ accuracy = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

6.2.5.3 VALIDATION

The data is separated into three sets: training (75%), testing (20%) and validation (5%).

6.3 METHOD DESCRIPTION: EXPERIMENTS WITH GENDER BIAS MITIGATION

The following sections will experiment with different machine learning methods to find how they affect gender bias measured by fairness metrics previously described. There are four experiments in all. The first experiment will investigate whether fairness through unawareness is a good approach to mitigate gender bias and achieve fairness by the defined metrics. The second experiment will test the bias mitigation technique called Reweighing described in section 5.4.1, where different weights are assigned so that gender is not considered as a variable. The third experiment tests different sampling strategies of gender class variables, to see how that affects fairness metrics. All experiments are conducted using the same data from the previous section 6.2, where the model is manipulated in different ways that will be further described for each experiment.

6.3.1 EXPERIMENT 1 DESCRIPTION

The goal of the first experiment is to check whether there is a difference in fairness and accuracy metrics when gender variables (protected attributes) are included in the data, compared to if they are excluded. From section 3.3.3 we know that protected attributes are not recommended to include as part of the features for training data, although it is often a good idea to collect them and include them in the pipeline so that it is possible to measure the bias against the groups or individuals. Excluding the protected attributes is called fairness through unawareness. In some cases, this improves fairness metrics and thus decrease bias, but in other cases it makes no difference. First, the metrics that are possible to obtain pre-training are measured. Then, a model is trained where gender variables are included (number of female founders and number of male founders), and tested on the test set, with both accuracy and a range of fairness metrics measured on the predictions the model made.

6.3.2 EXPERIMENT 1 RESULTS

The table below shows the metrics of Statistical Parity Difference and Disparate Impact on the data prior. The metrics is reported both on the training and test datasets.

	Dataset metrics	
	Train	Test
SPD	-0.07	0.08
DI	0.90	0.90

Table 4: SPD and DI of the dataset

	Accuracy	Balanced accuracy	AOD	DI	SPD	EOD
With gender attributes	0.77	0.70	-0.11	0.80	-0.15	-0.13
Without gender attributes (baseline)	0.77	0.70	-0.11	0.81	-0.15	-0.13

Table 5: Results from experiment 1

As can be seen from table 5 above, there is no difference by any metrics when including or removing gender as a variable in the model. There is no difference when measuring the input data or when measuring the predicted outcomes. These metrics will be used as a baseline for the next experiments, with the gender variables excluded in the further experiments. There is a decrease in SPD and DI from the prior metrics (original data).
An overview of feature importance is also included. The features below are the feature importance from the coefficients of the first Logistic Regression model, where all variables of gender are used. The feature *male_founders*, which is a count of the number of male founders for each company, is the fourth most important feature in the model. It is also interesting to note the other most important features. On top three most important in predicting success (status = 1), is the rank of the region the startup operates in, the sum of the education time and the rank of the city. The weights on predicting the non-successful companies (status = 0) are not as strong, but the sum of work experience is the most important one, followed by whether the company has a majority of male founders or not.

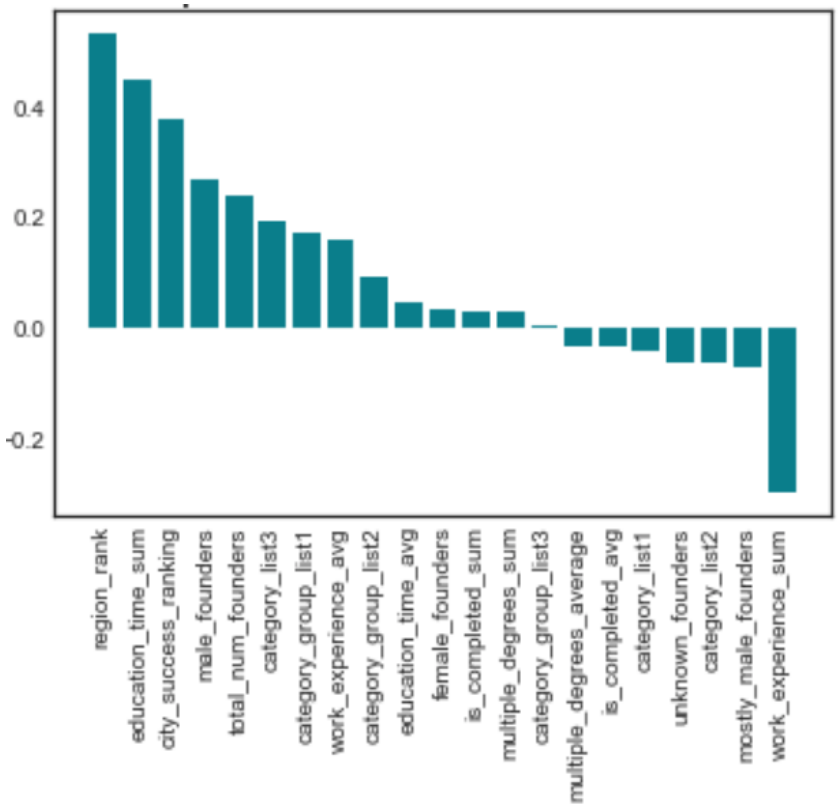

Figure 3: Feature importance from coefficients with gender variables

	Attribute	Importance
1	region_rank	0.529701
9	education_time_sum	0.446541
0	city_success_ranking	0.377787
10	male_founders	0.269142
13	total_num_founders	0.239432
16	category_list3	0.193423
17	category_group_list1	0.171851
6	work_experience_avg	0.157257
18	category_group_list2	0.092245
8	education_time_avg	0.045954
11	female_founders	0.032784
4	is_completed_sum	0.029396
2	multiple_degrees_sum	0.029396
19	category_group_list3	0.001296
3	multiple_degrees_average	-0.033782
5	is_completed_avg	-0.033782
14	category_list1	-0.042289
12	unknown_founders	-0.063786
15	category_list2	-0.063828
20	mostly_male_founders	-0.071942
7	work_experience_sum	-0.299272

Figure 4: Feature importance from coefficients as ordered list

6.3.3 EXPERIMENT 2 DESCRIPTION

In experiment 2, the effect of reweighing will be explored. Reweighing is explained in depth in section 5.1.4. To perform the reweighing on the data in Python, the *aif360.algorithms.preprocessing.reweighing* has been applied. As previously described, reweighing can be expected to have a significant result on increasing fairness metrics, without having too much of an effect on accuracy.

6.3.4 EXPERIMENT 2 RESULTS

The table below shows the results of experiment 2 on different accuracy and fairness metrics.

	Accuracy	Balanced accuracy	AOD	DI	SPD	EOD
Baseline	0.77	0.70	-0.11	0.81	-0.15	-0.13
Reweighing	0.77	0.70	0.01	0.94	-0.05	-0.04

Table 6: Results from experiment 2

As seen in the table above, all fairness metrics show a significant gender bias reduction. The reweighing has thus made this a model that is to be considered fair on most of the metrics, even more so than with the baseline model results. There is no measured effect on accuracy or balanced accuracy.

6.3.5 EXPERIMENT 3 DESCRIPTION

In the third experiment, both over- and under-sampling of the minority class will be explored with different methods. The minority class in this case will not refer to the outcome class, but to the variable of *mostly_male_founders*, referring to whether a company has strictly more male founders or not. The class of *mostly_male_founders* = 0 will be oversampled in the first part using SMOTE, then using a random over-sampler from *imblearn*. Both methods are explained in section 5.1.3. Next, the class of *mostly_male_founders* = 1 will be under-sampled, using random under-sampler from *imblearn*. Finally, a combination of the two will be explored: first the minority class (*mostly_female_founders* = 0) will be oversampled with a random over-sampler, then the majority class will be under-sampled. Over-sampling will be using *sample_strategy* = 0.3 and under-sampling will be with *sample_strategy* = 0.5, meaning that we first over-sample

the number of instances in the minority class by 30%, then under-sample the majority class by 50%. This will be attempted both with and without reweighing (as shown in experiment 2) in all instances.

There are two methods for oversampling, both will be attempted, and both are described in section 5.1.3. In the combination, only the *randomoversampler* will be used.

The various sampling methods give the following distribution.

	Number of companies with mostly male founders	Number of companies with mostly female founders
Original	16.519	1.948
Oversampling	16.519	16.519
Undersampling	1.948	1.948
Combination	9.910	4.955

Table 7: The number of companies by gender in different sample strategies

6.3.6 EXPERIMENT 3 RESULTS

The results from the experiment are shown in the table below.

	Accuracy	Balanced accuracy	AOD	DI	SPD	EOD
Baseline	0.77	0.70	-0.11	0.81	-0.15	-0.13
SMOTE Oversampling						
No reweighing	0.72	0.71	-0.37	0.44	-0.46	-0.41
Reweighing	0.72	0.69	0.02	0.89	-0.07	-0.12
Random Oversampling						
No reweighing	0.74	0.71	-0.14	0.77	-0.17	-0.14
Reweighing	0.74	0.71	0.02	0.98	-0.01	0.01
Random Undersampling						
No reweighing	0.69	0.69	-0.07	0.83	-0.10	-0.08
Reweighing	0.67	0.68	0.02	0.95	-0.03	-0.02
Random Combination Sampling						
No reweighing	0.74	0.72	-0.1	0.78	-0.15	-0.12
Reweighing	0.73	0.72	0.04	0.99	-0.01	0.02

Table 8: Results from experiment 3

From the table we can read that the worst strategy in terms of accuracy was under-sampling, and the best strategy was either random oversampling with reweighing or a combination with reweighing. The best strategy seems to be combination sampling or over-sampling, with reweighing in both cases. None achieve better accuracy than the baseline.

The worst strategy in terms of the fairness metrics was using SMOTE for over-sampling and not reweighing. This had significant negative results on all fairness metrics. With reweighing the results were much better. Not combining sampling with reweighing makes sampling an unfavorable strategy in terms of mitigating gender bias, as all the methods show worse results in fairness metrics compared to the baseline.

6.3.7 EXPERIMENT 4 DESCRIPTION

As explained in section 5.1.3 over-sampling is commonly used as a strategy when the target variables are imbalanced, as it is in the case of this dataset. If gender bias mitigation was not an important aspect, but rather accuracy, precision and recall, over-sampling would be a likely strategy to use to improve results of predicting the minority target variable. From the investor's perspective, it is not only important to be able to pick the successes, but also to avoid the non-successful startups. Thus, the next experiment will see how a combination of over-sampling of the minority target variable (non-successful startups) and under-sampling of the majority of the target variable will affect both accuracy metrics as well as fairness metrics. Based on the research found, gender bias in this context has not been a focus more than an approach of fairness through unawareness. It is thus interesting to check whether using this technique in the search of better accuracy metrics could impair or improve fairness metrics.

6.3.8 EXPERIMENT 4 RESULTS

The table below shows the most successful combination in terms of accuracy, along with fairness metrics, both with and without reweighing.

	Accuracy	Balanced accuracy	AOD	DI	SPD	EOD
Baseline	0.77	0.70	-0.11	0.81	-0.15	-0.13
Combination sampling of target variable						
No reweighing	0.74	0.74	-0.26	0.4	-0.32	-0.32
Reweighing	0.73	0.73	0.01	0.87	-0.07	0.1

Table 9: Results from experiment 4

The accuracy is not better than the baseline. However, the balanced accuracy is better than the baseline. The fairness metrics before reweighing is amongst the worst seen in the experiments so far. Reweighing improves metrics.

6.4 SUMMARY AND COMPARISON OF RESULTS ALL EXPERIMENTS

The table below gives a comparison of the results from the four experiments.

	Accuracy	Balanced accuracy	AOD	DI	SPD	EOD
Experiment 1						
Without gender attributes (baseline)	**0.77**	0.70	-0.11	0.80	-0.15	-0.13
With gender attributes	**0.77**	0.70	-0.11	0.81	-0.15	-0.13
Experiment 2						
Reweighing	**0.77**	0.70	**0.01**	0.94	-0.05	-0.04
Experiment 3						
SMOTE Oversampling						
No reweighing	0.72	0.71	-0.37	0.44	-0.46	-0.41
Reweighing	0.72	0.69	0.02	0.89	-0.07	-0.12
Random Oversampling						
No reweighing	0.74	0.71	-0.14	0.77	-0.17	-0.14
Reweighing	0.74	0.71	0.02	0.98	-0.01	**0.01**
Random Undersampling						
No reweighing	0.69	0.69	-0.07	0.83	-0.10	-0.08
Reweighing	0.67	0.68	0.02	0.95	-0.03	-0.02
Random Combination Sampling						
No reweighing	0.74	0.72	-0.10	0.78	-0.15	-0.12
Reweighing	0.73	0.72	0.04	**0.99**	**-0.01**	0.02
Experiment 4						
Combination sampling of target variable						
No reweighing	0.74	**0.74**	-0.26	0.40	-0.32	-0.32
Reweighing	0.73	0.73	**0.01**	0.87	-0.07	0.10

Table 10: Overview results all experiments

From the results, across the fairness metrics, reweighing is a highly successful technique for this dataset. The drawback of the technique, as has been mentioned in section 5.1.4, is that gender attributes must be available. Even better results in terms of fairness were found when a random sampling of the gender attribute is paired with reweighing. In this case, the balanced accuracy is improved from the baseline, while the overall accuracy is lower by 4 percentage points. All sampling strategies give better results in terms of fairness metrics from baseline, when combined with reweighing. When not combined with reweighing, fairness is in all cases except with random under-sampling is impaired from baseline. Particularly detrimental to fairness is the oversampling using SMOTE when no reweighing is applied.

7 DISCUSSION

The motivation for this research was to investigate how and if we can mitigate gender bias found in entrepreneurial financing when data from this field is used to create machine learning models. Female founders have been, and perhaps still are, disadvantaged when seeking funding from investors in early stages of a venture. It is important that the increasing number of machine learning models created to assist entrepreneurial financing decisions have a clear understanding of which types of bias that might be programmed into the model. This section will first look at some of the data limitations that are found in the data used for creating the model for this paper. Next it will discuss some of the successful bias mitigation techniques found. Third, there will be discussion of protected attributes, which is relevant both when using historical data, but also in the collection of future data. The fourth subsection will discuss fairness metrics in the context of gender bias in entrepreneurial financing and machine learning. As there is no previous work to base the use of fairness metrics, this paper discusses which ones could be relevant. Finally, real-life applications and relevance of work conducted in this paper will be discussed.

7.1 DISCUSSION OF DATA LIMITATIONS

Ideally, the data used should be more realistic, such as from a specific VC company. Unfortunately, this is both difficult as it is proprietary and valuable data, but also because many VC companies do not yet have this type of data made available. This will be discussed further in the next section.

The perhaps biggest limitation of the data is that the information it contains is quite far away from the information investors would use in a real-life situation, according to both the literature (Krishna et al., 2016; Ferrati & Muffato, 2021) and the VC investor interviews (section 6.1). There are some aspects that might be overlapping such as education time, time since education, and number of founded companies. However, there is little to no information on how the founder present themselves, how their personality is, and so on. It is not possible to measure whether they have the ‘inner drive’ or the ‘grit’ that investors mentioned in interviews for this paper, using the Crunchbase data. Neither is it possible to measure any details on the venture, the size or relevance of the market, whether it is a new product and so on. Further, a VC investor is making the decision to invest, they will often receive information about the startup through various channels. This could be an interview

with the founders, viewing a pitch or reviewing a pitch deck, reviewing some sort of application from the startup, looking at the startup's website, hearing about the startup from someone in their network, etc. Thus, the features used in this machine learning model, is unlikely to capture the decision-making points from real VC decisions. VC companies/investors often focus their efforts or analysis on smaller parts of the markets through their own market selections, rather than large ones such as crawling the entire database of Crunchbase as many of the discussed ML models do.

When creating machine learning models to assist decision-making we must at the same time address all the human bias that is found in the data, and not only focus on the outcomes of the machine learning models (Feuerriegel & Schwabe, 2020: Mehrabi et al., 2022). In the context of entrepreneurial financing, it is an ecosystem of education, capital, location, and overall possibilities given that makes a founder or a venture successful. The gender bias from the literature overview in section 3.2 makes it clear that male and female founders are not on equal footing, even in similar or identical ventures, and the literature show that historical data on funding favors male founders although the reasons are unclear (Kanze et al., 2018). Funding is decisive for whether a company can grow and be successful. Thus, one limitation of the chosen target variable is that successful startup companies need funding to succeed. The target variable only views companies that have succeeded or not. When female founders are less likely to get funded, or receive lower funding amounts, they will also have a harder time succeeding with their venture. Another possible target variable could be to predict which companies receive a VC investment or not. This could be used by a VC company who would like to create a model in which they make investments based on their own (and others') previous investments. Another limitation related to the choice of target variable, as Ferrati and Muffato (2021) mention, only including the companies that have been acquired or registered for IPO is not necessarily correct as a measure of startup success. Companies that have not been acquired or gone public might be successful in terms of other measures such as profit or growth. Additionally, from the investors' perspective, the acquisition or IPO might not have been a financial success.

In the context of startup investment decision-making, we are interested in predicting success to know which companies to invest in. Preferably, an investor would like to know as early as possible, as the purchase price of a startup will increase as they show signs of success. However, what defines success for a startup is not necessarily easily defined,

especially in the context of machine learning as there are availability constraints in the data, and in Crunchbase data as is commonly used (Ferrati & Muffato, 2021).

Another limitation of Crunchbase data is that it has a high number of missing data (Xiang et al., 2012), as is seen in the chosen dataset. As mentioned by Xiang et al. (2012) in their study of Crunchbase data, it is likely that successful companies have fewer missing data, as users are more likely to provide data on more successful companies. Information might be available in the dataset simply because successful companies are more likely to have people fill this out at a later stage.

7.2 DISCUSSION OF PROTECTED ATTRIBUTES

Experiment 1 results show the need to go beyond the approach that is commonly used in machine learning models, namely fairness through unawareness. Although it might intuitively seem a fair approach it does not have any effect on fairness metrics. This is also important to note when it comes to the models created in previous works mentioned in section 3.1.2, as many are using Crunchbase data. These do not explicitly measure fairness metrics or any gender variables or implement any fairness reduction techniques.

Another important aspect of collecting protected attributes, is shown through experiment 4. A possible step in the model pipeline with the aim of improving balanced accuracy, would be to use combination sampling of the target variable to balance out the successes and non-successes. As seen from the results, the combination sampling negatively impacts fairness, with DI at a low 0.4. If gender was an unknown attribute, it would not be possible to measure fairness or to perform reweighing, and one would be likely to use this method without knowing the effects on fairness.

Overall, none of the experiments, or measures of fairness, would be possible with unknown gender attributes. Haeri and Zweig (2020) argue that protected attributes should be included in data collection so that fairness can be measured. Based on the results of excluding them in this paper, where fairness metrics were unchanged, one could argue that it does not matter. However, using another dataset could yield different results, and without the fairness metrics, one cannot know. It is also relevant to track the development over time, which makes it necessary to collect the protected attributes, although it is not necessary to use as prediction features.

7.3 DISCUSSION OF FAIRNESS METRICS

"Without accounting for agents′ responses to a fairness requirement, any discussion of its implications is of limited usefulness" (Fu et al., 2020).

The discussion of which fairness metrics to use in this field is important, as it reflects what a fair outcome should be, and to some extent it could give an indication of when a fair outcome is reached – although on its own the metrics are not enough. The fairness metrics used in this paper have been explained in previous sections, but there is no literature or agreement in the industry as to which metrics should be used to measure when a fair outcome is reached. In a fairness assessment of AI in the financial industry, Zhang and Zhou (2019) chooses the following three metrics: Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD) and Disparate Impact (DI). They highlight EOD as an important metric of bias in the context of the financial institutions.

In the context of this paper, male founders are defined as the privileged group, while female founders are defined as the unprivileged group. A favorable outcome is obtaining an investment from a VC company, or, relatedly, having success as a startup in the sense of being acquired or reaching an IPO.
Disparate impact is relevant as it refers to the unintentional bias in a selection process.
As can be seen from the metrics of DI and SPD essentially measures the same, that the possibility of a positive prediction is the same between groups. It could thus be argued that if SPD = 0 then there is no disparate impact on the unprivileged group, and we have achieved statistical parity. If male and female founders have different probabilities of achieving an investment from a VC, the SPD will reflect this. SPD is not sufficient as a metric to ensure fairness on its own and can be especially flawed in data where there is little data on the favorable outcomes of a group (Dwork et al., 2011), as is mostly the case in the context of start-up investments. One of the major critiques of SP is that it does not consider individual qualifications. If there are more qualified male founders, then SP would require a model to reject qualified individuals from that group and instead approve female founders. However, as suggested by data scientist Cortez (2019), SP is a good metric if there is a desire to see more of the unprivileged groups make their way to a positive outcome or if there are known historical biases that may have affected the data as it is today. This can arguably be said to be the case of female founders in entrepreneurial financing.

Using SPD as a target fairness metric in a machine learning model would essentially force the VC to invest in the same number of women as men. It is not given that a VC company would agree that there are enough qualified female founders to justify statistical parity as a target metric. However, if VC companies are of the opinion that men and women are in general equally qualified as entrepreneurs, that there are not any inherent gender differences that makes one gender more qualified than the other, then SPD or DI is an ideal target metric.

EOD consider the true positive rates rather than just the positive outcomes. In the case of entrepreneurial finance, one example of fair EOD could be that the probability of receiving an investment from a VC should be the same for female founders who later have success, that it is for male founders who later have success. In machine learning terms it means that the true positives for both genders should be equal for the outcomes to be considered fair by EOD. This might seem fairer at first glance. However, it will then be based on historic success. We know that funding is important to succeed, and some of the ways female founders have been disadvantage have been made clear in section 3.2. Only relying on the EOD as a fairness metric might make sure that female founders are not more disadvantaged than today but is not likely to help correct the consequences of historic bias.

The topic of this paper is gender bias mitigation in entrepreneurial finance machine learning. The fairness metrics used must consider that to the early-stage investor, one major goal is to get a return on their investment, meaning that they would like to 'pick winners' and avoid the non-successful companies. The fairness/accuracy trade-off should not be too large. Thus, accuracy and in particular balanced accuracy with a special focus on precision is likely to be an important focus in the development of machine learning models in the industry. Perhaps the most important type to avoid for investors are the false positives. In this case, the investor might risk investing both time and financial resources in a company that end up as non-successful.

7.4 DISCUSSION OF APPLICATIONS AND RELEVANCE

"Machine learning should not be considered a black-box solution able to replace human experience but rather a valuable tool that can support practitioners in their decision-making process" (Ferrati and Muffato, 2021).

Machine learning models should not be created without regard for human interaction and influence. It needs to be recognized beyond the fairness metrics that data contains bias and ignoring that or using fairness through unawareness is not enough. Human decision makers put much trust in the outcome of AI systems, even more than they trust the outcomes of human-decision making, and despite not fully understanding the mechanisms behind the AI systems' outcome (Logg et al., 2019; Keding & Meissner, 2021). Dellermann et al. (2019) suggest that, instead of the sometime gloomy prediction that AI will supersede human intelligence and take over our jobs, the more likely paradigm to come is one of 'Hybrid Intelligence', where human and machine intelligence are complementary and together perform better than each would on their own. In the context of entrepreneurial finance and gender bias, an approach that also considers the human decision-makers appears most relevant. Machine learning models can at one hand help reduce the reliance on 'gut feel' that many investors report using in their decision-making. On the other hand, it is unlikely that reliance on a model alone will convince investors to invest in a startup venture. In the interviews with the VC investors, many mentioned that they had tools such as checklists to help evaluate founder teams, however, that they would often rely more on the overall gut feel than the tools.

Although there is little research on machine learning models in entrepreneurial finance, this does not mean that companies have not made their own models that are not openly available. Transparency (public) in a proprietary algorithm is unlikely to happen as it for many businesses will be a core business asset, and as pointed out by Fu et al. (2020) this is one of the biggest issues when it comes to identifying algorithmic bias.

When it comes to gender bias, we need to separate discrimination and gender bias.

The difference can be seen more clearly in studies such as Malmström et al. (2017): Discrimination is illegal in Sweden, but this does not keep the governmental VC investors in displaying a gender bias when they systematically describe male and female founders using different terms, with these descriptions ultimately affecting funding decisions. An understanding of the difference might be important to combat gender bias. The impression from the interviews is that it is difficult to talk about the female disadvantage in funding decisions. The investors interviewed are certain that they do not discriminate and that they want to invest in female founders, and that they actively seek this. However, as one investor said: they believe that there are gender differences that might contribute to female founders being less like the wanted entrepreneurial stereotype. It is important to understand that one can certainly have gender bias in judgments of people, without direct discrimination.

8 FUTURE WORK

To address some of the limitations discussed, some suggestions are made for future work. The first is to consider different data sources and modalities. Features that can be extracted from Crunchbase data does not necessarily resemble how entrepreneurial financing decisions are made, but the models created by VC companies is perhaps more likely to be based on the data they themselves collect. Many of the studies conducted on gender bias in entrepreneurial financing have revolved around language; the language used by investors when evaluating entrepreneurs, or the language used by entrepreneurs when presenting their companies in various formats (Balachandra et al., 2021; Edelman et al., 2018; Huang 2020; Johansson et al., 2021; Kanze et al., 2018; Malmström et al., 2017). Machine learning models created to assist investment decision-making are likely to use this type of data, especially as natural language processing (NLP) is getting increasingly useful and advanced. Thus, future work could focus on how to mitigate the gender bias within NLP-based models, potentially using data that is shown to exhibit a high level of gender bias such as Malmström (2017).

In combination with the above-mentioned data sources and modalities, it could be interesting to test more of the techniques described throughout section 5. This could also be interesting to test using the same Crunchbase data, but with a different target variable such as 'investor funding achieved'. Future work could also seek to test more models on the same data and techniques, to compare the effects on the fairness metrics.

Experiment 4 shows how a commonly used machine learning method could affect results on fairness. This could possibly be the case with other commonly used machine learning methods not tested in this paper and should be investigated further. This would help machine learning practitioners to know which techniques they should be more careful about using, also when the gender attributes are not known, as it might ultimately increase the gender funding gap.

9 CONCLUSION

This paper is a helpful starting point for anyone attempting to create a machine learning model for decision-making assistance in entrepreneurial finance, where there is an interest in mitigating gender bias in the model, both in real-life applications and in academia. The paper has given a thorough a theoretical background in the three domains of: 1) machine learning for entrepreneurial financing decisions in an early stage, 2) gender bias in early-stage investment decision making, and 3) gender bias in machine learning models. The paper gives useful contributions to each domain and build the early foundations of being its own domain. The theoretical background gave the basis for some important assumptions that the paper further builds on: that machine learning is a useful tool for early-stage investment decision-making assistance and its use is increasing, that there is gender bias found with investors when they make decisions regarding early-stage investments, and finally that machine learning models will copy and amplify bias that it identifies in its data.

Early-stage investors have been interviewed for this paper to give a context useful for applications in real-life settings. A machine learning model for early-stage investment decision-making assistance was created using data from Crunchbase and a Logistic Regression classifier, with protected attribute of gender available from the data. This model was then used to perform experiments using different machine learning techniques and registering how these techniques affect gender bias as measured through four different fairness metrics: Disparate Impact, Statistical Parity Difference, Equal Opportunity Difference and Average Odds Difference. Reweighing was found to be an effective technique across the fairness metrics and was also effective in combination with other techniques such as random combination sampling of the gender classes.

10 REFERENCES

Ahn, Y., & Lin, Y.-R. (2019). FairSight: Visual Analytics for Fairness in Decision Making. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 1086–1095. https://doi.org/10.1109/tvcg.2019.2934262

Ajunwa, I. (2020). The "black box" at work. *Big Data & Society*, *7*(2), 205395172096618. https://doi.org/10.1177/2053951720938093

Alsos, G. A., & Ljunggren, E. (2016). The Role of Gender in Entrepreneur-Investor Relationships: A Signaling Theory Approach. *Entrepreneurship Theory and Practice*, *41*(4), 567–590. https://doi.org/10.1111/etap.12226

Arroyo, J., Corea, F., Jimenez-Diaz, G., & Recio-Garcia, J. A. (2019). Assessment of Machine Learning Performance for Decision Support in Venture Capital Investments. *IEEE Access*, *7*, 124233–124243. https://doi.org/10.1109/ACCESS.2019.2938659

Bai, S., & Zhao, Y. (2021). Startup Investment Decision Support: Application of Venture Capital Scorecards Using Machine Learning Approaches. *Systems*, *9*(3), 55. https://doi.org/10.3390/systems9030055

Balachandra, L., Briggs, T., Eddleston, K., & Brush, C. (2017). Don't Pitch Like a Girl!: How Gender Stereotypes Influence Investor Decisions. *Entrepreneurship Theory and Practice*, *43*(1), 116–137. https://doi.org/10.1177/1042258717728028

Balachandra, L., Fischer, K., & Brush, C. (2021). Do (women's) words matter? The influence of gendered language in entrepreneurial pitching. *Journal of Business Venturing Insights*, *15*, e00224. https://doi.org/10.1016/j.jbvi.2021.e00224

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *104 California Law Review*, *104*(3), 671–732. https://doi.org/10.2139/ssrn.2477899

Bavey, N., Messel, T., Jessen, H., Schuyler, S., Di Fonzo, A., Lundqvist, M., & Renoldi, M. (2021). *The Startup Funding Report*. Unconventional Ventures. https://report2021.unconventional.vc

Bellamy, R. K. E., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., Zhang, Y., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., & Mehta, S. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4:1–4:15. https://doi.org/10.1147/jrd.2019.2942287

Besse, P., del Barrio, E., Gordaliza, P., Loubes, J.-M., & Risser, L. (2020). A survey of bias in Machine Learning through the prism of Statistical Parity for the Adult Data Set. In *arXiv:2003.14263 [cs, stat]*. https://arxiv.org/abs/2003.14263v2

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020). *Fairlearn: A toolkit for assessing and improving fairness in AI*. Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *ArXiv:1607.06520 [Cs, Stat]*. https://arxiv.org/abs/1607.06520

Brooks, A. W., Huang, L., Kearney, S. W., & Murray, F. E. (2014). Investors prefer entrepreneurial ventures pitched by attractive men. *Proceedings of the National Academy of Sciences*, *111*(12), 4427–4431. https://doi.org/10.1073/pnas.1321202111

Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building Classifiers with Independency Constraints. *2009 IEEE International Conference on Data Mining Workshops*, 13–18. https://doi.org/10.1109/icdmw.2009.83

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K., & Varshney, K. (2017). Optimized Pre-Processing for Discrimination Prevention. *Advances in Neural Information Processing Systems*, 3995–4004. https://facctconference.org/static/tutorials/calmon_preprocess18.pdf

Carli, L. L. (2010). Having it All: Women with Successful Careers and Families. *Sex Roles*, *62*(9-10), 696–698. https://doi.org/10.1007/s11199-009-9719-0

Chakraborty, J., Majumder, S., & Menzies, T. (2021). Bias in machine learning software: why? how? what to do? *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. https://doi.org/10.1145/3468264.3468537

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chouldechova, A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *FATML*. https://arxiv.org/abs/1610.07524

Cortez, V. (2019, September 24). *How to define fairness to detect and prevent discriminatory outcomes in Machine Learning*. https://towardsdatascience.com/how-to-define-fairness-to-detect-and-prevent-discriminatory-outcomes-in-machine-learning-ef23fd408ef2

Crawford, K. (2016, June 25). Opinion | Artificial Intelligence's White Guy Problem. *The New York Times*. https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

Croce, A., Martí, J., & Murtinu, S. (2012). The Impact of Venture Capital on the Productivity Growth of European Entrepreneurial Firms: "Screening" or "Value added" Effect?. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1705225

Crunchbase. (n.d.). *Crunchbase - Crunchbase Company Profile & Funding*. Crunchbase. Retrieved March 11, 2022, from https://www.crunchbase.com/organization/crunchbase

Dastin, J. (2018, October 10). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters; Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business & Information Systems Engineering*, *61*(5), 637–643. https://doi.org/10.1007/s12599-019-00595-2

d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification. *Big Data*, *5*(2), 120–134. https://doi.org/10.1089/big.2016.0048

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). Fairness Through Awareness. *ArXiv:1104.3913 [Cs]*. https://arxiv.org/abs/1104.3913v2

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*(3), 573–598. https://doi.org/10.1037//0033-295x.109.3.573

Edelman, L. F., Donnelly, R., Manolova, T., & Brush, C. G. (2018). Gender stereotypes in the angel investment process. *International Journal of Gender and Entrepreneurship*, *10*(2), 134–157. https://doi.org/10.1108/ijge-12-2017-0078

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

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and Removing Disparate Impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2783258.2783311

Feldman, T., & Peake, A. (2021). End-To-End Bias Mitigation: Removing Gender Bias in Deep Learning. *ArXiv:2104.02532 [Cs]*. https://arxiv.org/abs/2104.02532

Ferrati, F., & Muffatto, M. (2021). Entrepreneurial Finance: Emerging Approaches Using Machine Learning and Big Data. *Foundations and Trends® in Entrepreneurship*, *17*(3), 232–329. https://doi.org/10.1561/0300000099

Feuerriegel, S., Dolata, M., & Schwabe, G. (2020). Fair AI. *Business & Information Systems Engineering*, *62*(4). https://doi.org/10.1007/s12599-020-00650-3

Foulds, J. R., Islam, R., Keya, K. N., & Pan, S. (2019). An intersectional definition of fairness. *ArXiv Preprint*. http://arxiv.org/abs/1807.08362

Fu, R., Y. Huang, and P.V. Singh. (2020). Artificial Intelligence and Algorithmic Bias: Source, Detection, Mitigation, and Implications. *In Pushing the Boundaries: Frontiers in Impactful OR/OM Research*, 39-63. Informs.

Gartner. (2021, March 10). *Gartner Says Tech Investors Will Prioritize Data Science and Artificial Intelligence Above "Gut Feel" for Investment Decisions By 2025*. Gartner. https://www.gartner.com/en/newsroom/press-releases/2021-03-10-gartner-says-tech-investors-will-prioritize-data-science-and-artificial-intelligence-above-gut-feel-for-investment-decisions-by-20250

Gompers, P. A., Gornall, W., Kaplan, S. N., & Strebulaev, I. A. (2020). How do venture capitalists make decisions? *Journal of Financial Economics*, *135*(1), 169–190. https://doi.org/10.1016/j.jfineco.2019.06.011

Haeri, M. A., & Zweig, K. A. (2020). The Crucial Role of Sensitive Attributes in Fair Classification. *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. https://doi.org/10.1109/ssci47803.2020.9308585

Han, J. (2012). *DATA MINING : concepts and techniques.* (3rd ed.). Morgan Kaufmann.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *NIPS*. 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain. https://arxiv.org/abs/1610.02413

Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *ArXiv:1805.03677 [Cs]*. https://arxiv.org/abs/1805.03677

Holstein, K., Vaughan, J.W., Daumé, H., Dudik, M., and Wallach, H. (2019). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA, Paper 600, 1–16. DOI:https://doi.org/10.1145/3290605.3300830

Hooker, S. (2021). Moving beyond "algorithmic bias is a data problem." *Patterns*, *2*(4), 100241. https://doi.org/10.1016/j.patter.2021.100241

Huang, B.G. (2016). Predict startup success using network analysis and machine learning techniques. (CS224W: Machine Learning with Graphs, Fall 2016, Stanford University).

Huang, J., & Zhan, S. (2015). With a Little Help of My (Former) Employer: Past Employment and Entrepreneurs' External Financing. *Academy of Management Proceedings*, *2015*(1), 12050. https://doi.org/10.5465/ambpp.2015.12050abstract

Huang, L. (2018). The Role of Investor Gut Feel in Managing Complexity and Extreme Risk. *Academy of Management Journal*, *61*(5), 1821–1847. https://doi.org/10.5465/amj.2016.1009

Huang, L., Joshi, P., Wakslak, C., & Wu, A. (2020). Sizing Up Entrepreneurial Potential: Gender Differences in Communication and Investor Perceptions of Long-Term Growth and Scalability. *Academy of Management Journal*, *64*(3). https://doi.org/10.5465/amj.2018.1417

Johansson, J., Malmström, M., Lahti, T., & Wincent, J. (2021). Oh, it's complex to see women here, isn't it and this seems to take all my attention! A repertory grid approach to capture venture capitalists cognitive structures when evaluating women entrepreneurs. *Journal of Business Venturing Insights*, *15*, e00218. https://doi.org/10.1016/j.jbvi.2020.e00218

Johnson, M. A., Stevenson, R. M., & Letwin, C. R. (2018). A woman's place is in the… startup! Crowdfunder judgments, implicit bias, and the stereotype content model. *Journal of Business Venturing*, *33*(6), 813–831. https://doi.org/10.1016/j.jbusvent.2018.04.003

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer. *Machine Learning and Knowledge Discovery in Databases*, 35–50. https://doi.org/10.1007/978-3-642-33486-3_3

Kamishima, T., Akaho, S., & Sakuma, J. (2011). Fairness-aware Learning through Regularization Approach. *2011 IEEE 11th International Conference on Data Mining Workshops*. https://doi.org/10.1109/icdmw.2011.83

Kanze, D., Huang, L., Conley, M. A., & Higgins, E. T. (2018). We Ask Men to Win and Women Not to Lose: Closing the Gender Gap in Startup Funding. *Academy of Management Journal*, *61*(2), 586–614. https://doi.org/10.5465/amj.2016.1215

Keding, C., & Meissner, P. (2021). Managerial overreliance on AI-augmented decision-making processes: How the use of AI-based advisory systems shapes choice behavior in R&D investment decisions. *Technological Forecasting and Social Change*, *171*(8), 120970. https://doi.org/10.1016/j.techfore.2021.120970

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. In P. Christos H. (Ed.), *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*.

Kohavi, R., & Becker, B. (1996). *UCI adult data set*. UCI Meachine Learning Repository. https://archive.ics.uci.edu/ml/datasets/adult

Kortum, S., & Lerner, J. (2000). Assessing the Contribution of Venture Capital to Innovation. *The RAND Journal of Economics*, *31*(4), 674. https://doi.org/10.2307/2696354

Krishna, A., Agrawal, A., & Choudhary, A. (2016, December 1). *Predicting the Outcome of Startups: Less Failure, More Success*. IEEE Xplore. https://doi.org/10.1109/ICDMW.2016.0118

Leavy, S. (2018). Gender bias in artificial intelligence. *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering - GE '18*. https://doi.org/10.1145/3195570.3195580

Leavy, S., Meaney, G., Wade, K., & Greene, D. (2020). Mitigating Gender Bias in Machine Learning Data Sets. *Communications in Computer and Information Science*, *1245*, 12–26. https://doi.org/10.1007/978-3-030-52485-2_2

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Maalouf, M. (2011). Logistic regression in data analysis: an overview. *International Journal of Data Analysis Techniques and Strategies*, *3*(3), 281. https://doi.org/10.1504/ijdats.2011.041335

Malmström, M., Johansson, J., & Wincent, J. (2017). Gender Stereotypes and Venture Support Decisions: How Governmental Venture Capitalists Socially Construct Entrepreneurs' Potential. *Entrepreneurship Theory and Practice*, *41*(5), 833–860. https://doi.org/10.1111/etap.12275

Malmström, M., Voitkane, A., Johansson, J., & Wincent, J. (2020). What do they think and what do they say? Gender bias, entrepreneurial attitude in writing and venture capitalists' funding decisions. *Journal of Business Venturing Insights*, *13*, e00154. https://doi.org/10.1016/j.jbvi.2019.e00154

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. *ArXiv:1908.09635 [Cs]*. https://arxiv.org/abs/1908.09635

M. Lichman. (2013). UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

Nadeem, A., Abedin, B., & Marjanovic, O. (2020). Gender Bias in AI: A Review of Contributing Factors and Mitigating Strategies. *ACIS 2020 Proceedings*. https://aisel.aisnet.org/acis2020/27

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., & Broelemann, K. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining and Knowledge Discovery*, *10*(3). https://doi.org/10.1002/widm.1356

Obschonka, M., & Audretsch, D. B. (2019). Artificial intelligence and big data in entrepreneurship: a new era has begun. *Small Business Economics*, *19*(1). https://doi.org/10.1007/s11187-019-00202-4

Omidi, B. (2021). *Enhancing Fairness in Supervised Machine Learning* [Master thesis]. http://dx.doi.org/10.20381/ruor-26258

O'Sullivan, L. (2021, August 14). *How the law got it wrong with Apple Card*. TechCrunch. https://tcrn.ch/3yLUrdS

Pan, C., Gao, Y., & Luo, Y. (2018). *Machine Learning Prediction of Companies' Business Success*. https://cs229.stanford.edu/proj2018/report/88.pdf

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. (2017). On Fairness and Calibration. *Conference on Neural Information Processing Systems*.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., & Ghani, R. (2019). Aequitas: A Bias and Fairness Audit Toolkit. In *arXiv:1811.05577 [cs]*. https://arxiv.org/abs/1811.05577

Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., & Liu, Y. (2019). How Do Fairness Definitions Fare? *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3306618.3314248

Schneider, A., Tinsley, C., Cheldelin, S., & Amanatullah, E. (2010). Likeability v. Competence: The Impossible Choice Faced by Female Politicians, Attenuated by Lawyers. *Faculty Publications*, 529. https://scholarship.law.marquette.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1528&context=facpub

Smith, G., & Rustagi, I. (2021). When Good Algorithms Go Sexist: Why and How to Advance AI Gender Equity. *Stanford Social Innovation Review*. https://doi.org/10.48558/A179-B138

Smith, G., & Rustagi, I. (2021, March 21). When Good Algorithms Go Sexist: Why and How to Advance AI Gender Equity. *Stanford Social Innovation Review*. https://ssir.org/articles/entry/when_good_algorithms_go_sexist_why_and_how_to_advance_ai_gender_equity

Spence, M. (2002). Signaling in Retrospect and the Informational Structure of Markets. *The American Economic Review*, *92*(3), 434–459. http://www.jstor.org/stable/3083350

Ünal, C. (2019). *Searching for a Unicorn: A Machine Learning Approach Towards Startup Success Prediction* [Master thesis]. http://dx.doi.org/10.18452/20347

Żbikowski, K., & Antosiuk, P. (2021). A machine learning, bias-free approach for predicting business success using Crunchbase data. *Information Processing & Management*, *58*(4), 102555. https://doi.org/10.1016/j.ipm.2021.102555

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, *4*(2), 205395171774353. https://doi.org/10.1177/2053951717743530

Veer, T. H., & Bringmann, K. (2020). Everything is (Not) Negotiable: The Gender Startup Valuation Gap. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3738519

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *FairWare@ICSE* (pp. 1–7). ACM.

Xiang, G., Zheng, Z., Wen, M., Hong, J., Rose, C., & Liu, C. (2012). A Supervised Approach to Predict Company Acquisition with Factual and Topic Features Using Profiles and News Articles on TechCrunch. *Sixth International AAAI Conference on Weblogs and Social Media*. https://www.researchgate.net/publication/266224153_A_Supervised_Approach_to_Predict_Company_Acquisition_with_Factual_and_Topic_Features_Using_Profiles_and_News_Articles_on_TechCrunch

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning Fair Representations. *Proceedings of Machine Learning Research*, *28*(3), 325–333. http://proceedings.mlr.press/v28/zemel13.html

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *ArXiv:1801.07593 [Cs]*. https://arxiv.org/abs/1801.07593

Zhang, Y., & Zhou, L. (2019). Fairness Assessment for Artificial Intelligence in Financial Industry. *33rd Conference on Neural Information Processing Systems*. (NeurIPS 2019), Vancouver, Canada. https://arxiv.org/pdf/1912.07211.pdf

Zhao, H., & Gordon, G. J. (2022). Inherent Tradeoffs in Learning Fair Representations. *Journal of Machine Learning Research*, *23*(57), 1–26. http://jmlr.org/papers/v23/21-1427.html

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2017). *Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints*. ArXiv.org. https://arxiv.org/abs/1707.09457

Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics*, *10*(5), 593. https://doi.org/10.3390/electronics10050593