machine learning & knowledge extraction

Tracked for Impact Factor

Indexed in: Scopus

Article

Developing a Novel Fair-Loan Classifier through a Multi-Sensitive Debiasing Pipeline: DualFair

Arashdeep Singh, Jashandeep Singh, Ariba Khan and Amar Gupta

Special Issue

Fairness and Explanation for Trustworthy AI

Edited by

Dr. Jianlong Zhou, Prof. Dr. Andreas Holzinger and Prof. Dr. Fang Chen

MDPI

https://doi.org/10.3390/make4010011

machine learning & knowledge extraction

MDPI

Article

Developing a Novel Fair-Loan Classifier through a Multi-Sensitive Debiasing Pipeline: DualFair

Arashdeep Singh [1,*,†], Jashandeep Singh [1,†], Ariba Khan [2] and Amar Gupta [2]

[1] Floyd B. Buchanan High School, 1560 N Minnewawa Ave, Clovis, CA 93619, USA; jashan@mit.edu
[2] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar St, Cambridge, MA 02139, USA; akhan02@mit.edu (A.K.); agupta@mit.edu (A.G.)
* Correspondence: arash2348@gmail.com
† These authors contributed equally to this work.

Abstract: Machine learning (ML) models are increasingly being used for high-stake applications that can greatly impact people's lives. Sometimes, these models can be biased toward certain social groups on the basis of race, gender, or ethnicity. Many prior works have attempted to mitigate this "model discrimination" by updating the training data (pre-processing), altering the model learning process (in-processing), or manipulating the model output (post-processing). However, more work can be done in extending this situation to intersectional fairness, where we consider multiple sensitive parameters (e.g., race) and sensitive options (e.g., black or white), thus allowing for greater real-world usability. Prior work in fairness has also suffered from an accuracy–fairness trade-off that prevents both accuracy and fairness from being high. Moreover, the previous literature has not clearly presented holistic fairness metrics that work with intersectional fairness. In this paper, we address all three of these problems by (a) creating a bias mitigation technique called DualFair and (b) developing a new fairness metric (i.e., AWI, a measure of bias of an algorithm based upon inconsistent counterfactual predictions) that can handle intersectional fairness. Lastly, we test our novel mitigation method using a comprehensive U.S. mortgage lending dataset and show that our classifier, or fair loan predictor, obtains relatively high fairness and accuracy metrics.

Keywords: machine learning; algorithmic fairness; bias mitigation; mortgage lending; accuracy–fairness trade-off

Citation: Singh, A.; Singh, J.; Khan, A.; Gupta, A. Developing a Novel Fair-Loan Classifier through a Multi-Sensitive Debiasing Pipeline: DualFair. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 240–253. https://doi.org/10.3390/make4010011

Academic Editors: Jianlong Zhou, Andreas Holzinger and Fang Chen

Received: 23 November 2021
Accepted: 24 February 2022
Published: 12 March 2022

1. Introduction

Machine learning (ML) models have enabled automated decision making in a variety of fields, ranging from lending to hiring to criminal justice. However, the data often used to train these ML models contain many societal biases. These biased models have the potential to perpetuate stereotypes and promote discriminatory practices, therefore giving privileged groups undue advantages. As a result, it has become increasingly important for ML researchers and engineers to work together in eliminating this algorithmic unfairness.

Despite an awareness of the need for these fair models, there still exist many examples of models exhibiting prevalent biases, such as the following:

- In 2016, ProPublica reported that the ML models used by judges to decide whether to keep criminals in jail were discriminatory against African-American males, labeling them with relatively high recidivism (tendency to re-offend) scores [1].
- Amazon discovered that its automated recruiting system was biased against female job applicants, rendering them far less successful in the application process [2].
- A healthcare algorithm evaluated on 200 million individuals to predict whether patients needed extra medical care was highly discriminatory against African Americans while prioritizing white individuals [3].

One particular domain where bias mitigation has become especially crucial is mortgage lending. It has been reported that over 3.2 million mortgage loan applications

and 10 million refinance applications exhibited bias against Latin and African American lenders [4]. In another study, it was shown that minority groups were charged significantly higher interest rates (by 7.9 basis points) and were rejected 13% more often than their privileged counterparts [5]. These biases, when trained upon, lead to discriminatory loan classifiers that can cause greater gaps in suitable housing, wealth, and property between unprivileged (African Americans) and privileged (White and Asian American) groups. This problem becomes exacerbated as companies rely on these biased models in making real-life decisions. Additionally, these biased models are unlawful under the Equal Credit Opportunity Act (ECOA) of the United States, which forbids the discrimination of individuals based upon sensitive attributes (e.g., race, gender, national origin, and ethnicity) by any private or public institution. It has thus become the moral and legal duty for researchers and software developers to find a solution to this problem of "algorithmic unfairness".

Fortunately, work has been done in approaching this bias. There are currently three ways in which bias mitigation has been approached, correlating with before, in, and after the data usage pipeline:

- *Pre-processing*—the transformation of data (e.g., the alteration of balancing distributions) to "repair" inherent biases [6–9].
- *lIn-processing*—the use of classifier optimization or fairness regularization terms to affect model learning and maximize a model's fairness [10–12].
- *lPost-processing*—the manipulation of model output to improve performance and fairness metrics [13].

While prior works gained relatively high fairness and performance metrics using one or more of these mitigation techniques, the current literature contains three main problems that hamper their adaptability and deployment: (1) more work can be done on extending mitigation techniques to situations with intersectional fairness, (2) an accuracy and fairness trade-off still exists, and (3) an absence of accepted fairness metrics for data with multiple sensitive parameters and sensitive options. For point 1, it may be important to clarify that in our study, intersectional fairness extends from its current definition by acknowledging the existence of "joint" as a valid sensitive option under various sensitive parameters since the category is commonplace in the realm of financial documents (e.g., mortgage loan applications and tax returns).

Our Contributions: In this paper, we target all three of the previously stated problems to develop a novel and real-world applicable fair ML classifier in the mortgage-lending domain that obtains relatively high fairness and accuracy metrics. Through this process, we coin a bias mitigation pipeline called DualFair (a pre-processing strategy), which approaches intersectional fairness through data sampling techniques, and solves problems hindering the growth of the "Fairness, Accountability, and Transparency", or FAT, field.

More concretely, the main insights we provide within this paper include the following:

- Creating a bias mitigation strategy coined DualFair targeted toward the mortgage domain, which debiases data through oversampling and undersampling techniques that target the root causes of bias.
- Extending our mitigation approach to intersectional fairness by subdividing datasets and then balancing by class and labels.
- Developing a novel fairness metric called AWI (alternate world index) that is generalizable to intersectional fairness and an accurate representation of model fairness.
- Eliminating the *accuracy–fairness trade-off* by debiasing our mortgage lending data using DualFair.
- Creating a fair loan classifier that achieves relatively high fairness and accuracy metrics and can potentially be used by practitioners.

The rest of this paper is structured as follows: Section 2 provides an overview of the prior work and contributions in the FAT field, particularly in relation to our own. Section 3 explains fairness terminology and metrics used throughout our paper. Section 4 gives a detailed outline of our data, debiasing strategy, novel fairness metric AWI, and

experimental design. Section 5 summarizes the results of our bias mitigation pipeline and documents the success of our approach, compared to the previous state-of-the-art based upon *accuracy, precision, false alarm rate, recall, F1 score and AWI*. In Section 6, we give a brief overview on the potential directions for future work. Finally, Section 7 concludes the paper.

2. Related Work

Fairness in ML models is a largely explored topic within the AI community. Recently, major industries have begun to put a greater priority on AI fairness. IBM developed AI Fairness 360, which is a fairness toolkit that contains commonly used fairness metrics and debiasing techniques to aid researchers working in ML fairness [14]. Microsoft has established FATE [15], a research group dedicated to fairness, accountability, transparency, and ethics in AI. Google [16], Microsoft [17], IEEE [18], and The European Union [19] each respectively published on ethical principles in AI, which are general guidelines on what the companies define as "responsible AI". Facebook created bias detection tools for its own internal AI systems [20]. The research community has started to take an interest in fair AI as well. ASE 2019 and ICSE 2018 hosted workshops on software fairness [21]. Mehrabi et al. studied various notions of fairness and fundamental causes of bias [22]. ACM established the FAccT '21 as a conference to spearhead work on fairness and transparency of ML algorithms [23].

Thus far, achieving algorithmic fairness has been addressed through pre-processing, in-processing, and post-processing approaches. Prior work has proposed a variety of bias-mitigating methods. *Optimized pre-processing* [9] is a pre-processing method that seeks to achieve group fairness by altering labels and features, using probabilistic transformations. Zhang et al. presented the in-processing approach, *adversarial debiasing* [24], which increases accuracy and strips away an adversary's ability to make decisions based upon protected attributes using GANs. *Reject option classification* [25] is a post-processing strategy that translates favorable outcomes from the privileged group to the unprivileged group and unfavorable outcomes from the unprivileged group to the privileged group based upon a certain level of confidence and uncertainty. Chakraborty et al. proposed *Fair-SMOTE* [26], a pre-processing and in-processing approach, which balances class and label distributions and performs *situation testing* (i.e., testing individual fairness through alternate "worlds").

Our experience has shown, however, that these approaches lack mainly in their ability to extend to intersectional fairness. Chakraborty et al. noted that the consideration of intersectional fairness would divide data into unmanageable small regions [27]. Salerio et al. attempted to approach intersectional fairness by creating AEQUITIS, a fairness toolkit, that uses parameter-specific fairness metrics to systematically view bias within one sensitive parameter at a time [28]. Gill et al., Chakraborty et al., and Kusner et al. approached fairness in their own separate domains by using a singular sensitive parameter, designating one privileged group and one unprivileged group as a way to compare mitigation results [6]. Ghosh et al. introduced a new intersectional fairness metric coined the worst-case disparity framework that finds the largest difference in fairness between two subgroups and then minimizes this difference by utilizing existing single-value fairness metrics [29]. We refrain from using this approach in our work because it relies on previous single-value metrics that require the designation of unprivileged and privileged groups. Through our experience with DualFair, we argue that it is possible to debias data with multiple sensitive parameters and sensitive options, given a proper pipeline, approach, and data. This allows for deployability and scalability within real-world systems. We also show that given this type of data, one could devise a fairness metric (AWI), which considers bias from all parameters and options cohesively.

The mortgage domain has seen its own work in the realm of AI fairness as well. Fuster et al. and Bartlet et al. showed disparity in over 92% of loans, spanning origination, interest rate charges, and preapprovals across the United States on the basis of sex, race, and ethnicity [5,30]. Gill et al. built upon these conclusions and proposed a state-of-the-art machine learning workflow that can mitigate discriminatory risk in singular sensitive parameter and sensitive option mortgage data while maintaining interpretability [31]. This

framework was used to create a fair loan classifier. Lee et al. also presented a theoretical discussion of mortgage domain fairness through relational trade-off optimization [32]. That is, the paper discussed a method to achieve a balance between accuracy and fairness within the *accuracy–fairness trade-off* on mortgage lending data rather than maximizing both.

Our work builds on the foundation created by these previous works in mortgage lending and bias mitigation in AI systems at large. It is important to note that the literature varies in two main aspects, however: (1) finding bias and (2) mitigating bias.

While most prior work has centralized on finding bias, our study seeks to mitigate bias through creating a debiasing pipeline and training a fair loan classifier.

3. Fairness Terminology

In this section, we outline the fairness terminology that will be used within this work. An *unprivileged group* is one that is discriminated against by a ML model. *Privileged groups* are favored by a ML model due to some sensitive parameter. These groups usually receive the *favorable label* (i.e., the label wanted), which, for our purposes, is a mortgage loan application being accepted. A *sensitive parameter*, also known as a *protected attribute*, is a feature that distinguishes a population of people into two groups, an unprivileged group and a privileged group. This parameter was historically discriminated against (e.g., race and sex). *Sensitive options* are sub-groups, or options, within sensitive parameters (e.g., for race: White, Black, or Asian). The distribution of all sensitive parameters and sensitive options (e.g., White males, Black males, White females, Black females) is referred to as a *class distribution*. The distribution of favorable outcomes and unfavorable outcomes for a particular group as represented by the ground truth is its *label balance*. *Label bias* is a type of societal bias which can shift the label balance (e.g, a mortgage underwriter subconsciously denying reliable African-American lenders for loans). *Selection bias* is another type of bias that is created when selecting a sample. For example, suppose that researchers are collecting car insurance data for a particular location. However, the particular location they are collecting data from has historical discrimination that causes it to have a low annual income per person. The insurance data collected in this location, therefore, would contain implicit economic biases. *Fairness metrics* are a quantitative measure of bias within a specific dataset.

Finally, there are two main types of fairness denoted within the literature: *individual fairness* and *group fairness* [22].

- Individual fairness is when similar individuals obtain similar results.
- Group fairness is when the unprivileged and privileged groups, based upon a particular sensitive parameter, are treated similarly.

Before beginning our discussion on DualFair, we would like to note that in this work, we use a binary classification model for all of our inferences and methods of achieving these notions of fairness. Future work could make an effort of looking into algorithmic fairness with regression models instead.

4. Materials and Methods

4.1. Mortgage Data

Previous domain-specific ML studies have faced many challenges in acquiring large-scale datasets for comprehensive work. The mortgage domain conveniently offers a solution to this problem. For our study, we use the HMDA dataset, which was publicly made available by the Home Mortgage Disclosure Act (HMDA) of 1975 [33]. HMDA data have been used in various studies to outline and understand recent mortgage trends. They span 90% of all loan origination and rejections in the United States and contain over 21 distinct features (e.g., race, sex, and ethnicity).

It is important to note that HMDA has prevalent racial and gender prejudices within the data. Table 1 shows a quantitative distribution of these biases through particular sensitive, or biased, features. Our evaluation shows that features, such as loan amount, income, and interest rate, already give certain groups an undue advantage. For instance,

females are given higher mean and median interest rates, lower loan amounts, and lower property values compared to those of males. This is a primary indication of bias against females applicants when compared to male applicants. The following issue will be dealt downstream during DualFair's bias mitigation pipeline.

Table 1. Population statistics for HMDA (note: all values are expressed in USD 1000; the interest rate and LTV, the loan-to-value ratio, between the loan amount and property value are expressed as a percentage; lastly, groups are split via sensitive parameters of sex, race, and ethnicity, respectively).

Group		Income	Loan Amt.	Interest Rate	LTV Ratio	Property Value
Male	Mean	124	267	3.36%	76.4%	405
(n = 1,224,719)	Median	84	225	3.19%	79.8%	315
Female	Mean	89	225	3.68%	74.4%	346
(n = 805,583)	Median	69	195	3.25%	78.6%	275
White	Mean	131	270	3.41%	73.6%	428
(n = 3,474,562)	Median	96	235	3.13%	76.1%	335
Black	Mean	94	240	3.47%	83.5%	319
(n = 220,517)	Median	76	215	3.25%	90.0%	265
Non. Hisp.	Mean	133	271	3.42%	73.6%	430
(n = 3,360,377)	Median	97	235	3.13%	76.2%	335
Hispanic	Mean	96	256	3.36%	79.3%	362
(n = 334,184)	Median	76	235	3.25%	80.0%	305

In this study, we use HMDA national data from 2018 to 2020 and 2020 data from two small states (<150,000 rows of data), two medium states (>150,000 to <250,000 rows of data), and two large states (>250,000 rows of data). These data are unique compared to past years', as, in 2018, the Dodd–Frank Wall Street Reform and Consumer Protection Act (Dodd–Frank) mandated more expansive updates to mortgage loan data from all applicable institutions. Dodd–Frank led to the addition of features such as credit score, debt-to-income ratio, interest rate, and loan term, providing a more comprehensive review of loan applicants and studies for algorithmic fairness. In our work, we build a fair ML classifier on the HMDA data using DualFair, where our classifier predicts whether an individual originated (i.e., $y = 1$) or was denied (i.e., $y = 0$) a loan. The following steps are taken to facilitate the creation of this classifier and analysis of HMDA data:

- 755,000 loan records were randomly sampled without replacement from each year 2018–2020 to form a combined HMDA dataset, spanning over 2.2 million loan applications, for analysis.
- Features with more than 25% data missing, exempt, or not available were removed during data pre-processing. When deciding upon this value, we tested the threshold values 20%, 25%, 30%, 35%, and 40%. We found that setting the threshold value to 25% produced the best accuracy. It is essential to note that this value was *most* optimal for our data and prediction model. However, it is important to tune this value when using alternative methods or materials.
- Only White, Black, and joint labels from the race category, male, female, and joint labels from the sex category, and Non-Hispanic or Latino, Hispanic or Latino, and joint labels from the ethnicity category were used in the study.

Note: Joint is defined as a co-applicant sharing a different feature option than the main applicant. For instance, a White male applicant and a Black co-applicant would be joint for race. Future research is highly encouraged in implementing all sensitive parameters and sensitive options.

4.2. *Debiasing*

It has been shown that data bias mainly derives from two factors: *label bias* and *selection bias* [34,35]. In this section, we eliminate both of these biases from our dataset in the debiasing process through a novel pre-processing approach that we coin DualFair. More concretely, DualFair removes the *accuracy–fairness trade-off* and increases the algorithmic fairness metrics by approaching the task of selection bias and label bias through balancing at the class and label levels.

At first, DualFair splits the central dataset into sub-datasets based upon all combinations of sensitive parameters and sensitive options. The following methodology is used to designate the sub-datasets: suppose a dataset contains two sensitive parameters, sex and race. Additionally, suppose that there are only two sensitive options for each sensitive parameter, *male (M)* or *female (F)* and *Black (B)* or *White (W)*. After being split, the dataset is broken into four distinct sub-datasets *WM (White males)*, *BM (Black males)*, *WF (White females)*, and *BF (Black females)*.

In the case of HMDA, DualFair results in 27 sub-datasets. We start with three sensitive parameters: race, sex, and ethnicity. Then upon each parameter, there is a division by sensitive options, where each sensitive parameter has three different options. For race, an individual could be White, Black, or joint; for sex, individuals could be male, female, or joint; and, lastly, for ethnicity, individuals could be Non-Hispanic or Latino, Hispanic or Latino, or joint. Through the combination of these sensitive parameters and options, a total of 27 unique datasets (i.e., groups) are formed and utilized extensively throughout the pipeline.

We can generate an equation to represent the count of sub-datasets given by multiplying each sensitive option count, o_i, of a sensitive parameter:

$$\prod_{i=1}^{n} o_i \tag{1}$$

$$n = number\ of\ sensitive\ parameters$$

$$o_i = number\ of\ sensitive\ options\ of\ the\ ith\ sensitive\ parameter$$

In our case, $n = 3$ and each o_i is 3; therefore, multiplication (i.e., $o_1 \cdot o_2 \cdot o_3$) gives us 27 sub-datasets.

After dividing into sub-datasets, in each subset of data, we obtain the number of accepted ($y = 1$) and rejected ($y = 0$) labels. We take the median number of accepted and rejected labels in all subsets of data and synthetically oversample or undersample each class in the sub-datasets to that value, using SMOTE (synthetic minority oversampling technique) or RUS (random undersampling). The result of this process is a class-balanced HMDA dataset that has no selection bias.

Figure 1 captures the class and label distributions before and after selection bias are removed. Observing the figure, it can be seen that HMDA is unbalanced within the class distributions (i.e., between different subsets of data) and label distribution (i.e., balance of accepted and rejected labels within a class). Some sub-datasets are far more common in the data, while others are less represented. Thus, there is a huge imbalance that propagate root biases when trained on a model, which need to be taken into account. In the after class and label balance figure (right graph), it is shown that all sub-datasets have become balanced at both the label and class distribution level. This strips away selection bias by regulating the way that the model perceives the data. That is, the data are repaired so that no class is overly trained upon and each group has an equivalent amount of favorable labels.

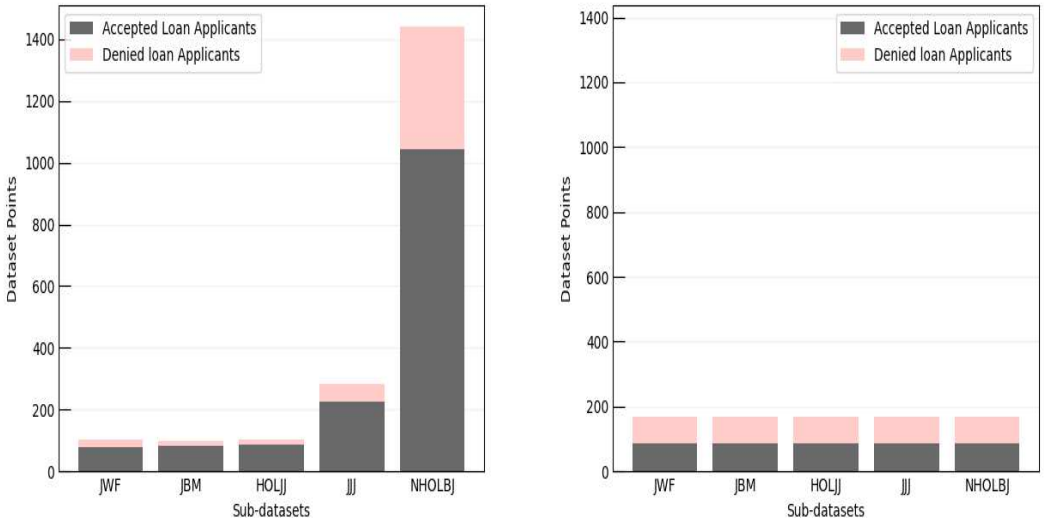

Figure 1. Class and label balancing before (**left** graph) and after (**right** graph) using DualFair. (Note: only 5 sub-datasets from the balancing procedure of Connecticut, a smaller sample of HMDA, are shown by the normal 27; note that J denotes joint, B denotes Black, W denotes White, F denotes female, M denotes male, HOL denotes Hispanic or Latino, and NHOL denotes Non-Hispanic or Latino. These denotations form sub-groups shown above. For instance, JJJ is a joint, joint, joint individual and HOLJJ is a Hispanic or Latino, joint, joint individual).

In using oversampling techniques to achieve this equal distribution, we follow the general guidelines taken from Chakraborty et al. Thus, we preserve valuable associations between values when oversampling. When creating data, we make sure they are close to the neighboring examples. This allows for the average case association that may be between two particular variables. We use two hyperparameters called "mutation amount" (f) and "crossover frequency" (cr) to carefully use SMOTE. These parameters lie in between [0, 1]. "Mutation amount" controls the probability the new data point is different from the parent point, while "crossover frequency" denotes the probability of how different the new data point is to its paternal point. When tuning, we determine that from the options of 0.2, 0.5, and 0.8, using 0.8 (80%) as the value for both of these parameters provides the best results for data generation in regards to preserving vital associations.

After balancing, our debiasing process uses a method known as situation testing, coined by Chakrabokty et al., to reduce label bias [27]. Situation testing finds biased points within the training data by probing all combinations of sensitive options. More clearly, situation testing will test all combinations of sensitive options on an ML model trained on the balanced dataset. If all combinations of sensitive option do not result in the same prediction, then that data point is removed. For instance, given that sex is a sensitive parameter, if changing a mortgage loan applicant's sex from male to female alters the mortgage loan approval prediction, then that data point is considered biased. This process removes biased data points from the dataset and decreases label bias.

The use of balancing and situation testing is illustrated by the DualFair process outlined in Figure 2. After removing label and selection bias from our dataset, we have a debiased dataset that can be used in our testing framework for yielding metrics. DualFair removes the accuracy–fairness trade-off within this debiasing process, as it creates fair data to evaluate and train from. Fairness metrics are shown before and after DualFair in Section 5.1.

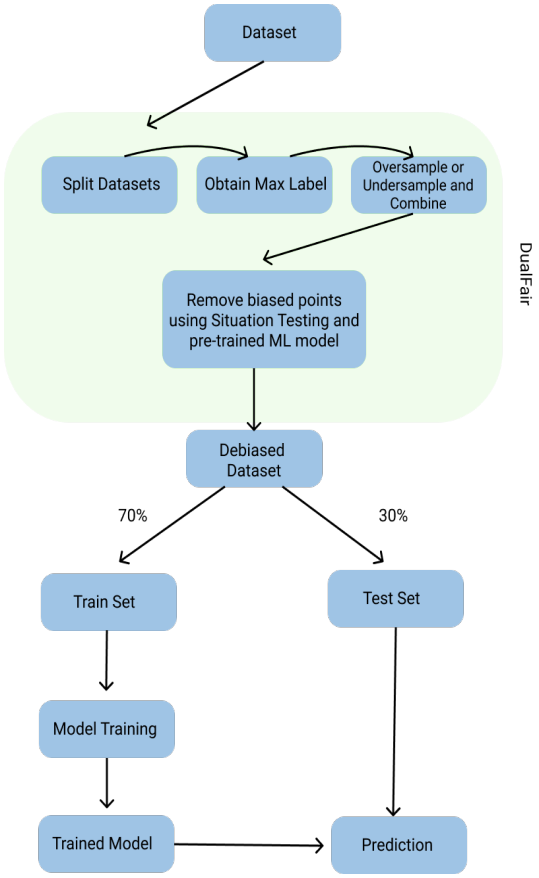

Figure 2. DualFair bias mitigation pipeline.

4.3. Novel Fairness Metrics

Although there is a significant amount of fairness metrics proposed by the literature, all of the fairness metrics lack in one general area: non-applicability to data with multiple sensitive parameters and sensitive options. We remedy this issue by creating a new fairness metric, the alternate world index (AWI), based upon computational truths and previous literature.

Let us begin by defining fairness. Because of varying definitions of fairness in prior work, for our purposes, we will define fairness as having different sensitive groups (e.g., male group and female group) being treated equivalently and similar individuals (i.e., possess similar statistics) being treated equivalently. For this definition of fairness to occur, both *group fairness* and *individual fairness* must be met.

Mathematically, this fairness is defined as $U_g|y = 1$ being similar to $P_g|y = 1$, where U_g is the underprivileged groups, P_g is the privileged groups, and $y = 1$ represents the desired outcome.

Using our previous definition of fairness, we can generate a metric that satisfies the interpretation of the previous requirements for intersectional fairness work; we coin this metric AWI (alternate world index).

$$\frac{\sum_{w=1}^{n}[\frac{\sum_{i=1}^{k} p_i}{k} \neq \{0,1\}]_w}{n} \tag{2}$$

$$n = number\ of\ data\ points$$

$$k = subsets\ of\ data$$

$$p_i = prediction\ in\ one\ world$$

Specifically, AWI is a count of the number of biased points within a dataset normalized by the dataset size. A biased point is identified by iterating a point through all respective counterfactual worlds (i.e., similar individuals with different sensitive parameters and options) and evaluating prediction constancy with the model in use. If, under all counterfactuals, the model prediction is constant, we call this point devoid of bias; however, if one situation yields a unique result, the point is marked as ambiguous (biased). Figure 3 illustrates this procedure. After repeating this process for all points in the test dataset, we use the count of total bias points normalized by the total dataset points to yield AWI. AWI captures model bias by inferring all biased points as synonymous to biased predictions the model made. In our study, we report AWI to be 10 times larger than its value to more accurately represent its difference, whether beneficial or harmful.

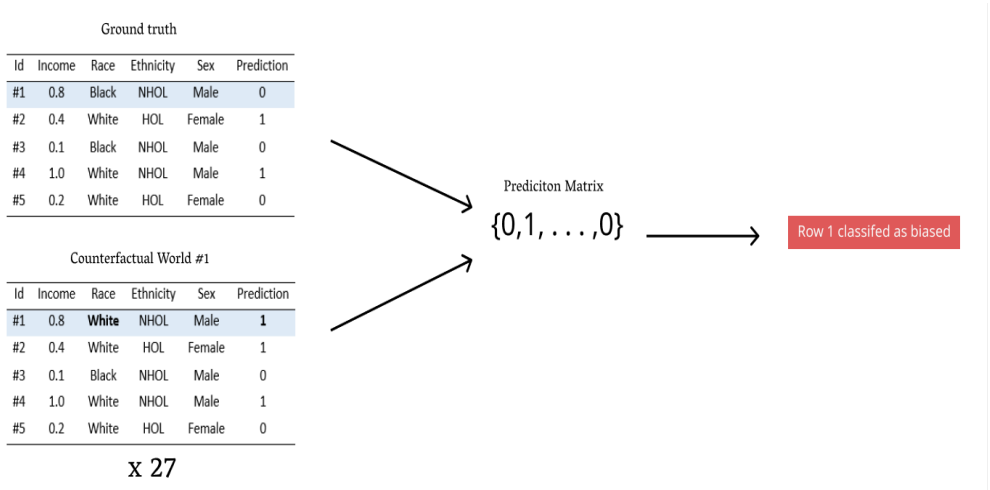

Figure 3. AWI bias classification procedure. (Note that only 2 counterfactual worlds are shown; however, a select data point will, for our use case, iterate over 27 counterfactual worlds. Bold indicates a change in parameters and results with respect to the ground truth.)

AWI extends the fairness metrics to the realm of intersectional fairness by quantifying the biased predictions within a dataset using counterfactuals. By doing this, we solve two major problems when applying fairness metrics to data with multiple sensitive parameters and sensitive options: (1) an unequal amount of privileged to unprivileged group and (2) the lack of one holistic fairness value for an entire dataset.

AWI takes a pragmatic approach to fairness evaluation, specifically targeting similar individuals and their different realities. It is a versatile metric for any work looking to achieve individual or group fairness with or without multiple sensitive parameters and sensitive options. One pitfall of AWI is its computational expense, especially with large volumes of data, as all class distributions must be predicted upon. Future research could look for directions in optimization.

Since AWI is a metric of our own creation, we utilize the average odds difference (*AOD*), a method of equalized odds fairness from Chakraborty et al., as a method of comparison in testing [26]. *AOD* measures the average of difference in false positive rates (*FPR*) and true positive rates (*TPR*) for unprivileged and privileged groups [27]. Mathematically, *AOD* represents the following:

$$TPR = \frac{TP}{P} = \frac{TP}{(TP + FN)} \tag{3}$$

$$FPR = \frac{FP}{N} = \frac{FP}{(FP + TN)} \tag{4}$$

$$AOD = [(FPR_u - FPR_p) + (TPR_u - TPR_p)] * 0.5 \tag{5}$$

Two major flaws with *AOD* for our work is that (1) an unprivileged and privileged group must be designated and (2) *AOD* can only account bias for one sensitive parameter at a time (i.e., cannot collectively handle data with multiple sensitive parameters and sensitive options). Point 1 serves an issue when using multiple option datasets (i.e., datasets including multiple options, such as White, Black, and joint for a parameter such as race) because only two options can be designated as privileged and unprivileged. To circumvent this issue, we refrain from including joint in all of our *AOD* testing calculations. Point 2 raises the issue that all biases within a dataset are not measured at once. However, for our use case, we attempt to deal with this issue by measuring three *AOD* metrics simultaneously, AOD_{sex}, AOD_{race}, and $AOD_{ethnicity}$, for all runs.

4.4. Experimental Design

Here, we describe the process we used to prepare our data for our experiments. Our study used the 2020 HMDA data for the training and testing set and logistic regression (LSR) as the classification model. We decided to use logistic regression with default settings as our classification model because it is a simple model that does not require a large amount of high-dimensional data, which is not a common occurrence in the fairness domain due to the process of debiasing datasets. Furthermore, due to the model's simplicity and interoperability, previous literature has frequently used the model for classification problems and, in most cases, concluded it to be the best performing classifier for their task. For each experiment, the dataset is split using 5-fold cross validation (train—70%, test—30%). This step is repeated 10 times with a random seed and then the median is reported. The feature columns that have at most 25% of the values as missing or not applicable are kept, but any rows containing said values that are missing or not applicable are removed. Additionally, non-numerical features are converted to numerical (e.g., female—0, male—1, and joint—2) values. It is important to note that any data points that do not contain White, Black or joint as the race are removed. Finally, all feature values are normalized between 0 and 1.

Now, we describe how we obtained the results for each of our experiments. In DualFair, the training and testing data are both repaired during the bias mitigation pipeline. We evaluated AWI, AOD_{sex}, AOD_{race}, and $AOD_{ethnicity}$ fairness metrics before and after the bias mitigation process for comparison. To do this, we first trained a classification model on the training data (i.e., either before or after DualFair) and then measured its fairness and accuracy on testing data. Our accuracy was measured in terms of *recall*, *specificity*, *accuracy*, and *F1 score*. Fairness was measured in *AWI*, AOD_{sex}, AOD_{race}, and $AOD_{ethnicity}$. *Recall*, *specificity*, *accuracy*, and *F1 score* are better at larger values (i.e., closer to 1), while *AWI*, AOD_{sex}, AOD_{race}, and $AOD_{ethnicity}$ are better at smaller values (i.e., closer to 0).

In this work, we perform experiments using DualFair on 2020 state-level data from two small states (<150,000 rows of data), two medium states (>150,000 to <250,000 rows of data), and two large states (>250,000 rows of data). We also perform an experiment using DualFair on 2018–2020 nationwide HMDA (2,265,000 rows of data). For this experiment, we randomly sample 755,000 rows from each year and then apply DualFair. For selecting our two small, medium, and large states, we group all states according to category and randomly sample two states from each group.

5. Results

We structure our results around three central research questions (RQ).

5.1. RQ1: How Well Does DualFair Create an Intersectional Fair Loan Classifier?

RQ1 explores the performance of our pipeline in debiasing mortgage data and creating a fair loan classifier. It is reasonable to believe that a fair loan classifier should perform two things proficiently: prediction and fairness. Accordingly, we test DualFair for both performance metrics (e.g., accuracy, recall, precision, and F1 score) and fairness metrics

(e.g., AWI, AOD_{sex}, AOD_{race}, and $AOD_{ethnicity}$). A summary of all the metrics we use can be found in Table 2.

Table 2. Fairness and performance metrics. (Note: AOD represents AOD_{sex}, AOD_{race}, and $AOD_{ethnicity}$; FPR denotes false positive rate and TPR denotes true positive rate. The subscript U denotes underprivileged group and P denotes privileged group.)

Metric	Equation	Ideal Value
Accuracy	$\frac{(TP+TN)}{(TP+FP+TN+FN)}$	1
Precision	$TP/(TP+FP)$	1
Specificity	$TN/N = TN/(TN+FP)$	1
Recall	$TP/P = TP/(TP+FN)$	1
F1 Score	$\frac{2*(Precision*Recall)}{(Precision+Recall)}$	1
Alternate World Index (AWI)	$\frac{\sum_{w=1}^{n}[\frac{\sum_{i=1}^{k} p_i}{k} \neq \{0,1\}]_w}{n}$	0
AOD	$[(FPR_u - FPR_p) + (TPR_u - TPR_p)] * 0.5$	0

In Table 3, we give the performance and fairness before and after the DualFair pipeline. We run seven different trials from a range of small, medium, and large states as well as nationwide. Columns 1, 2, 3, 4, 5, 6, and 7 summarize the results of DualFair on the trials. Measured in terms of AWI, DualFair is successful in increasing fairness for 5 of 7 states varying in data size. Measured in terms of AOD_{sex}, AOD_{race}, and $AOD_{ethnicity}$, DualFair is successful in increasing fairness for 7 of 7 states. A state is considered successful if a majority of the AOD metrics show improvement. In terms of transitions (i.e., changes between before and after the pipeline), AOD_{sex}, AOD_{race}, and $AOD_{ethnicity}$ increase fairness in 18 of 21 transitions. It is important to note that AWI, AOD_{sex}, AOD_{race}, and $AOD_{ethnicity}$ are all multiplied by 10 to more accurately represent their differences before and after the bias mitigation pipeline. In addition to fairness, DualFair benefits precision and specificity on all occasions while damaging recall and F1 score only slightly.

Table 3. Results before and after DualFair. AWI, AOD_{sex}, AOD_{race}, and $AOD_{ethnicity}$ the lower the better. Accuracy, Precision, Recall, Specificity, and F1 Score the higher the better. All values are rounded. AWI, AOD_{sex}, AOD_{race}, and $AOD_{ethnicity}$ are multiplied by 10 to accurately show the change.

# of Rows	Nationwide HMDA 2,265,000		CA (Large) 2,000,000		TX (Large) 1,964,077		IL (Medium) 864,270		WA (Medium) 823,323		NV (Small) 316,969		CT (Small) 231,251	
	Before	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**
AWI (−)	0.17	0.04	0.15	0.13	0.13	0.02	0.06	0.11	0.18	0.26	0.13	0.09	0.24	0.05
AOD Sex (−)	0.08	0.01	0.03	0.01	0.07	0.04	0.11	0.01	0.07	0.05	0.10	0.03	0.12	0.02
AOD Race (−)	0.17	0.00	0.09	0.02	0.07	0.06	0.10	0.06	0.05	0.03	0.08	0.04	0.12	0.06
AOD Ethnicity (−)	0.07	0.01	0.12	0.05	0.06	0.01	0.03	0.03	0.01	0.04	0.03	0.07	0.04	0.01
Accuracy (+)	0.98	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.97	0.97	0.98	0.98	0.98
Precision (+)	0.99	0.99	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.98	1.00	0.99	1.00
Recall (+)	0.98	0.95	0.97	0.94	0.98	0.97	0.99	0.97	0.98	0.95	0.98	0.97	0.98	0.96
Specificity (+)	0.97	0.99	0.96	1.00	0.98	1.00	0.96	1.00	0.95	1.00	0.95	1.00	0.96	1.00
F1 Score (+)	0.99	0.97	0.99	0.97	0.99	0.98	0.99	0.98	0.99	0.97	0.99	0.98	0.99	0.98

We hypothesize that medium states, columns 4 and 5, falter in their ability to improve fairness, AWI, as their median oversampling and undersampling value tends to be small compared to counts of individuals in certain classes. For robustness, taking an average rather than a median for the oversampling and undersampling value may provide improved results in the future.

Thus the answer to **RQ1** is "DualFair establishes an intersectional fair loan classifier that achieves both high levels of accurate prediction and fairness". That is, DualFair can

mitigate bias within data with multiple sensitive parameters and sensitive options while maintaining high levels of accuracy. This is one of the biggest achievements of our work and is pivotal toward real-world applicability.

5.2. RQ2: Does DualFair Eliminate the Accuracy–Fairness Trade-off?

RQ2 seeks to consider if DualFair effectively removes the accuracy–fairness trade-off. That is, we hope to explore if our pipeline resulted in consistent accuracy while increasing fairness simultaneously.

In all our testing, including our seven rigorous trials in Table 3, we found that our accuracy remained consistent prior to debiasing. It also remained stagnate after debiasing, albeit showing an increase in fairness metrics. We hypothesize that the reason for this derives from our pipeline which "repairs" both training and testing data. We believe the repair process ensures that a fair model is evaluated upon fair data. This leads the model performance to remain equivalent both prior to and after DualFair.

Hence, the answer to **RQ2** is "Yes, DualFair simultaneously removes the accuracy–fairness trade-off while achieving individual fairness".

5.3. RQ3: Is DualFair Capable of Capturing All Sensitive Parameters and Sensitive Options in HMDA Data?

To answer **RQ3**, we analyzed DualFair's time complexity with increasing multiple sensitive parameters and sensitive options for HMDA data. The literature has suggested that all sensitive parameters and sensitive options may not be both computationally and logically feasible due to data division into very small, unmanageable regions.

Our analysis of DualFair's time complexity tells us that it is possible for DualFair to scale to all sensitive parameters and sensitive options. We determined that although DualFair's computational expense compounds at a rate of $O(n^2)$, our mitigation strategy is still adequate at scaling with the demand of more sensitive parameters or options for HMDA data, as it only contains about six sensitive parameters with an average of three sensitive options, each.

One limitation of this conclusion is that adequate data must be provided. That is, at least two individuals from all sub-datasets of the original data must be present. In addition, each sub-dataset must contain at least a rejected individual ($y = 0$) and an accepted individual ($y = 1$) to generate oversamples, using SMOTE. Thus, there arises a problem with the data fragmentation into small sub-datasets when large data with multiple sensitive parameters and sensitive options may not be readily available.

Overall, DualFair has the capacity to scale to all of HMDA, given its limited quantity of sensitive parameters and options and large dataset size that can provide adequate data.

6. Future Works

The following are future directions and limitations for researchers to consider:

- Direction 1: Use of DualFair within other AI-related domains (e.g., healthcare, job applications, and car insurance).
- Direction 2: Performing a widespread analysis of bias mitigating methods (including DualFair) with AWI as a fairness metric for comparison.
- Direction 3: Applying DualFair or another bias-mitigating method to ML regression models in various domains.
- Limitation 1: Instability of DualFair on low amounts of data with multiple sensitive parameters and sensitive options.
- Limitation 2: DualFair, although it removes the accuracy–fairness trade-off, needs large quantities of data to achieve high-performance metrics.

We provide our GitHub repository below to aid practitioners and researchers in enhancing the domain of AI fairness and/or adopting DualFair for their own purposes. The repository contains the source code for DualFair and AWI as well as their application

on HMDA data (https://github.com/ariba-k/fair-loan-predictor accessed on 6 February 2022).

7. Conclusions

This paper tested the DualFair process, which removes label bias and selection bias within the working data (*pre-processing*). As shown above, we showed that DualFair can be applied to the HMDA dataset to create a high-performing fair ML classifier in the mortgage-lending domain. Unlike other ML fairness pipelines, DualFair is capable of such results, even if the data contain non-binary *sensitive parameters* and *sensitive options*, such as in the case of HMDA. We showed that DualFair is a comprehensive bias mitigation tool targeted specifically for the mortgage domain. In summary, we achieved the following in this work:

1. Created a novel bias mitigating method called DualFair for data with with multiple sensitive parameters and sensitive options.
2. Developed a new fairness metric (AWI) that can be applied to data with multiple sensitive parameters and sensitive options.
3. Established a fair machine learning classifier in the mortgage-lending domain.

We aim for our work to cause the creation of fair ML models in other domains of work and solve the dilemma of linking research with real-world deployment. We hope our work is adopted by mortgage-lending organizations wanting to implement a state-of-the-art non-discriminatory model for loan prediction and comply with the ECOA.

Author Contributions: Conceptualization, A.S., J.S., A.K. and A.G.; methodology, A.S. and J.S.; software, A.S., J.S. and A.K.; validation, A.S., J.S. and A.K.; formal analysis, A.S., J.S. and A.K.; investigation, A.S., J.S. and A.K.; resources, A.S., J.S. and A.K.; data curation, A.S., J.S. and A.K.; writing—original draft preparation, A.S., J.S. and A.K.; writing—review and editing, A.S., J.S., A.K. and A.G.; visualization, A.S., J.S. and A.K.; supervision, A.G.; project administration, A.G. All authors have read and agreed to the published version of the manuscript.

Funding: Not applicable.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that were used within this work can be found at [33]. These data are publicly available, as made possible by the Home Mortgage and Disclosure Act (HMDA). For our experiments, we used 2020 HMDA data when working with state versions of HMDA while employing 2018–2020 HMDA data when using its nationwide level.

Acknowledgments: We would like to acknowledge Ye Zhu for providing integral advise on this work and helping in the ideation process. We would also like to thank the anonymous reviewers for their invaluable feedback, which were critical to improving the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. Machine Bias—ProPublica. 2016. Available online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed on 6 February 2022).
2. Dastin, J. Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women. 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (accessed on 6 February 2022).
3. Ledford, H. Millions of Black People Affected by Racial Bias in Health-Care Algorithms. *Nature* **2019**, *574*, 608–609. [CrossRef] [PubMed]
4. Weber, M.; Yurochkin, M.; Sherif, B.; Vanio, M. Black Loans Matter: Fighting Bias for AI Fairness in Lending. Available online: https://mitibmwatsonailab.mit.edu/research/blog/black-loans-matter-fighting-bias-for-ai-fairness-in-lending/ (accessed on accessed on 6 February 2022).
5. Bartlett, R.; Morse, A.; Stanton, R.; Wallace, N. Consumer-Lending Discrimination in the FinTech Era. *J. Financ. Econ.* **2022**, *143*, 30–56. [CrossRef]
6. Kusner, M.J.; Loftus, J.R.; Russell, C.; Silva, R. Counterfactual Fairness. *arXiv* **2018**, arXiv:1703.06856.

7. Chiappa, S.; Gillam, T.P.S. Path-Specific Counterfactual Fairness. *arXiv* **2018**, arXiv:1802.08139.
8. Brunet, M.E.; Alkalay-Houlihan, C.; Anderson, A.; Zemel, R. Understanding the Origins of Bias in Word Embeddings. In Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 803–811.
9. Calmon, F.; Wei, D.; Vinzamuri, B.; Natesan Ramamurthy, K.; Varshney, K.R. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Nice, France, 2017; Volume 30.
10. Zhang, B.H.; Lemoine, B.; Mitchell, M. Mitigating Unwanted Biases with Adversarial Learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA, 2–3 February 2018.
11. Wu, Y.; Zhang, L.; Wu, X. Fairness-Aware Classification: Criterion, Convexity, and Bounds. *arXiv* **2018**, arXiv:1809.04737.
12. Berk, R.; Heidari, H.; Jabbari, S.; Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; Roth, A. A Convex Framework for Fair Regression. *arXiv* **2017**, arXiv:1706.02409.
13. Hardt, M.; Price, E.; Srebro, N. Equality of Opportunity in Supervised Learning. *arXiv* **2016**, arXiv:1610.02413.
14. Bellamy, R.K.E.; Dey, K.; Hind, M.; Hoffman, S.C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; et al. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv* **2018**, arXiv:1810.01943.
15. Microsoft Trust in FATE to Keep AI Safe. Available online: https://gulfnews.com/technology/microsoft-trust-in-fate-to-keep-ai-safe-1.2289745 (accessed on accessed on 6 February 2022).
16. Responsible AI: Putting Our Principles into Action. 2019. Available online: https://blog.google/technology/ai/responsible-ai-principles/ (accessed on accessed on 6 February 2022).
17. Madaio, M.A.; Stark, L.; Wortman Vaughan, J.; Wallach, H. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, New York, NY, USA, 29 April–5 May 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 1–14. [CrossRef]
18. Chatila, R.; Havens, J.C. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. In *Robotics and Well-Being*; Aldinhas Ferreira, M.I., Silva Sequeira, J., Singh Virk, G., Tokhi, M.O., Kadar, E., Eds.; Springer International Publishing: Cham, Switzerland, 2019; Volume 95, pp. 11–16. [CrossRef]
19. Weiser, S. Building Trust in Human-Centric AI. 2019. Available online: https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines (accessed on accessed on 6 February 2022).
20. Gershgorn, D. Facebook Says It Has a Tool to Detect Bias in Its Artificial Intelligence. Available online: https://qz.com/1268520/facebook-says-it-has-a-tool-to-detect-bias-in-its-artificial-intelligence/ (accessed on accessed on 6 February 2022).
21. EXPLAIN 2019—ASE 2019. Available online: https://2019.ase-conferences.org/home/explain-2019 (accessed on accessed on 6 February 2022).
22. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* **2021**, *54*, 115:1–115:35. [CrossRef]
23. ACM FAccT. Available online: https://facctconference.org/ (accessed on accessed on 6 February 2022).
24. Kenna, D. Using Adversarial Debiasing to Remove Bias from Word Embeddings. *arXiv* **2021**, arXiv:2107.10251.
25. Kamiran, F.; Mansha, S.; Karim, A.; Zhang, X. Exploiting Reject Option in Classification for Social Discrimination Control. *Inf. Sci.* **2018**, *425*, 18–33. [CrossRef]
26. Chakraborty, J.; Majumder, S.; Menzies, T. Bias in Machine Learning Software: Why? How? What to Do? In Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Athens, Greece, 23–28 August 2021; ACM: Athens, Greece, 2021; pp. 429–440. [CrossRef]
27. Chakraborty, J.; Majumder, S.; Yu, Z.; Menzies, T. Fairway: A Way to Build Fair ML Software. In Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2020, Jabalpur, India, 27–29 February 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 654–665. [CrossRef]
28. Saleiro, P.; Kuester, B.; Hinkson, L.; London, J.; Stevens, A.; Anisfeld, A.; Rodolfa, K.T.; Ghani, R. Aequitas: A Bias and Fairness Audit Toolkit. *arXiv* **2019**, arXiv:1811.05577.
29. Ghosh, A.; Genuit, L.; Reagan, M. Characterizing Intersectional Group Fairness with Worst-Case Comparisons. *arXiv* **2021**, arXiv:2101.01673.
30. Fuster, A.; Goldsmith-Pinkham, P.; Ramadorai, T.; Walther, A. *Predictably Unequal? The Effects of Machine Learning on Credit Markets*; SSRN Scholarly Paper ID 3072038; Social Science Research Network: Rochester, NY, USA, 2021. [CrossRef]
31. Gill, N.; Hall, P.; Montgomery, K.; Schmidt, N. A Responsible Machine Learning Workflow with Focus on Interpretable Models, Post-hoc Explanation, and Discrimination Testing. *Information* **2020**, *11*, 137. [CrossRef]
32. Lee, M.S.A.; Floridi, L. Algorithmic Fairness in Mortgage Lending: From Absolute Conditions to Relational Trade-offs. *Minds Mach.* **2021**, *31*, 165–191. [CrossRef]
33. HMDA Data Browser. Available online: https://ffiec.cfpb.gov/data-browser/data/2020?category=states (accessed on 6 February 2022).
34. Wick, M.; Panda, S.; Tristan, J.B. Unlocking Fairness: A Trade-off Revisited. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Nice, France, 2019; Volume 32.
35. Zhang, Y.; Zhou, F.; Li, Z.; Wang, Y.; Chen, F. Bias-Tolerant Fair Classification. *arXiv* **2021**, arXiv:2107.03207.