Mitigating Discriminatory Biases in Success Prediction Models for Venture Capitals

Yiea-Funk Te[1], Michèle Wieland[1], Martin Frey, Helmut Grabner
ZHAW Zurich University of Applied Sciences, School of Engineering, Switzerland
{yfunk.te}@gmail.com
{michele.wieland, martin.frey, helmut.grabner}@zhaw.ch

***Abstract*—The fairness of machine learning-based decision support systems has become a critical issue, also in the field of predicting the success of venture capital investment startups. Inappropriate allocation of venture capital, fueled by discriminatory biases, can lead to missed investment opportunities and poor investment decisions. Despite numerous studies that have addressed the prevalence of biases in venture capital allocation and decision support models, few have addressed the importance of incorporating fairness into the modeling process. In this study, we leverage invariant feature representation learning to develop a startup success prediction model using Crunchbase data, while satisfying group fairness. Our results show that discriminatory bias can be significantly reduced with minimal impact on model performance. Additionally, we demonstrate the versatility of our approach by mitigating multiple biases simultaneously. This work highlights the significance of addressing fairness in decision-support models to ensure equitable outcomes in venture capital investments.**

***Index Terms*—model fairness, gradient reversal, venture capital, success modeling**

I. Introduction

Venture capitalists (VCs) play a crucial role in providing funding to private companies by managing a pool of capital and acting as financial intermediaries. They connect investors with entrepreneurs holding promising venture ideas, often serving as the only source of capital for startups. Despite this important role, identifying promising startups remains a challenge for VCs as the screening process relies heavily on subjective assessments and is prone to human error. In addition, the startup screening process in traditional investment practices is very time-consuming due to the complexity of the assessment process.

Machine learning has seen a significant upsurge in data-driven investment approaches [1]. Numerous studies have been conducted to better understand and predict the success of a startup by leveraging machine learning. In particular, several studies have been conducted to build predictive models with accuracy consistently exceeding 85% based on large datasets from Crunchbase which contain extensive information about the company, founding team, investor, and funding [2]–[5].

However, these models are subject to demographic biases that tend to overestimate the success of male founders, founding teams with higher education, and graduates of prestigious universities, while underestimating the success of non-white founders. Moreover, these models are biased toward startups from economically developed countries, such as the United States, Germany, the United Kingdom, and China. The lack of model generalization is likely due to the under-representation of minority populations in the Crunchbase datasets, e.g., fewer female entrepreneurs in the tech industry or fewer startups from countries with lower levels of entrepreneurial activity.

Biases in the underlying machine learning models lead to discrimination against certain groups of founders and startups. This can lead to an unfair distribution of venture capital, resulting not only in missed investment opportunities, but also in poor investment decisions with fatal consequences for VCs [6].

Previous works towards fairness in the VC industry focused on the identification of discriminatory biases [7], [8]. Surprisingly, there has been little effort in mitigating these biases when creating success prediction models for VCs. Simple approaches such as removing the attributes driving the biases before a model is trained will still result in a biased model because the correlations with the omitted attributes are not eliminated from the dataset [9].

This paper therefore investigates the mitigation of discriminatory biases during the modeling of startup success. Gradient Reversal Layer by Ganin et al. [10] is applied to effectively learn fair feature representations before a binary classification model for success is trained. We compare our models to a baseline model that does not incorporate fairness (BL1) and another baseline model that attempts to account for fairness by omitting attributes related to the biases under study (BL2). We show that our models are substantially fairer while providing a nearly equivalent level of accuracy.

The rest of the paper is structured as follows. Section II provides an overview about biases in venture capital and approaches for achieving fairness. Section III outlines the methodology employed in the study. Section IV demonstrates selected experiments and finally, section V and VI discuss the results and conclusions of the study.

II. Related work

A. Biases in venture capital

There has been a great deal of effort to examine biases in venture capital [7], [8]. In this paper, we discuss discriminatory

This work was funded by Innosuisse under Grant 53126.1 IP-ICT.
[1]Both authors have equally contributed.

This is the author accepted manuscript version of a conference paper published by IEEE. The Version of Record is available at IEEE *Xplore*: https://doi.org/10.1109/SDS57534.2023.00011

biases which are both well documented in the literature and are also present in the Crunchbase dataset to be investigated.

Kanze et al. [11] investigated the well known gender gap in startup funding and concluded that female entrepreneurs often position their startups to "playing not to lose" while male entrepreneurs position their startups to "playing to win", causing the gender gap in startup funding to perpetuate. Zhang [8] examined whether early-stage investors are biased toward gender, educational background, and certain racial minority groups of founders during the investment process. Zhang found that investors have implicit biases against female and Asian founders, especially when evaluating high-growth startups. Moreover, investors are more likely to establish the first contact with founders with a high level of education. Bengtsson & Hsu [12] studied whether there is racial and ethnic discrimination in crowdfunding. Using data from one of the largest crowdfunding platforms, they found that funding is lower on average for African Americans and that African Americans receive less funding from U.S. funders compared to whites.

A much debated bias related to the company itself is location [13]. Although venture capitalists claim that they do not intentionally distinguish between geographic regions in their investment decisions, studies have shown that venture capitalists actually invest predominantly in startups that are located in their own immediate region [14]. The phenomenon in which venture capitalists tend to invest predominantly in startups that are located near their own offices is also referred to as local bias. This bias is driven by information asymmetries and access to local networks. The presence of local bias has significant implications for the venture capital industry, such as missing out on opportunities to invest in promising startups located outside the immediate region and leading to inefficient allocation of resources.

B. Approaches for achieving fairness

Despite considerable interest in the ethical implications of biases in venture capital, little work exists describing the extent to which predictive models - developed to support investment due diligence - satisfy the requirements of fairness.

Datasets commonly used to model success predictions are often biased because they disproportionately represent demographic groups or startups from major countries, which can lead to performance disparities in trained models. A straightforward mitigation strategy is to balance samples from a dataset by adjusting sampling frequency or weight for each datapoint based on the proportion of its group in the dataset [15]. Another approach is "fairness by unawareness", in which sensitive attributes are omitted in the input and feature space such that the trained models cannot use the excluded attributes to make predictions [9]. However, omitting an input variable may not always eliminate the bias in the model, as the bias can still be inferred from other proxy variables. In general, bias mitigation algorithms can be categorized into three techniques, namely pre-processing, in-processing, and post-processing, depending on their application timing [16]. In-processing methods differ from pre-processing and post-processing techniques as they directly incorporate fairness during model optimization. This means that even with biased data as input, the model can attain fairness [17].

Another line of work focuses on invariant feature representation learning to ensure model fairness. In particular, recent studies explored the gradient reversal approach in non-investment domains. In the context of credit risk modeling, Zheng et al. [18] proposed a framework for privacy-preserving risk modeling based on gradient reversal. They showed that latent representations can be effectively generated where sensitive information is obstructed, providing a solid foundation for privacy-aware machine learning for credit risk analysis. The gradient reversal approach for fairness was further explored by [19] for data that is imbalanced in both the distribution of outcomes and sensitive attributes.

III. Methodology

A. Fairness metrics

The definition of fairness in machine learning has attracted much attention. In particular, two different notions of fairness are widely used [20]: (1) group fairness, which requires similar treatment of the disadvantaged group and the advantaged group, and (2) individual fairness, which requires similar treatment of observations with similar characteristics.

In this paper, we use *equal opportunity* to evaluate group fairness and *consistency* for individual fairness. To simplify the following definitions we focus on a binary outcome. In the following definitions $\hat{y}$ corresponds to the predicted label, z to the sensitive attribute and y to the true label.

1) Equal opportunity: A classifier satisfies *equal opportunity* if the groups defined by the protected feature $z \in \{z_1, z_2\}$ achieve equal true positive rates. Formally, *equal opportunity* is defined as

$$P(\hat{y} = 1|z = z_1, y = 1) = P(\hat{y} = 1|z = z_2, y = 1) \quad (1)$$

Therefore, we aim to minimize the *equal opportunity gap*, which is defined as the absolute difference between the two groups

$$|P(\hat{y} = 1|z = z_1, y = 1) - P(\hat{y} = 1|z = z_2, y = 1)| \quad (2)$$

In the context of venture capital, this means that startups should have an equal chance of success regardless of their sensitive attributes, such as the gender or race of the founders.

2) Consistency: For each observation x_i, its prediction $\hat{y_i}$ is compared to the average of its k nearest neighbors, and the average of this score is reported over the entire dataset, containing N elements. Formally, *consistency* is defined as:

$$1 - \frac{1}{N}\sum_{i=1}^{N}\left|\hat{y_i} - \frac{1}{k}\sum_{j \in k-\mathrm{NN}(x_i)}^{N}\hat{y_j}\right| \quad (3)$$

Individual fairness is fully established if the *consistency* is equal to 1.

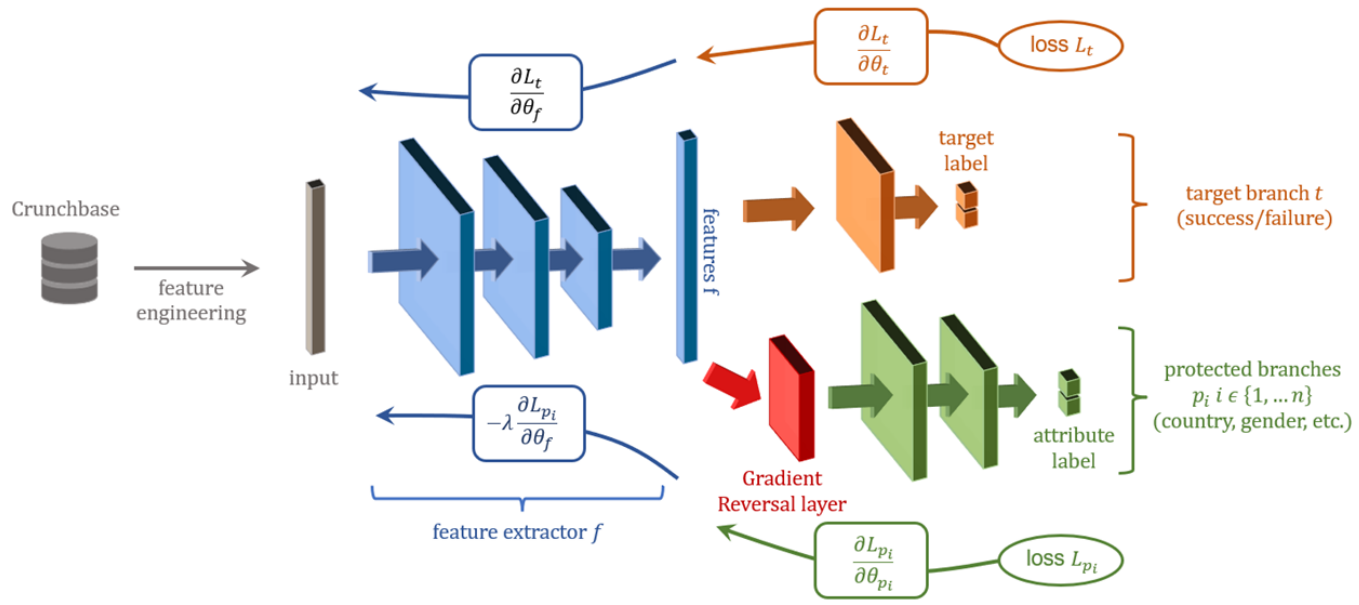


Fig. 1: Illustration of the applied neural network architecture. The feature extractor f contains three dense layers (128, 64, 32) to extract features from the input data. The network is then divided into two types of branches: target branch and attribute branch. The target branch t is responsible for predicting the success of a startup. The protected branch p_i is used to unlearn the prediction of the sensitive attribute by employing a Gradient Reseral layer. A protected branch is used for each sensitive attribute.

B. Gradient Reversal

We suggest utilizing an in-processing approach known as Gradient Reversal. The concept of a Gradient Reversal layer has been first introduced in domain adaptation [10] and since then generalized to fair classification [21], [22]. In essence, the domain adaptation approach proposes to learn representations of the data which are domain-invariant. In this setting, Ganin et al. [10] proposed a neural network model which is trained to optimize two training objectives at the same time by two sub-networks, one optimized to predict the labels and another to predict the domain. When backpropagating the domain gradients into the network, the gradient's sign is reversed, effectively removing domain-specific information from the feature representation shared by both networks. Ganin et al. show that this learning scheme is able to find a saddle point equilibrium between the two training objectives. By treating the sensitive attribute as the new domain, we can use the same approach to prevent the network from being biased by the sensitive attribute.

We employ Gradient Reversal to learn feature representations which are invariant to the sensitive attributes. Gradient Reversal allows us to simultaneously train the network to predict our target variable, while it aims to unlearn the prediction of the sensitive features. Figure 1 depicts our model architecture.

Our network consists of three fully connected layers (128, 64, 32) for feature extraction (feature extractor), which is then divided into several branches. The first branch is for the prediction of the target variables (target branch), while the other branches are used to unlearn the prediction of the sensitive attributes (protected branches). One branch is used for every sensitive attribute to be protected. Each protected branch consists of a Gradient Reversal layer with a scaling factor λ, followed by two fully connected layers (64, 32) which are then connected to a softmax layer before the attribute output. The parameter λ controls the influence of the target branch t and protected branch p responses on the total loss [23]. The target branch is directly connected to the feature extractor shared by all branches and consists only of a softmax layer before the target output.

The total loss of the whole network is calculated as the sum of the target variable $loss_t$ and the weighted losses for the n protected attributes $loss_{p_i}$, $i \in \{1...n\}$:

$$loss = loss_t + w_1 \cdot loss_{p_1} + ... + w_n \cdot loss_{p_n} \quad (4)$$

The coefficients $w_1, ..., w_n$ specify the weights for the losses of the protected branches and can be optimized according to task at hand.

IV. Experiments

A. The Dataset and Success Definition

We use a dataset from Crunchbase consisting of approximately 20,000 startups. The definition of success used by researchers and practitioners is not uniform. Various definitions have been used in studies that have attempted to

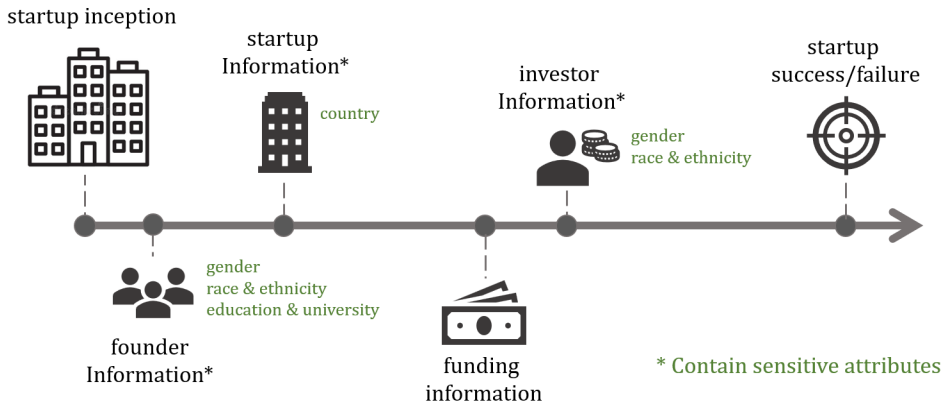


Fig. 2: Data selection for feature engineering follows the definition of feature engineering in [5]. Only data prior to the occurence of Series A funding is considered, including company, founder, and investment information. Features related to sensitive attributes are found in all information groups, e.g., location in company data and gender, ethnicity, and education in founder and investor data.

explain business success. Following [5], obtaining a Series A financing is used as the definition of business success because it is of great importance to VCs [24]. It represents a significant early investment into a startup and signals a successful demonstration of progress and a clear path to revenue growth. Moreover, obtaining a Series A financing is an objective measure that reflects the potential future business value of the startup. Therefore, the main objective of this study is to classify startups into successful and failed startups. Startups are defined as successful if they have received a Series A funding. The definition of a failed startup requires several additional steps in which the details of the previous funding as well as the company size were examined in more closely. For example, recently founded startups and those that cannot be conclusively categorized as either successful or failed are excluded from this study. More details on the definition can be found in [5]. Consequently, the dataset consists of 8,537 successful and 12,457 failed startups.

B. Sensitive Attributes

We consider country, gender, education, university, race and ethnicity as sensitive attributes, as discussed in section II-A. For country, we consider the origin of the startup's foundation. Country can be considered sensitive as it may reveal information related to the founder's race or religion. For gender, we consider whether a team consists of mixed genders, only female or only male founders. Education is evaluated based on the subjects studied (such as law, economics, or technology), as well as the level (BSc, MSc, PhD, etc.) and number of degrees achieved. University takes into account the reputation of the university according to QSWorld University Rankings [25]. For race, the dataset is partitioned into six groups: Asian, Black, Hispanic, Other, Unknown, and White. For ethnicity, we consider 19 groups including: African, Asian, British, Germanic, GreaterAfrican, IndianSubcontinent, Jewish, Muslim, Nordic, to name a few. Race and ethnicity are not explicitely available in the Crunchbase dataset. We therefore use a transformer-based model to derive race and ethnicity based on founder name [26]. It is crucial to acknowledge that this serves as an approximation and should not be regarded as an absolute truth.

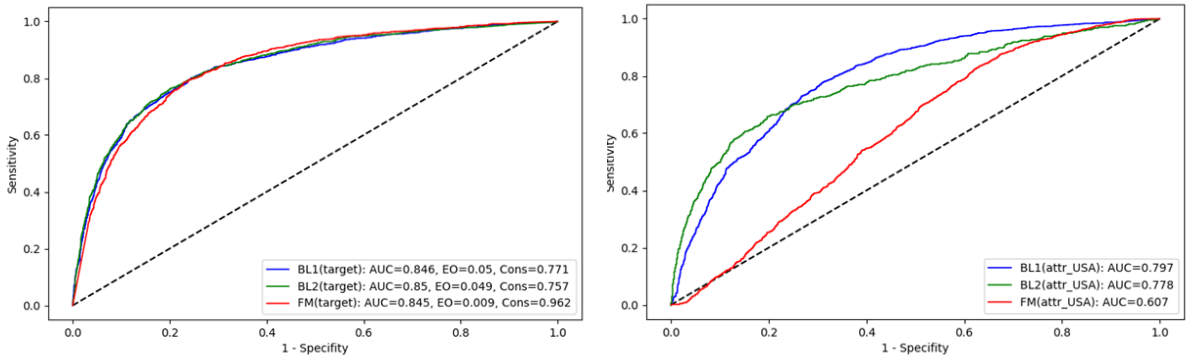
C. Feature Engineering

Fig. 2 illustrates the data selection process for feature engineering. For each startup, we first filter the historical data to include only the information that was available prior to the occurrence of a Series A financing (if available). Then, extensive feature engineering is performed to cover a wide range of factors influencing company success. In total, over 400 features are generated, which can be categorized into three macro groups according to [27]: features related to (1) the company, such as location and industry, (2) founder characteristics, including demographic information, education background, and prior work experience, and (3) investment factors, including funding details, investor demographic information, and investment track record. The created features include information related to the sensitive attributes, which are converted into binary features by one-hot encoding for the application of Gradient Reversal learning.

D. Fair models protecting a binary sensitive attribute

In this study, a comprehensive evaluation of our approach is conducted through a series of experiments. The objective of these experiments is to demonstrate the validity of our approach and provide guidance to practitioners.

First, we demonstrate through a simple example that omitting sensitive attributes during model training alone does not eliminate discriminatory biases. Instead, we find that explicit mitigation techniques such as the Gradient Reversal are necessary to effectively address such biases. To this end, two baseline models are considered. The first baseline model (BL1) incorporates the sensitive attributes in the dataset, while the second baseline model (BL2) discards them. A fair model (FM) is trained using Gradient Reversal to effectively remove



(a) BL1 includes sensitive attributes, while BL2 omits them. FM uses Gradient Reversal to mitigate local bias related to the origin of U.S. startups. The experiment indicates that the equal opportunity gap (EO) decreases for FM while its model performance remains stable.

(b) Re-training all three models to predict the sensitive attribute shows that BL1 and BL2 still implicitly contain the information related to the sensitive attribute, while FM suppresses it.

Fig. 3: Sample illustrations of experiments with one discriminatory attribute.

information related to the sensitive attributes from the feature representation, as illustrated in Fig. 1. To ensure a consistent model comparison, all three models use the same neural network architecture for the feature extractor f and target branch t. For the purpose of illustration, we demonstrate our findings using a single sensitive attribute based on the local bias, specifically whether a startup was founded in the United States or not.

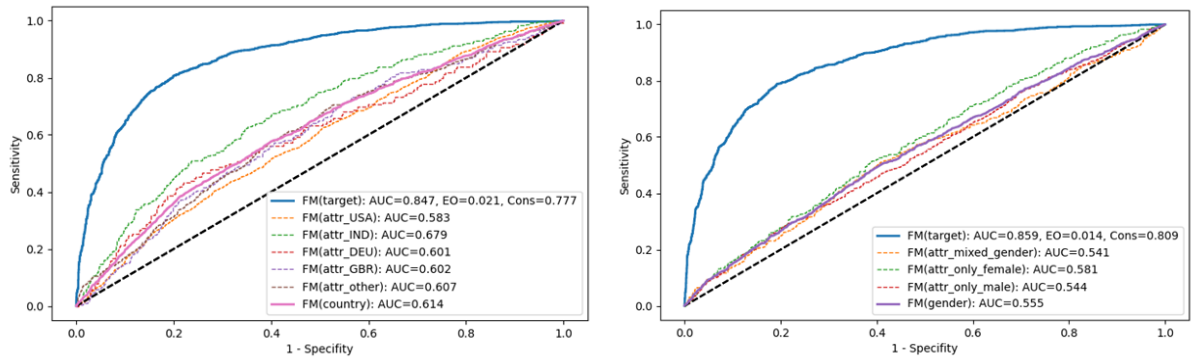
Fig. 3a depicts the AUC-ROC curve along with the auc score (AUC), equal opportunity gap (EO) and consistency (Cons) of BL1, BL2 and FM (i.e., de-biased with regards to whether the startup was founded in the United States or not). The results show that all three models exhibit comparable performance in terms of predicting the success of startups, as indicated by the similar AUC scores. However, it can be observed that the equal opportunity gap of the FM model is significantly lower than those of BL1 and BL2, while its consistency is significantly higher. This suggests that our fair model clearly outperforms both baseline models in terms of fairness while maintaining a comparable level of performance in predicting startup success.

To further corroborate our findings, a second experiment was conducted to emphasize the presence of latent information related to the sensitive attribute in the feature representation. This was achieved by retraining the target branch t to predict the sensitive attribute, while keeping the weights of the hidden layers in the feature extractor f fixed. The results of this experiment are depicted in Fig. 3b and reveal that both BL1 and BL2 are capable of accurately predicting the sensitive attribute, i.e., whether the startup is founded in the United States or not, with AUC scores of 79.7% and 77.8%, respectively. On the other hand, FM is only able to moderately predict the sensitive attribute, with an AUC score of 60.7%. These results indicate that even though the sensitive attribute was omitted during model training in BL2, the information related to it remains implicit in the feature representation, due to its correlation with other features such as the type of funding currency (e.g., seed financing conducted in U.S. dollars). The application of Gradient Reversal in FM effectively suppresses this information in the feature representation of the feature extractor f.

E. Fair models protecting a categorical sensitive attribute

In another experiment, we demonstrate the application of Gradient Reversal learning on a categorical sensitive attribute to evaluate model fairness. The startups in the dataset originate from over 15 countries, including the United States, Germany, the United Kingdom, and India, to name a few. To apply Gradient Reversal to the categorical country variable, a protected branch p for each country is created and attached to the feature extractor f. Therefore, one-hot encoding is applied to the country variable (thereafter called *country codes*). The fair model is trained by optimizing the parameters of the underlying feature mapping of the feature extractor f by minimizing the loss of the target branch t and to maximize the loss of all protected branches p simultaneously [10].

To report the fairness performance of the model, the EO and Cons are first calculated for each sensitive attribute and then the weighted average is obtained based on the sample size with respect to *country code*. The results are illustrated in Fig. 4a. In addition, the presence of latent information related to country codes in the feature extractor f is evaluated by retraining the target branch to predict the countries (dashed lines). The AUC-ROC curve for the country is calculated using the weighted average of the AUC-ROC curves of the each *country code*. The EO score suggests that our FM is relatively fair in terms of local bias (EO equal to 0 indicates complete fairness). The result is further supported by the finding that the retrained FM predicts the sensitive attributes poorly, with a mean AUC score of 0.614.



(a) This experiment examines the categorical sensitive attribute country, which includes five levels: United States, India, Germany, Great Britain and others.

(b) This experiment explores the categorical sensitive attribute gender, which contains three levels: mixed gender, only female, and only male.

Fig. 4: Two sample illustrations of Gradient Reversal on the categorical attribute. The blue line shows model performance for the target outcome. In addition, the presence of latent information related to the sensitive attribute in the feature extractor f is evaluated by retraining the target branch to predict the protected attributes (dashed lines). The AUC-ROC curve for the sensitive attribute is calculated using the weighted average of the AUC-ROC curves of each level of the categorical attribute (pink line).

Fig. 4b illustrates the results of the fair model with regards to the gender of the founders. Three protected branches are used for the binary sensitive features *mixed gender*, *only female* and *only male*.

F. Fair models protecting multiple sensitive attributes

The methodology explained in section IV-E can be extended such that multiple sensitive attributes can be protected simultaneously. This enables the construction of a model that is fair with respect to country, gender, education, university, race, and ethnicity. A comprehensive summary of the results from all experiments can be found in Table I and will be thoroughly discussed in section V. The approach demonstrated for binary and categorical attributes can be applied to continous attributes as well, by using a regression head with an appropriate loss (e.g. mean squared error) in the target branch.

V. Results

The results of the experiments are given in Table I. For all experiments, we conduct an extensive hyperparameter tuning for the scaling factor $\lambda \in [1, 100]$ and the loss weights $w_i \in [0.01, 1]$, $i \in \{1...n\}$, with a fixed set of hyperparameters for the employed neural network. The first two models (BL1 and BL2) are the baselines, with the BL1 model incorporating sensitive attributes during training and the BL2 model excluding these attributes. The remaining models (FM_country, FM_gender, etc.) are fair models that protect a specific sensitive attribute group during training, including country, gender, education, university, race, and ethnicity. The last model (FM) is a fair model that protects all sensitive attributes simultaneously. The performance metrics reported in the table include the area under the curve (AUC), accuracy, equal opportunity gap, and consistency.

A comparison of BL1 and BL2 shows that simply removing the sensitive attributes from the dataset slightly improves model fairness without significantly decreasing the predictive ability of the model. The group fairness of BL2 increased (i.e., the equal opportunity gap decreased for all attribute groups), whereas the individual fairness of BL2 decreased slightly (i.e., consistency decreased) compared with BL1.

The metrics of the fair models FM_* suggest that model fairness can be further improved by training models that aim to protect specific sensitive attribute groups without compromising predictive performance. For all FM_*, it can be shown that the equal opportunity gap decreases with respect to the protected attribute group, while the consistency remains more or less the same compared to BL1 and BL2. For example, the value for the equal opportunity gap with respect to country is 0.021 for FM_country, compared to 0.049 and 0.093 for BL1 and BL2, respectively. Moreover, to our surprise, AUC and accuracy of fair models can even increase when protecting specific attribute groups, such as the auc and accuracy scores reported for FM_race.

Furthermore, the results suggest that protecting a single sensitive attribute group leads to an increase in the equal opportunity gap for other (i.e., unprotected) attribute groups. As demonstrated in the case of FM_gender, where only gender is protected, the equal opportunity gap for gender decreases, but higher equal opportunity gaps are observed in comparison to BL1 and BL2 for the unprotected attribute groups, including country, education, and university. This trend is also evident when evaluating the average equal opportunity gap for all sensitive attributes in models where only one attribute group is protected (refer to Table V, column "All attributes"). However, if all sensitive attribute groups are protected simultaneously (FM), the equal opportunity gap for each attribute group

TABLE I: Summary of experimental outcomes. The first baseline model BL1 incorporates the sensitive attributes during training, whereas the second baseline model BL2 excludes them. The fair models FM_* are trained by protecting the respective attribute of focus, such as country, gender, etc. The fair model FM is trained by protecting all discriminatory biases simultaneously. Evaluation metrics including AUC, accuracy, equal opportunity gap, and consistency are reported. While equal opportunity gap is optimized at 0 (indicated by a downward pointing arrow), the other metrics are optimized at 1 (indicated by an upward pointing arrow). The best results are highlighted in bold, and the scores corresponding to the protected attribute are denoted in italics.

			Equal opportunity gap ↓							
Model	AUC ↑	Accuracy ↑	Country	Gender	Education	University	Race	Ethnicity	All attributes	Consistency ↑
BL 1	0.849	0.782	0.093	0.056	0.063	0.077	0.046	0.055	0.061	0.772
BL 2	0.839	0.777	0.049	0.050	0.054	0.055	0.050	0.038	0.041	0.769
FM_country	0.847	0.777	***0.021***	0.095	0.128	0.127	0.052	0.042	0.070	0.777
FM_gender	0.859	0.794	0.097	***0.014***	0.122	0.129	0.024	0.042	0.065	**0.809**
FM_education	0.861	0.795	0.092	0.113	***0.029***	**0.013**	0.025	0.048	0.052	0.779
FM_university	0.853	0.781	0.078	0.122	0.043	*0.017*	0.032	0.043	0.054	0.786
FM_race	**0.871**	**0.798**	0.108	0.138	0.079	0.097	*0.019*	0.050	0.071	0.779
FM_ethnicity	0.843	0.777	0.114	0.123	0.100	0.114	0.040	***0.027***	0.071	0.783
FM	0.827	0.761	0.028	0.033	0.040	0.041	**0.015**	0.038	***0.033***	0.781

can be reduced, leading to a reduction in the average equal opportunity gap and thus to an overall improvement of the model fairness.

In summary, our findings suggest that the explicit simultaneous protection of all sensitive attributes is the only reliable method for reducing discriminatory bias in the model. For FM, we did not observe a significant change in the AUC and accuracy of the target when compared to BL1 and BL2. The results were also similar for consistency, which measures individual fairness. Although our approach has an impact on group fairness, it does not necessarily affect individual fairness.

VI. Conclusions

In this work, we demonstrate the capabilities of Gradient Reversal learning to build startup success prediction models satisfying group fairness using high-dimensional Crunchbase data including multiple sensitive attributes. Inline with prior research, our findings suggest that it is not sufficient to simply remove sensitive attributes to ensure group fairness. Furthermore, we demonstrate that protecting individual sensitive attribute groups may worsen the group fairness of other sensitive attributes and thus, simultaneously protecting all sensitive attributes is the only safe way to increase the overall group fairness of the model. In addition, our results suggest that Gradient Reversal can be effectively applied to tackle model fairness with minimal adversarial effect on the model performance.

However, there are still some key challenges that need to be addressed. There exists a tradeoff between fairness and prediction accuracy. A completely fair model may not be able to make useful predictions. Our approach allows for balancing prediction accuracy and fairness by assigning different weights to the total loss (see Equation 4). Practioners must determine the acceptable level of standard accuracy and, therefore, the extent to which the model can be debiased.

Gradient Reversal can be used to improve group fairness, but it does not necessarily increase individual fairness i.e., consistency. To establish VC investor confidence in a success prediction model, it is important that similar startups are given similar predictions. Therefore, different approaches must be explored and evaluated to ensure individual fairness. These key challenges require further investigation in future research. Nevertheless, the methodology presented can serve as a basis for developing a predictive model for startup success that incorporates fairness considerations.

Acknowledgment

The authors would like to thank Penny Schiffer for her continued support, as well as business partner Raized.ai for supplying the data set utilized in this paper.

References

[1] C. M. Schmidt, "The impact of artificial intelligence on decision-making in venture capital firms," Ph.D. dissertation, 2019. [Online]. Available: http://hdl.handle.net/10400.14/29250

[2] K. Żbikowski and P. Antosiuk, "A machine learning, bias-free approach for predicting business success using crunchbase data," *Information Processing & Management*, vol. 58, no. 4, p. 102555, 2021.

[3] J. Arroyo, F. Corea, G. Jimenez-Diaz, and J. A. Recio-Garcia, "Assessment of machine learning performance for decision support in venture capital investments," *Ieee Access*, vol. 7, pp. 124 233–124 243, 2019.

[4] J. Li, "Prediction of the success of startup companies based on support vector machine and random forset," in *2020 2nd International Workshop on Artificial Intelligence and Education*, 2020, pp. 5–11.

[5] Y.-F. Te, M. Wieland, M. Frey, A. Pyatigorskaya, P. Schiffer, and H. Grabner, "Making it into a successful series a funding: An analysis of crunchbase and linkedin data," *Available at SSRN 4217648*.

[6] C. Jain, "Artificial intelligence in venture capital industry: opportunities and risks," Ph.D. dissertation, Massachusetts Institute of Technology, 2018. [Online]. Available: http://hdl.handle.net/1721.1/118544

[7] S. A. Zahera and R. Bansal, "Do investors exhibit behavioral biases in investment decision making? a systematic review," *Qualitative Research in Financial Markets*, 2018.

[8] Y. Zhang, "Discrimination in the venture capital industry: Evidence from two randomized controlled trials," *arXiv preprint arXiv:2010.16084*, 2020.

[9] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 560–568. [Online]. Available: http://doi.acm.org/10.1145/1401890.1401959

[10] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016. [Online]. Available: https://www.jmlr.org/papers/volume17/15-239/15-239.pdf

[11] D. Kanze, L. Huang, M. A. Conley, and E. T. Higgins, "We ask men to win and women not to lose: Closing the gender gap in startup funding," *Academy of Management Journal*, vol. 61, no. 2, pp. 586–614, 2018. [Online]. Available: https://doi.org/10.5465/amj.2016.1215

[12] O. Bengtsson and D. H. Hsu, "Ethnic matching in the us venture capital market," *Journal of Business Venturing*, vol. 30, no. 2, pp. 338–354, 2015. [Online]. Available: https://doi.org/10.1016/j.jbusvent.2014.09.001

[13] J. D. Coval and T. J. Moskowitz, "Home bias at home: Local equity preference in domestic portfolios," *The Journal of Finance*, vol. 54, no. 6, pp. 2045–2073, 1999. [Online]. Available: https://doi.org/10.1111/0022-1082.00181

[14] D. Cumming and N. Dai, "Local bias in venture capital investments," *Journal of empirical finance*, vol. 17, no. 3, pp. 362–380, 2010. [Online]. Available: https://doi.org/10.1016/j.jempfin.2009.11.001

[15] D. Cao, X. Zhu, X. Huang, J. Guo, and Z. Lei, "Domain balancing: Face recognition on long-tailed domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5671–5679.

[16] B. d'Alessandro, C. O'Neil, and T. LaGatta, "Conscientious classification: A data scientist's guide to discrimination-aware classification," *Big Data*, vol. 5, no. 2, pp. 120–134, jun 2017.

[17] M. Wan, D. Zha, N. Liu, and N. Zou, "Modeling techniques for machine learning fairness: A survey," 2022.

[18] Y. Zheng, Z. Wu, Y. Yuan, T. Chen, and Z. Wang, "Pcal: A privacy-preserving intelligent credit risk modeling framework based on adversarial learning," *arXiv preprint arXiv:2010.02529*, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2010.02529

[19] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi, "Data decisions and theoretical implications when adversarially learning fair representations," *arXiv preprint arXiv:1707.00075*, 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1707.00075

[20] S. Barocas, M. Hardt, and A. Narayanan, "Fairness in machine learning," *Nips tutorial*, vol. 1, p. 2, 2017.

[21] D. McNamara, C. S. Ong, and R. C. Williamson, "Provably fair representations," *arXiv preprint arXiv:1710.04394*, 2017.

[22] Q. Xie, Z. Dai, Y. Du, E. Hovy, and G. Neubig, "Controllable invariance through adversarial feature learning," *Advances in neural information processing systems*, vol. 30, 2017.

[23] J. M. Clavijo, P. Glaysher, J. Jitsev, and J. M. Katzy, "Adversarial domain adaptation to reduce sample bias of a high energy physics event classifier," *Machine learning: science and technology*, vol. 3, no. 1, p. 015014, 2021.

[24] V. Wu and C. Gnanasambandam, "A machine-learning approach to venture capital," *McKinsey Quarterly*, vol. 27, 2017. [Online]. Available: https://www.proquest.com/scholarly-journals/machine-learning-approach-venture-capital/docview/2371885841/se-2

[25] "QS World University Rankings 2021: Top global universities. Top Universities," 2021, accessed on November 15, 2021. [Online]. Available: https://www.topuniversities.com/university-rankings/world-university-rankings/2021

[26] P. Parasurama, "racebert–a transformer-based model for predicting race from names," *arXiv preprint arXiv:2112.03807*, 2021.

[27] F. Corea, "Ai and venture capital," in *An introduction to data*. Springer, 2019, pp. 101–110.