Machine Learning Fairness in Finance: An Application to Credit Scoring

Student details

Majda Lalla Kasmi
STUDENT NUMBER: U407044

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee

Supervisor: dr. S. Collin
Second reader: dr. G. Saygili

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
June 2021

Acknowledgement

I would like to thank my supervisor, Prof. S. Collin for her guidance, advice, and words of encouragement throughout the trajectory of this research. I would also like to thank my friends and family for their moral support, as well as Tilburg University for their flexibility during these times, continuously providing high quality education online and on campus.

Machine Learning Fairness in Finance: An Application to Credit Scoring

Majda Lalla Kasmi

Abstract

Machine learning has already established itself in the world of finance and is more and more used in risk analytics and credit scoring. As automated credit scoring is further developing itself through the use of more data and more advanced machine learning models, fairness in this application is still lagging behind. Research on inducing fairness has been done and multiple methods have been proposed, however not specifically tailored to this context. In this paper we apply preprocessing and postprocessing methods for inducing fairness in an effort to compare their effectiveness in a credit scoring application.
The results suggest that Preferential Sampling perform best in the pre-processing for fairness category, while Calibrated Equal Odds performs best in the post-processing category. Overall, it is concluded that both pre-processing and post-processing methods can be used effectively for addressing discrimination in the automated credit-scoring application.

1. Introduction

The aim of this research is to identify effective methods for bias mitigation and fairness induction in the machine learning application for credit scoring in the financial services industry. Machine learning has already made it possible to automate the credit scoring process through the use of classification models. Individuals are either classified as creditworthy or not creditworthy. Dastile et al. (2020) elaborate on the current models used industry wide, while also shedding a light on opportunities for the future. From the literature it is evident that machine learning and its application is quickly evolving (Lessmann et al., 2015). The application of credit scoring is no different, resulting in benefits such as time efficiency. While the field is rapidly evolving, there is limited expertise in machine learning fairness. Biases are easily introduced to a model because it is often trained on datasets, containing data from previous clients and past decisions, resulting in models that discriminate against marginal groups/ minorities. (Prabhakar & Weber, 2020) With automated decision making applied in situations that can majorly impact the lives of those the decisions are made about, such as loan approvals, it is critical that decisions are made fairly. The study of machine learning fairness itself is, relatively new, with little research on the application in industry specific situations (Mehrabi et al., 2019). This while, regulators increasingly emphasize the importance of fairness and propose and impose legislation, requiring financial institutions to not only impose fairness but also prove the fairness of their practices (Aggarwal, 2021).

In light of the more and more demanding regulatory requirements for financial institutions and the gap in the literature, the following research questions were formulated:

Research Question 1: Do sensitive attributes like gender and age influence decision making in the application of credit scoring and how can this be measured? This question will be answered by identifying fairness metrics and applying them to an industry standard machine learning method for credit scoring, specifically a logistic regression model.

Research Question 2: What fairness inducing method(s) cause(s) sensitive attributes such as age and gender to be of the smallest influence on the decision making? This question will be answered by applying several methods for fairness, both preprocessing and postprocessing, to our industry standard model and comparing them based on model performance and fairness.

The results showed that in credit scoring, group attribution does influence the probability of being assigned a positive classification and thus deemed "creditworthy". Our findings suggest that the use behavioral data, reduces the presence of discrimination, while this cannot be said definitively. The results also suggest that Preferential Sampling perform best in the pre-processing for fairness category, while Calibrated Equal Odds performs best in the post-processing category. Overall, it is concluded that both pre-processing and post-processing methods can be used effectively for addressing discrimination in the automated credit-scoring application.

The structure of the paper is as follows. In section 2 an inclusive theoretical background on the topic is provided. Section 3 then covers the different methods and metrics used for the purpose of answering the research questions. Section 4 describes the experimental set up in several steps, including pre-processing, model building, and testing. In section 5 the results are presented in detail. After which a discussion of the results in relation to the research questions will be presented in section 6.

2. Related Work

Machine learning in the application of credit-scoring is not uncommon practice, with research on effectiveness of several machine learning methods already taking place in 2003 (Baesens et al.). Credit scoring is a supervised learning problem, often a binary classification problem, aiming at identifying good borrowers and bad borrowers (Dastile et al., 2020). Other terms used are credit approval versus credit rejection and creditworthiness versus non-creditworthiness. Though automated credit-scoring is a well-studied field, fairness in this application is less studied. Indeed, research in fairness methods have included credit-scoring data but have not looked into specific requirements for industry implementation and practicality (Kamiran & Calders, 2011) (Calmon et al., 2017).

2.1 Fairness

Generally speaking, fairness in society is discussed in terms of non-discrimination and equality. Britannica (2020) describes discrimination as "the intended or accomplished differential treatment of persons or social groups for reasons of certain generalized traits. (…) For the most part, discrimination results in some form of harm or disadvantage to the targeted persons or groups." Equality can be defined as "an ideal of uniformity in treatment or status by those in a position to affect others" (Britannica, 2020). Oneto et al. (2020) describe Machine learning fairness as the study of methods to guarantee no unfavorable treatment of individuals by models, based on sensitive characteristics, such as gender, age, sexual or political orientation. Though there is currently no consensus on what fairness in machine learning means, several definitions have been set out as well as measures of fairness that can be used for the evaluation

of machine learning models. (Jones et al., 2017) (Speicher et al. (2018). Before looking at such measures, we will explore the current understanding of how unfairness in machine occurs.

Data that is used for training algorithms can contain many different types of biases, influencing the decision making. Mehrabi et al. (2019) created an overview of potential biases in data, including historical bias. Historical bias occurs when existing societal biases as well as socio-technical issues are included in the data generation process, regardless of accurate sampling or adequate feature selection. Furthermore, the use of unbalanced datasets for training algorithms can result in the presence of biases against underrepresent groups when employing the algorithm, the latter are usually dealt with through the use of reweighing or resampling methods. Dealing with the first however, can be more challenging, yet important. When the bias is not adequately handled, the resulting model, trained on the generated data, will likely be discriminatory in its classification. Though there are many more different biases that could potentially lead to unfair models, the focus of this research will be on the reduction of historical and societal bias in the decision-making process of machine learning models, and therefore this section will herein be limited

Dealing with biases is a difficult task, not in the least because there is no one definition of fairness. In their paper Mehrabi et al. (2019) discuss 10 different definitions of fairness. Amongst them are Fairness through (un)awareness, treatment equality, test fairness and counterfactual fairness. They state that all fairness definitions fall in one of three groups, being individual fairness, group fairness and subgroup fairness. The first two groups respectively mean, classifiers give similar predictions to similar individuals and different groups are treated equally. Subgroup fairness aims at using the most suitable properties of both individual and group fairness concepts to create an optimal definition. Both individual and group fairness are relatively well researched types of fairness, though current available methods for fairness mainly deal with group fairness. Methods dealing with subgroup fairness are much rarer. This can be explained by the trade-off existing when it comes to fairness. When ensuring group fairness is present, it comes at the cost of individual fairness. Even when choosing a group fairness definition, there is a trade-off to be made between different measures of group fairness or fairness and accuracy(Valdivia et al. 2021). In their paper, Kleinberg et al. (2016) prove and explain this phenomenon. Two guarantees an algorithm should ideally have for it to be considered fair are described. Subsequently it is proven that these guarantees or definitions of fairness are incompatible. Though this proof is far out of the scope of this research, we will discuss the two guarantees that represent different forms of group fairness. Firstly, Kleinberg et al. (2016) explains that a model should be "well-calibrated", meaning that if a model identifies a sample set to have a certain probability z to have positive classification, then approximately a z fraction of said group should have a positive classification. This then should also hold for subgroups in a sample. If e.g., men are found to have a certain probability z to obtain positive classification, this should mean that z percent of men have this classification. Secondly, there should be, what they call a "balance for the positive class". People belonging to different groups should have on average the same probability for positive classification. The reason for this is that if one group (e.g., men) consistently receives higher probability estimates, another group (e.g., women) will consequently receive lower probability estimates. The same holds for the "balance of the negative class". In practice what this means is that there should be both **statistical parity** and there should be **equal rates of false positives and false negatives between groups.** Fairness will be further elaborated on in the next part.

2.1.2 Defining and measuring fairness

In this part, we will further elaborate on fairness in a binary classification context. We will measure discrimination based on binary group attribution. One either belongs to the privileged (unprotected) group or the unprivileged (protected/ deprived) group.

Group fairness

Choulderou (2017) also discusses fairness and provides four definitions: calibration, predictive parity, error rate balance and statistical parity. This in the application of risk assessment for post-conviction supervision in the United States. These formulas can be generalized, effectively resulting in a further split of the definitions mentioned above. Considering a dataset (X, Y), calibration is described as follows:

$$P(Y = 1 \mid S = s, A = 0) = P(Y = 1 \mid S = s, A = 1)$$

Where S = S(*x*) translates to the probability score of being classified in a binary classification 1 (privileged class). $A \in \{0, 1\}$ indicating the sensitive attribute with groups *0* and *1* an individual can belong to. $Y \in \{0, 1\}$ denote the classification labels. Where 1 equates a positive classification. Calibration means that the classification outcome should be independent of the sensitive attribute *R* conditional on the probability of positive classification. In other words, *S(x)* is considered to be well calibrated when it represents the same score irrespective of an individual's group membership in *A.* This is in line with the definition provided by Corbett-Davies et al. (2018)

Predictive Parity is satisfied if individuals, irrespective of their group ascription, have equal positive predictive value (PPV) of classifier *Y = 1* given threshold *t* and can be denoted as:

$$P(Y = 1 \mid S > t, A = 0) = P(Y = 1 \mid S > t, A = 0)$$

Values of S above the threshold will be qualified as *Y(1)*. While PPV is denoted as:

$$\frac{True\ Positive}{True\ Positive + False\ Positve}$$

Simply put we speak of predictive parity when protected and unprotected groups have equal probability of belonging to Y(1) the positive class at a threshold *t* for *S*. Though it is similar to calibration it is not actually the same. At different thresholds, predictive parity can be failed to uphold. Calibration is different in that it considers all values of *S.* (Verma et Rubin, 2018)

Error Rate Balance simply means that false negative and false positive rates are equal in both the protected and unprotected group where $S \leq t$ and $S > t$ describe the prediction decision at a threshold *t*:

$$P(S > t \mid Y = 0, A = 0) = P(S > t \mid Y = 0, A = 1)$$
$$P(S \leq t \mid Y = 1,\ A = 0) = P(S \leq t \mid Y = 1,\ A = 1)$$

Statistical Parity is satisfied if individuals in both protected and unprotected groups have the same probability of being positively classified. Choulderou (2017)

$$P(S > t \mid A = 0) = P(S > t \mid A = 1)$$

Individual fairness

In addition to the above-mentioned fairness measures all pertain to group fairness, individual fairness can also be defined. One such definitions is causal discrimination. If two individuals from different groups have the same classification when all other attributes X are the same (Verma & Rubin, 2018).

In the same paper they also introduce "fairness through awareness" and argue that similar individuals should have similar classification predictions, they call this is fairness through awareness. It is measured employing a distance metric and formally results in the assumption that the distance between the distribution of outputs for individuals ought to be at most the distance between the individuals.

2.2 Tools and methods for inducing fairness

2.2.1 Pre-processing methods

Pre-processing for fairness machine learning is the processing of data before using it in a machine learning model so as to minimize the influence of bias and sensitive attributes. After the processing of the data, the machine learning model can then be trained on the adjusted data. (Oneto et Chiappa. 2020). (Kamiran et Calders, 2011) discussed several methods for this purpose, describing it as discrimination-aware classifiers. They concluded that though effective, bias reducing methods come at the cost of accuracy. Calmon et al. (2017) proposed a different model, building an optimization framework with the purpose of probabilistically transforming data. They also found that, while effective, it does come at the cost of accuracy.

2.2.2 In-processing methods

In-processing methods impose fairness constraints in the learning process of a model in order to enforce model fairness. Hardt et al.(2016) propose a so-called Equality of Opportunity method that can be applied in supervised learning, they state that they designed a "fairness measure that accomplishes two important desiderata. First, it remedies the main conceptual shortcomings of demographic parity as a fairness notion. Second, it is fully aligned with the central goal of supervised machine learning, that is, to build higher accuracy classifiers" (Hardt et al. (2016). Kusner et al. (2017) propose a method named "Counterfactual Fairness". According to them "It allows us to propose algorithms that, rather than simply ignoring protected attributes, are able to take into account the different social biases that may arise towards individuals based on ethically sensitive attributes and compensate for these biases effectively". Chiappa(2018) builds on this by introducing a path -specific element meaning, fairness occurs if the decision made, would have been the same in a counterfactual world (for sensitive attribute) along the unfair pathways(Chiappa, 2018).

2.2.3 Post-processing methods

Post-processing aims at the same results as the previous methods but does so after running the model. In the literature when looking at post-processing models, we first notice that this domain has been researched the least and the literature tends to suggest models for improving group fairness. Lohia et al. (2019) introduced a model for improving both individual and group fairness. They found their model to be effective, while only having a limited impact on accuracy. Other models include those introduced by Hard et al. (2016), Pleiss et al (2017) and

Kamishima et al., 2012. These methods will be further elaborated on in section 3.

2.3 Machine learning based decision making in credit scoring

Credit scoring (CS) started being widely used after the arrival of credit cards in the 1960s. With credit card applications rising rapidly, automation of the process was needed to meet the demand. After witnessing the success rates of credit scoring with credit cards (a decrease of 50% in defaults), banks started using credit scoring for other products such as personal loans (Thomas et al., 2002). Since then, there have been many developments in the process of CS and many methods have been researched (Trivedi,(2020), aiming at predicting the credit worthiness of individuals and organizations. The ultimate purpose of CS is to reduce the number of loans resulting in default, and therefore manage the risk of credit providers.

The study by Baesens et al. (2003) is regarded in the literature as the most highly valued benchmarking classification study and is therefore most used in the application of credit scoring. These are logistic regression (LOG), linear and quadratic discriminant analysis, linear programming, support vector machines (SVM), neural networks (NN), Naïve Bayes (NB) and k nearest neighbor (KNN) classifiers. It was found that the more complex methods such as SVM and NN classifiers perform well when measured in regards of PCC(Probability of correct classification) and AUC (Area under the ROC Curve), they do not perform significantly better than the more straightforward models, like linear discriminant analysis (LDA) and LOG. These models are thus competitive with one another. Years later the study was revisited by Lessman et al. (2015) where they compared 41 different algorithms that can be split into three groups: individual classifiers, homogenous ensemble classifiers and heterogenous classifiers. The first group uses a single classifier to develop a classification model, the latter two combine the prediction of multiple models to come to a final prediction. The difference between them lies in the fact that homogenous ensemble classifiers use multiple base models, whereas the heterogenous ones use different classification models. Lessman et al. (2015) conclude that some models, especially heterogenous ensemble classifiers perform better than the logistic regression classifier. The latter is used more industry wide. They also found that ANN classifiers function better than ELM ones and RF achieves better results than RotFor. Dynamic selective ensembles do worse than most other classifiers. Furthermore, Destile et al. (2020) found that good benchmarking models in the application of CS are Logistic regression and decision trees, as the LOG performs equally as well as traditional machine learning models, while DT have high explainability. It is important to note that a better performance was observed for deep learning models, compared to their statistical, and traditional machine learning counterparts.

Bequé et Lessmann (2017) revisited the ELM model and examined its potential by focusing on three criteria: ease of use, computational complexity, and predictive performance. Their findings show that ELM performs competitively with regard to other benchmark models. Though, some limitations of the model remain; random weight initialization used in ELM causes variation and the use of other sampling methods in the application of the model are not yet well researched.

2.3.1 Fairness in credit scoring

Much like there is no one definition of fairness in machine learning, there is no one method or rule to classify an individual's creditworthiness and predict their likelihood of repayment. As a

result, different companies will use a different set of attributes and requirements for classification. Aside from the use of the more traditional data such as income, number of dependents and disposable income, companies might also use behavioral data for decision making. Especially banks who have such information readily available, through their customers' past spending habits, might make use of this. Though one might think that this would lead to less biased decision making, this is not necessarily true. Hurley et Adebayo (2016) illustrate this with the following example. Kevin Johnson, who is black, is an exemplary credit customer. He has never missed a payment and practices responsible spending. Yet, his credit limit was reduced with more than 65%. American Express had classified him a risk because he had recently spent money at shops where other customers with poor repayment history had used their cards. Though some call this behavioral scoring, a better suited name is "creditworthiness by association." "Rather than being judged on their individual merits and actions, consumers may find that access to credit depends on a lender's opaque predictions about a consumer's friends, neighbors, and people with similar interests, income levels, and backgrounds" (Hurley et Adebayo, 2016). More traditional behavioral scoring entails, besides the traditional credit scoring attributes, like the ones mentioned earlier, repayment, and ordering history (Thomas & Crook, 2000). The extent to which the ordering history or payment behavior is used can thus clarify whether a person's spending habits are used for classification directly or if there's classification by association , which leads to a further instatement of bias. Hurley et Adebayo (2016) compare the situation of Kevin Johnson to the 'redlining" practice. In redlining, individuals are classified based on their postcodes instead of their ability to repay a loan, resulting in racial discrimination, as excluded postcodes were usually those of areas where relatively more ethnic minorities lived. Consumer's risk being punished for participating in activities that are in association with ethnic, racial, or socioeconomic groups.

3. Methods

In this section we will describe the methods used for this research, the reasoning/ motivation for the choice of methods and the metrics employed. We will first introduce the baseline model, that we will use as a current standard to which we will compare the other models. Then we will elaborate on the methods and models chosen for inducing fairness as well as which exact measures will be used for quantifying fairness.

3.1 Fairness methods: Automated Credit Scoring and Payment Default Prediction

For the purpose of this research, we will look at automated credit scoring as well as payment default prediction through the use of supervised learning classification methods. The baseline model will be a logistic regression. The logistic regression model will allow us to compare methods addressing societal bias in machine learning to simply ignoring it, in order to come to conclusion about the effectiveness of such bias reduction methods.

3.1.1 preprocessing methods

several preprocessing methods will be used. The first of them is called *suppression.* In suppression the sensitive attributes are simply removed from the dataset. It is generally understood that this is not the most effective method for bias reduction as it does not deal with correlated attributes and effects of belonging to a group whether protected or not, can influence the other attributes in less obvious ways. Being a woman for example might influence career

paths available and will thus influence job and income (Kusner et al, 2017). Nonetheless it serves as a good second baseline model as we address the bias directly visible. This not only for the pre-processing methods but also the post-processing methods.

Secondly, we will use a method called *massaging* that was introduced by Kamiran et Calders (2011). In essence what this method does is, it relabels some observations to remove discrimination in the dataset. Observations with $A = 0$ (the sensitive attribute) also referred to observations in the deprived community, will be changed from a negative to a positive classification (promoted), while observations with $A = 1$ will be changed from a positive to a negative classification (demoted). The observations that are changed are not chosen at random. A ranker orders observations based on their positive class probability, where higher scores mean a higher chance of belonging to a positive class. Observations from A = 0 will be ordered in descending order while the others will be ordered in ascending order. those with the highest probability in the promotion group(being those with a protected attribute) will be promoted and those with the lowest probability in the demotion group will be demoted. The number of observations that need to be changed is determined by the formula:

$$M = \frac{disc(D) \times |D_0| \times |D_1|}{|D|}$$

Where D_0 denotes the observations in D (dataset) with A = 0 (protected attribute) and A = 1 (unprotected attribute), respectively. Discrimination *(disc(D))* is defined by Kamiran et Calders (2011) as follows:

$$disc_{A=0}(D) = \frac{|\{X \in D \mid X(A) = 1, X(Class) = +\}|}{|\{X \in D \mid X(A) = 1\}|} - \frac{|\{X \in D \mid X(S) = 0, X(Class) = +\}|}{|\{X \in D \mid X(A) = 0\}|}$$

Where discrimination with respect to the deprived community (those with a protected attribute) is measured. The formula represents the difference of the probability of being in the positive class between those belonging to the deprived community and those who do not.

It is important to note that massaging is only applied to the training data. After the class labels are changed, the model can be fitted to the adjusted training dataset.

Our third preprocessing model is *preferential sampling*, also introduced by Kamiran et Calders (2011) It is a sampling approach that duplicates and removes samples based on their ranking and builds on the discrimination formula presented above. In this method it is assumed that objects close to the decision boundary of a model are more likely to be discriminated against or favored, depending on whether they belong to the deprived community or not. In figure 1 we see this visualized. First observations are split into four groups:

$DP: = \{X \in D \mid X(A) = 0 \wedge X(Class) = +\}$

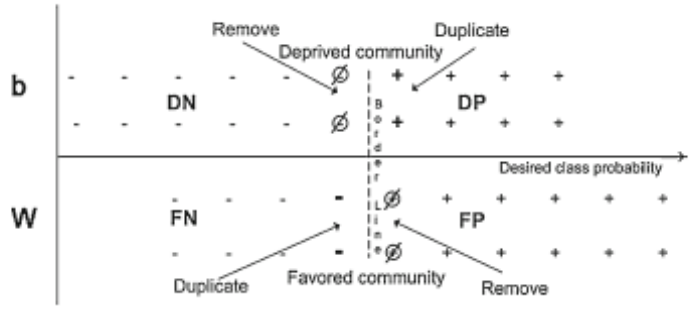
$DN: = \{X \in D \mid X(A) = 0 \wedge X(Class) = -\}$

$FP: = \{X \in D \mid X(A) = 1 \wedge X(Class) = +\}$

$FN: = \{X \in D \mid X(A) = 1 \wedge X(Class) = -\}$

DN and DP represent those in the deprived community with a negative and positive classification, respectively. FN and FP represent those in the favored community with negative and positive classification, respectively.

Figure 1
Visualisation of the organization of observations in accordance with the Preferential Sampling method. b represents the protected group and w the unprotected group.



Kamiran et Calders (2011)

First the observations are ranked, using the same ranker as in the the massaging method. However, DP and FP are ranked in ascending order, while DN and FN are sorted in descending order.

Preferential sampling uses the original training set and iteratively duplicates for the groups DP and FN and removes observations for DN and FP groups. It does so adhering to the following two rules.

1. Decreasing a group size happens through the removal of an observations, closest to the decision boundary, meaning those with the highest ranking.
2. Increasing a group size is done by duplicating the observations closest to the boundary. (those with the lowest scores in the ranking list).

Duplications and removals are done till all groups are of equal size. The desired size of each group is the total amount of observations divided by four. When the group sizes are more "beneficial" for the sensitive attribute, then no adjustments are made, this because the model aims at only addressing situations in which fairness for observations $A = 0$ are reduced.

3.1.2 post processing methods

In their paper Hard et al. (2016) propose a post-processing model called *equalized odds*. In essence their method strives at creating a predictor that satisfies equalized odds. This is satisfied when $\hat{Y}$ (predictions) and protected attribute A independent conditional on Y (true labels.

In the binary application we use this looks as follows:

$$P\{\hat{Y} = 1 \mid A = 0, Y = y\} = P\{\hat{Y} = 1 \mid A = 1, Y = y\}, \qquad y \in \{0,1\}$$

There should be equal true positive rates for both *A = 0* and *A = 1* as well as equal false positive rates. Therefore, it can be said that in equal odds both equal bias and equal accuracy are enforced. The model also considers equal opportunity, in which non-discrimination is only required in the privileged group. The formula for this is written below. This is a more relaxed measure than equal odds and can be useful in some situations, where equal odds is difficult to achieve.

$$\mathrm{P}\{\hat{Y} = 1 \mid \mathrm{A} = 0, \mathrm{Y} = 1\} = \mathrm{P}\{\hat{Y} = 1 \mid \mathrm{A} = 1, \mathrm{Y} = 1\}$$

Based on these definitions Hard et al. (2016) formulate a derived predictor $\tilde{Y}$. The predictor is applied after the model is ran and is oblivious in that it only considers,Y, $\hat{Y}$ and A.

Calibrated equalized odds is a method proposed by Pleiss et al (2017). While trying to obtain equal odds as in the abovementioned method, calibration is taken into consideration. Besides error types not disproportionately affecting protected groups, in their paper, they also take into account the proportion of individuals receiving positive and negative classifications, irrespective of the subgroup they belong to. This is the calibration. It was concluded that satisfying both calibration and equal odds fully is not possible. It is therefore necessary to choose a cost constraint i.e., while satisfying equal odds, it is possible to either satisfy equal false negative, false positive, true positive or true negative rates, but never all at once. Additionally, it is assumed that different subgroups (in our A) have different base probabilities (μ) to belong to a group. This looks as follows:

$$\mu 1 = P(x, y) \sim (A = 1)\,[y = 1] \neq P(x, y) \sim G2\,[y = 1] = \mu 2$$

Then given a cost constraint, calibrated classifiers h_1 and h_2 can be determined (the base probabilities are determined based on a validation set, not the training set). The classifiers are decided on while satisfying $A_0 h_1 = A_1 h_2$. One classifier belongs to the protected group A_0 and the other to the privileged group A_1. The classifiers are then applied in the same manner as in the equalized odds method, taking into account only $A, \hat{Y}$ and Y.

The last method that will be used in this research is called *Reject Option Classification*. This method was introduced by Kamishima et al., 2012. This method is rather straightforward. The method considers the classification probabilities, where probabilities closer to 1 have a higher degree of certainty. Based on this principle, it defines a critical region between (0,5 and a certain threshold t) where observations for which $max[P(Y = 1|\,X), 1 - P(Y = 1\,|X)] \leq t$ where $(\,0.5 < t < 1)$ is the critical region. Observations that fall in that region and are labeled reject. If an instance from the protected group is labeled reject, then it receives a positive classification, whereas those in the other group are negatively classified.

3.2 Model and method measures

3.2.1 Model performance

the model performances will be measured through accuracy, precision, recall and specificity. We will also look at balanced accuracy as inducing fairness often involves relabeling values in one way or another, possibly leading to imbalances. In our application we ideally would like to have a low *false positive* rate as false positives lead to losses for loan providers. Keeping the *false negative* low benefits loan providers as more profitable loans are then offered. As such the recall value is especially important.

3.2.2 Fairness measures

Fairness will be measured along different definitions. Namely, statistical parity difference, disparate impact, average odds difference and equal opportunity difference. Below these measures will be further elaborated on. We will not look into individual fairness measures as we address fairness associated with belonging to different groups, e.g., gender. These fairness metrics were retrieved from the AIF360 python library (Bellami et al. 2018).

- *Statistical parity difference* measures the difference in probability for the protected group to be classified positively and the probability for the unprotected group to be classified positively:

$$P(\hat{Y} = 1|A = 0) - P(\hat{Y} = 1|A = 1)$$

  A score of 0 is ideal, while negative scores mean that discrimination against the protected group occurs. Positive values indicate positive discrimination.

- *Disparate impact* is the ratio of positive classifications between the protected and unprotected group, where 1 is ideal and values below 1 indicate discrimination against the protected class. Values above 1 indicate positive discriminations :

$$\frac{P(\hat{Y} = 1|A = 0)}{P(\hat{Y} = 1|A = 1)}$$

- *Average odds difference* is the average difference in false positive and true positive rates in privileged and unprivileged groups. When the value for this measure is 0 it means that both groups are treated fairly. FPR is the *false positive* rate and TPR is the *true positive rate* . A negative value indicates discrimination against the protected group*:*

$$\frac{FPR_{A=0} - FPR_{A=1}) + (TPR_{A=0} - TPR_{A=1})}{2}$$

- *Equal opportunity difference* measures the difference in true positive rates between protected and unprotected groups, where 0 values are ideal and negative values indicate discrimination against the protected group:

$$TPR_{A=0} - TPR_{A=1}$$

4. Experimental Setup

In this part the experiment procedures, underlying steps and tools used will be described. First, we elaborate on the chosen datasets, after which we discuss the data pre-processing and parameter tuning for training the baseline model. Then this section will expand on the chosen fairness methods and how they were applied.

4.1 Model data

For this research two datasets were used, both retrieved from the UCI Machine Learning Repository. The first dataset is used for credit approval decision making, the second is used for default payment prediction, still both datasets speak to the financial credibility of individuals. The fact that both datasets make use of classification in a lending/borrowing context and both contain binary classification, while also containing sensitive attributes, make it suitable datasets for this research. Additionally, considering the scarcity of datasets in these applications, it is extra valuable that both datasets have a traceable source, assuring legitimacy.

4.1.1 Datasets

The first dataset used for this research is the **German Credit Data**. It was made available by Hamburg University and was compiled in 1994 for the use of binary classification, more exactly, the classification for having either good or bad credit risk. There are 1000 instances present and 20 attributes, most of which are relevant for determining the credit risk of an individual, such as credit amount requested, purpose of the loan requested and current bank balance. Furthermore, there are three attributes present on the basis of which discrimination could happen, these are referred to as sensitive attributes and they are, age, gender and "foreign worker". The latter describes whether the individual is an immigrant worker. Unfortunately, this attribute cannot be included in this research as only 3,84% of the instances concern a foreign worker.

The second dataset used is the **Default of Credit Card Clients Dataset**, also from the UCI Machine Learning Repository. Though the dataset contains similar attributes to the first one (gender, education, marital status, and age, its majority consists of attributes, describing payment behavior. These attributes are the presence of late payments, amount of bill statements and amounts paid in different months. Furthermore, it contains 30.000 instances and 24 attributes. Binary classification results in credible or not credible clients.

4.2 Pre-processing

The data pre-processing before applying the models and methods was fairly straight-forward. The attributes in table 1 were initially present for each dataset.

Table 1
Attributes per dataset before pre-processing the data.

German Credit Data	*Default of Credit Clients Data (Taiwan)*
Status existing checking account	Amount of given credit
Duration in months	Gender
Credit history	Education
Purpose (of requested credit)	Marital status
Credit amount	Age
Savings account/ bond	The repayment status in September, 2005
Present employment since	The repayment status in August, 2005
Installment rate in percentage of disposable income	The repayment status in July, 2005

Personal status and sex	The repayment status in June, 2005
Other debtors/ guarantors	The repayment status in May, 2005
Present residence since	The repayment status in April, 2005
Property	Amount of bill statement in September, 2005
Age in years	Amount of bill statement in August, 2005
Other installment plans	Amount of bill statement in July, 2005
Housing	Amount of bill statement in June, 2005
Number of existing credits at this bank	Amount of bill statement in May, 2005
Job	Amount of bill statement in April, 2005
Number of people being liable to provide maintenance for	Amount paid in September, 2005
Telephone	Amount paid in April, 2005
Foreign worker	Amount paid in September, 2005
	Amount paid in August, 2005
	Amount paid in April, 2005
Classification attribute	
0 = no credit approval 1 = credit approval	0 = no payment default 1 = payment default

First, in the German Credit Data, all categorical variables (which are most) were changed from object to integer in order to allow for easier application of the models. Secondly, 'personal status and sex' were separated in order to isolate the sensitive attribute of *sex*. Lastly, the objects '*foreign worker*' and '*personal status*' were removed. This, because the first was only present in 3,84% of the instances and the latter contained unequal information for men and women. For men it was registered whether they were 1) married/ widowed, 2) divorced/ separated or 3) single, while for women the only options were 1) single or 2) married/widowed/separated. After checking for missing values and outliers, the data was separated in a train (60%), validation(20%) and test set(20%) for models where needed, in other cases the dataset was simply split in a train (80%) and test set (20%) . Lastly, because there was an imbalanced class distribution of approximately 30% vs 70%, Synthetic Minority Over-sampling (SMOTE) was applied. SMOTE creates synthetic observations of the minority class, instead of simply duplicating them. This technique was chosen over other options because undersampling the dataset by at random removing observations form the majority class would make the small dataset even smaller, while using oversampling by randomly duplicating observations could more easily lead to overfitting of the models, worsening the performance.

For the Credit Card Clients Data Set fairly little preprocessing was needed. The positive and negative classes were switched, resulting in 0 = payment default (not credible) and 1 = no payment default (credible) in order to create uniformity in the datasets, but also because we are interested in identifying good clients for lending credit to. The label name was changed to "creditworthiness', representing the new classifications. Lastly, smote was also applied here as there was approximately 20% vs 80% imbalanced class distribution. All pre-processing was done in Python, using the libraries Scikit-learn, NumPy, Math and Pandas.

4.3 model building

All our models use all attributes present in the dataset after pre-processing, unless stated differently, as they are all deemed important for credit scoring. SMOTE was also applied before all fairness models as we wanted to maintain as much consistency as possible. The preprocessing fairness methods were all exclusively applied to the training set. The postprocessing methods were applied to the predictions made

We first built our baseline model, which was a logistic regression in scikit-learn, using different penalties, solvers and max iterations to come to an optimal model. This was done for both datasets. After building our model we applied our first fairness model. This was the suppression model. The only difference between the baseline model and the suppression model, was that the sensitive attributes (sex and age) were removed.

For the other two preprocessing fairness models, the code was built from scratch, following the explanations as provided in the methods section. This was done using the Scikit-learn, Math, Pandas and NumPy libraries the same rankers were used for both models (a knn- classifier with 15 nearest neighbors). The massaging model was built by first calculating the discrimination present in the datasets. Then the size of the protected group was counted, after which the size, M, for observations that needed to be massaged was calculated. The ranker was used to calculate probabilities for positive classification to all observations. Before massaging was applied on the unprivileged group, the dataset was organized in descending order of probability for positive classification (provided by the ranker). For the privileged group the opposite was done.

The Preferential sampling method was applied similarly to the massaging method. First the ranker assigned probabilities for positive classification to the observations. The dataset was then sorted for each group DN, DP, FN and FP as the method prescribes. Then group sizes for these groups were identified. Considering the desired group sizes, observations were identified either for duplication, deletion. Those for duplication were labeled 1, while those for deletion were labeled 2. Observations that did not require action were labeled 0. After the resampling was done, the logistic model was fitted on the new training data.

All fairness post-processing methods were built using the AIF360library, created by Bellami et al. (2018). This library provides metrics and methods for imposing fairness. The methods it has available are built based on methods proposed in the literature, including the postprocessing models chosen for this research. The library requires the use for special datasets in which sensitive attributes are identified, as well as the labels. The datasets were therefore transformed into the so called ‘BinaryLabelDataset’.

4.4 Metrics

We also used the AIF360 library for obtaining our fairness metrics. The preprocessing data, which was not yet in the proper format, were transformed to ‘StandardDatasets’, which the library also support. The full code for the postprocessing and metric calculations can be found on https://github.com/majdaksm/Thesis-Machine-Learning-Fairness All other code pertaining to this research can be found there too.

4 Results

In this section we will discuss the traditional performance of the fairness induction models against the baseline model and each other. We also will have a look at the level of fairness of each model, considering the two different datasets. We do so by assigning all models a score, where 0 is neutral. For each performance measure, the difference in performance between the baseline model and the fairness model will be calculated. All individual scores will be added, yielding a final score. Positive scores mean that the fairness model performed better. Negative scores mean they performed worse.

4.2 Model performance

In table 2 the performance scores for the different models on the *German Credit Data* can be seen. The results are separated for the two different sensitive attributes 'age' and 'sex'. Averages are also provided. The individual performance measure scores per sensitive attribute can be found in Appendix B. From the table we can read that performance is affected slightly by the induction of fairness models. Massaging has the biggest negative effect on performance, while Suppression and Preferential Sampling, seem to slightly improve the performance. All post-processing models slightly affect performance. In table 1 of Appendix B we can see that accuracy is not affected too much, while sometimes even improving. When solely looking at accuracy, Massaging negatively influences performance most. Especially when addressing age.

Table 2

Performance scores per model for the German Credit Dataset, separated on fairness induction for the two different sensitive attributes in the dataset 'sex' and 'age. The protected groups are respectively female and old age group.

German Dataset			
model	**performance score**		**Average**
sensitive attribute	***Sex***	***Age***	
Logistic Regression	0	0	
Pre-processing models			
Suppression	0	0.04	0.02
Massaging	-0.13	-0.4	-0.265
Preferential Sampling	0.11	0.06	0.085
Post-processing			
Equalized Odds	-0.13	0.01	-0.06
Calibrated Equalized Odds	-0.02	-0.07	-0.045
Reject Option Classification	0.05	-0.31	-0.13

In table 3 the performance scores for the Default of Credit Clients Data (Taiwan) are presented. Pre-processing models slightly affect performance, with the exception of suppression which resulted in a very small improvement. However, this improvement is insignificant. Unlike the in the application on the German Credit Dataset, post-processing significantly improved

performance. In particular Reject Option Classification and Calibrated Equalized odds caused better performance. When solely looking at accuracy (Appendix B, table 2), a different picture is painted. Equalized Odds negatively affected accuracy, Calibrated Equalized Odds left accuracy unchanged. Reject Option Classification improved accuracy by 0.04.

Table 3

Performance scores per model for the Default of Credit Clients Data (Taiwan), separated on fairness induction methods for the two different sensitive attributes 'sex' and 'age in the dataset. The protected groups are respectively female and old age group. Unlike the German Credit Data, this dataset considers behavioral scoring.

Default of Credit Card Clients Data			
model	**performance score**		**Average**
sensitive attribute	***Sex***	***Age***	
Logistic Regression	0	0	
Pre-processing models			
Suppression	-0.02	0.03	0.005
Massaging	-0.02	-0.04	-0.03
Preferential Sampling	-0.05	-0.03	-0.04
Post-processing models			
Equalized Odds	0.03	0.14	0.085
Calibrated Equalized Odds	0.27	0.21	0.24
Reject Option Classification	0.29	0.32	0.305

4.3 Fairness performance

Fairness was measured using Statistical Parity Difference, Disparate Impact, Average Odds Difference and Equal Opportunity difference. The final fairness scores for the German credit Dataset can be found below (table 4). We found that Suppression negatively influenced fairness rather significantly, as did Massaging. Preferential Sampling improves fairness quite significantly, across the measures, there is an improvement of 36% . When looking at the individual measures (Appendix B, table 4) , this improvement in fairness can be ascribed to Statistical Parity Difference. Based on this measure, it can be concluded that positive discrimination is taking place, where the protected group is being favored in classification. In the post-processing methods, Calibrated Equal Odds, is the only one improving fairness, this by 0.33 . This can be explained by the improvement in disparate impact. While the Reject Option Classification model performs well on the sensitive attribute 'sex', it performs worse than the baseline on the 'age' attribute. The Calibrated Equal Odds model does not have this issue.

Table 4

Fairness scores per model for the German Credit Dataset, separated on fairness induction for the two different sensitive attributes in the dataset 'sex' and 'age. The protected groups are respectively female and old age group.

German Credit Dataset			
model	**Fairness score**		**Average**
sensitive attribute	***Sex***	***Age***	
Logistic Regression	0	0	0
Pre-processing models			
Suppression	-0.37	-0.49	-0.43
Massaging	-0.81	-3.68	-2.25
Preferential Sampling	0.68	0.03	0.36
Post-processing models			
Equalized Odds	-0.17	-0.46	-0.315
Calibrated Equalized Odds	-0.04	0.7	0.33
Reject Option Classification	0.18	-0.54	-0.18

Table 5 shows the fairness scores on the behavioral dataset. With the exception of the Suppression, the Pre-processing methods do improve fairness, the improvement by the Massaging model can be neglected. Calibrated Equal Odds, performs best in the post-processing methods, again because it positively discriminates in favor of the protected group. In particular the Disparate Impact has a high score. (Appendix B, table 4)

Table 5

Fairness scores per model for the Default of Credit Card Clients Data (Taiwan), separated on fairness induction for the two different sensitive attributes in the dataset 'sex' and 'age. The protected groups are respectively female and old age group.

Default of Credit Card Clients Data			
model	**Fairness score**		**Average**
sensitive attribute	***Sex***	***Age***	
Logistic Regression	0	0	0
Pre-processing models			
Suppression	-0.02	0	-0.01
Massaging	0	0.01	0.01
Preferential Sampling	0.01	0.28	0.15
Post-processing models			
Equalized Odds	-0.09	0.22	0.07
Calibrated Equalized Odds	1.32	1.89	1.61
Reject Option Classification	0.01	0.28	0.15

Overall, we see the Preferential Sampling and Calibrated Equal Odds improve fairness most (when allowing for positive discrimination). Additionally, performance of these methods is similar to the baseline model. The reason for the occurrence of positive discrimination might be explained when having a closer look at the dataset. Positive discrimination occurs more in the behavioral dataset. The baseline model yields fairer outcomes for the Default of Credit Clients Data than the German Credit Dataset. This might have to do with the fact that the first dataset contains behavioral data but might also be partially explained by the size difference between the different groups. Figure 2 shows that the group size of the protected group 'female' in the sensitive attribute "sex" is larger than the 'male' group. Discrimination is more present for the sensitive attribute 'age', where we see that the protected group is significantly smaller. In figure 2 we see the group sizes for the German Credit Dataset can be seen in figure 3. There the protected group is consistently smaller.

Figure 2

Group sizes of the sensitive attributes in the Default of Credit Clients Data. 1 represents the protected group and 0 the unprotected group. Left we see the sensitive attribute 'sex' and right 'age'.

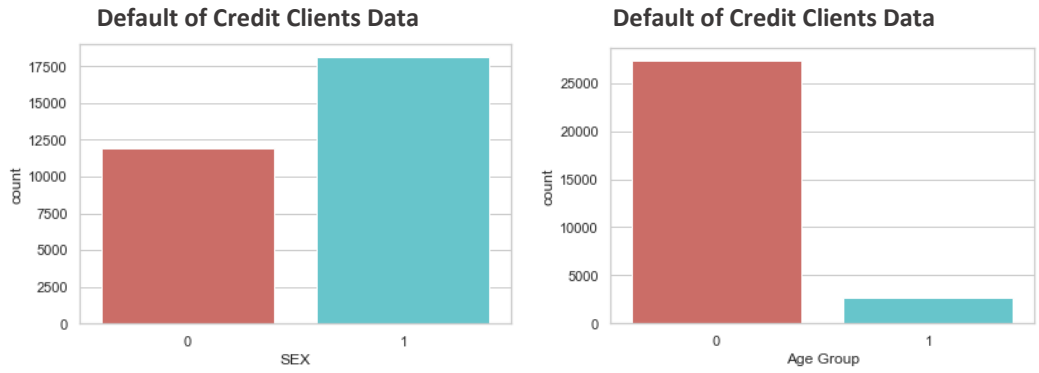
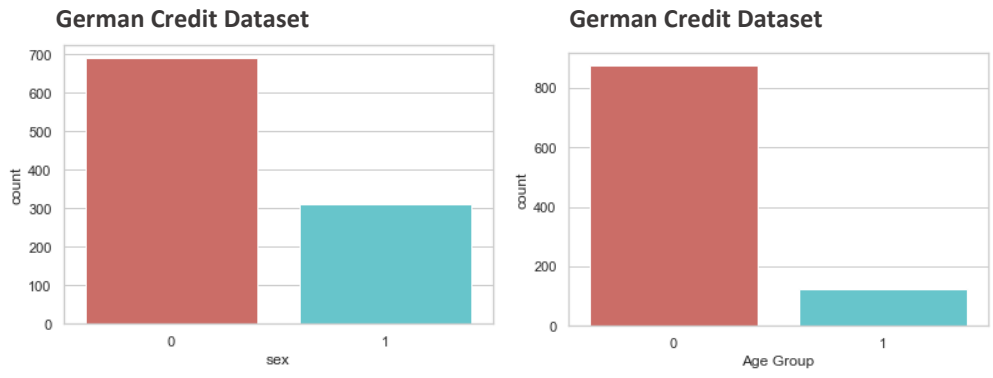


Figure 3

Group sizes of the sensitive attributes in the German Credit Data. 1 represents the protected group and 0 the unprotected group. Left we see the sensitive attribute 'sex' and right 'age'.



5 Discussion

In this study fairness inducing models were explored in the applied setting of credit scoring. Looking both at behavioral data and the more 'traditional data'. We looked at both pre-processing and post-processing fairness inducing models. The Massaging and Preferential Sampling methods by Kamiran et Calders (2011) were used, in addition to the more widely known suppression technique. For our post-processing models, we used Equalized Odds (Hard et al., 2016), Calibrated Equalized Odds (Pleiss et al., 2017) and Reject Option Classification (Kamishima et al., 2012). We compared these methods to a baseline model, looking at both performance and fairness metrics. This research was conducted to answer the following research questions.

- **Do sensitive attributes like gender and age influence decision making in the application of credit scoring and how can this be measured?**
- **What fairness inducing method(s) cause(s) sensitive attributes such as age and gender to be of the smallest influence on the decision making?**

From our results we can conclude that sensitive attributes do influence decision making in credit scoring. The fairness metrics for our logistic regression (the baseline) show that between protected and unprotected group there is inequality in decision making (Appendix B, table 3 and 4). We do see that unfairness is less present in the behavioral dataset than in the dataset not containing this data. This might in part have to do with the fact that, in particular, the protected group 'female' is overrepresented in the dataset. We also acknowledge the measures of equality between groups and see that measures of fairness do present a tradeoff. Higher fairness values measured by statistical parity difference and disparate impact ratio, lead to lower fairness measured by average odds difference and equal opportunity difference. Furthermore, measuring fairness, particularly in age was challenging as, most measures assume a binary inequality problem. The sensitive attribute 'age' was therefore reduced to two groups, young and old, while providing more age groups would have offered more nuance. Further research on multi group fairness is needed to understand the consequences of this better.

Methods combatting unfairness in this application do show their effectiveness. Overall, Preferential Sampling and Calibrated Equalized Odds have shown to be the most effective. Though, it can still not definitively be said that one model is best. Different models might have better use with different datasets. As we saw, performance can be influenced by group sizes within the sensitive attributes, the sensitive attributes itself and all attributes in the dataset (e.g., behavioral data vs non behavioral data). Additionally, the objectives of a financial institution can influence the decision for fairness models they choose to implement. If risk management and cost reduction are the highest priority (having low false positive rates), they will likely value precision more than other performance metrics. Our study suggests that when using behavioral data, Calibrated Equal Odds is more beneficial in this scenario, while for non-behavioral data, Preferential Sampling is the better choice. If a loan provider prioritizes generating more income over risk management, looking at recall performance in addition to fairness, can be more important. Using Reject Option Classification for non-behavioral data could then be a good choice. The choice of fairness metric also influences the decision making. Depending on what definition of non-discrimination we chose to uphold, we rate different fairness methods differently. Additionally, our fairness inducing models are often limited to addressing fairness based on one sensitive attribute only, while individuals can belong to

different protected and unprotected groups simultaneously. The choice of what fairness method to use, thus also greatly depends on what discrimination is chosen to address. If Discrimination based on gender is selected, then discrimination based on age or perhaps race cannot be treated anymore. Further research on the topic, could result in more accessible models that address multiple forms of discrimination at a time.

6 Conclusion

This research explored fairness inducing models in the applied setting of credit scoring. Differently from previous studies, this paper looked both behavioral data more 'traditional data'. Where models have been introduced in the past, this research compares different methods, both pre-processing and post-processing, that can more easily be implemented in practice. The results showed that in credit scoring, group attribution does influence the probability of being assigned a positive classification and thus deemed "creditworthy". Our findings suggest that the use behavioral data, reduces the presence of discrimination, while this cannot be said definitively.

While we recognize that even within the application of credit-scoring, different scenarios require different models, we find that the pre-processing method of Preferential Sampling and the post-processing method of Calibrated Equal Odds overall perform best. In coming to this conclusion, we also consider the different definitions of fairness and the tradeoffs that exist between them. We suggest two directions of future research. Firstly, we suggest research on multi group fairness within one sensitive attribute such as age or race. Secondly, research on treating discrimination based on multiple sensitive attributes at a time could lead to more fairness.

7 References

Aggarwal, N. (2021). The Norms of Algorithmic Credit Scoring. *The Cambridge Law Journal*, *80*(1), 42–73. https://doi.org/10.1017/s0008197321000015

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, *54*(6), 627–635. https://doi.org/10.1057/palgrave.jors.2601545

Bellamy, R. K. E, Dey, K., Hind, M., Hoffman, S. C, Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R, & Zhang, Yunfeng. (2018). *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*.

Bequé, A., & Lessmann, S. (2017). Extreme learning machines for credit scoring: An empirical evaluation. *Expert Systems with Applications*, *86*, 42–53. https://doi.org/10.1016/j.eswa.2017.05.050

Calmon, F. P., Wei, D., Ramamurthy , K. N., & Varshney , K. R. (2018). Conference on Neural Information Processing Systems (NIPS 2017). In *Optimized Pre-Processing for Discrimination Prevention*. Red Hook, New York; Curran Associates, Inc.

Chiappa, S. (2019). Path-Specific Counterfactual Fairness. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 7801–7808. https://doi.org/10.1609/aaai.v33i01.33017801

Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments, 153–163. https://doi.org/10.1089/big.2016.0047

Corbett-Davies, S., & Goel, S. (2018). International Conference on Machine Learning (ICML 2018). In *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*. Stockholm

Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, *91*, 106263. https://doi.org/10.1016/j.asoc.2020.106263

Hardt, M., Price, E., & Srebro, N. (2016). 30th Annual Conference on Neural Information Processing Systems . In *Equality of Opportunity in Supervised Learning*. La Jolla, CA; NIPS Foundation.

Hurley, M., & Adebayp, J. (2016). Credit Scoring in th Era of Big Data. *Yale Journal of Law and Technology*, *18*, 148–216.

Jones, G. P., Hickey, J. M., Dhanjal, C., Stoddart, L. C., & Vasileiou, V. (2017). ACM Conference (Conference'17). In *Metrics and methods for a systematic comparison of fairness-aware machine learning algorithms*. Wshington DC, Washington.

Kamiran, F., & Calders, T. (2011). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8

Kleinberg, Jon, Mullainathan, Sendhil, & Raghavan, Manish. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores.

Kusner, Matt J, Loftus, Joshua R, Russell, Chris, & Silva, Ricardo. (2017). *Counterfactual Fairness.*

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Lohia, P. K., Natesan Ramamurthy, K., Bhide, M., Saha, D., Varshney, K. R., & Puri, R. (2019). 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). *Bias Mitigation Post-processing for Individual and Group Fairness.* https://doi.org/10.1109/icassp.2019.8682620

Mehrabi, Ninareh, Morstatter, Fred, Saxena, Nripsuta, Lerman, Kristina, & Galstyan, Aram. (2019). *A Survey on Bias and Fairness in Machine Learning*.

Oneto, L., & Chiappa, S. (2020). INNS Big Data and Deep Learning Conference (INNSBDDL2019). In L. Oneto, Navarin Nicolò, A. Sperduti, & D. Anguita (Eds.), *Recent Trends in Learning From Data* (Vol. 896, pp. 155–172). Cham, Switzerland; Springer Nature.

Prabhakar, T., & Weber, S. (2020). Financial Inclusion as a Fairness Criterion in Credit Risk Assessment. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3579695

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., & Zafar, M. B. (2018). A Unified Approach to Quantifying Algorithmic Unfairness. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. https://doi.org/10.1145/3219819.3220046

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, *16*(2), 149–172. https://doi.org/10.1016/s0169-2070(00)00034-0

Thomas, L. C., Crook, J. N., & Edelman, D. B. (2002). Credit scoring and its applications. Siam.

Trivedi, S. K. (2020). A study on credit scoring modeling with different feature selection and machine learning approaches. *Technology in Society*, *63*, 101413. https://doi.org/10.1016/j.techsoc.2020.101413

Valdivia, A., Sánchez-Monedero, J., & Casillas, J. (2021). How fair can we go in machine learning ? Assessing the boundaries of accuracy and fairness. *International Journal of Intelligent Systems*, *36*(4), 1619–1643. https://doi.org/10.1002/int.22354

Verma, S., & Rubin, J. (2018). 2018 ACM/IEEE International Workshop on Software Fairness. In *Fairness Definitions Explained*. Gothenburg.

8 Appendices and Supplementary Materials

Appendix A

Description of the German Credit Data

```
1. Title: German Credit data

2. Source Information

Professor Dr. Hans Hofmann
Institut f"ur Statistik und "Okonometrie
Universit"at Hamburg
FB Wirtschaftswissenschaften
Von-Melle-Park 5
2000 Hamburg 13

3. Number of Instances:  1000

Two datasets are provided.  the original dataset, in the form
provided
by Prof. Hofmann, contains categorical/symbolic attributes and
is in the file "german.data".

For algorithms that need numerical attributes, Strathclyde
University
produced the file "german.data-numeric".  This file has been
edited
and several indicator variables added to make it suitable for
algorithms which cannot cope with categorical variables.
Several
attributes that are ordered categorical (such as attribute 17)
have
been coded as integer.    This was the form used by StatLog.


6. Number of Attributes german: 20 (7 numerical, 13 categorical)
   Number of Attributes german.numer: 24 (24 numerical)


7.  Attribute description for german

Attribute 1:  (qualitative)
              Status of existing checking account
                A11 :      ... <    0 DM
              A12 : 0 <= ... <  200 DM
              A13 :      ... >= 200 DM /
                  salary assignments for at least 1 year
                A14 : no checking account

Attribute 2:  (numerical)
            Duration in month
```

```
Attribute 3:  (qualitative)
            Credit history
            A30 : no credits taken/
                all credits paid back duly
              A31 : all credits at this bank paid back duly
            A32 : existing credits paid back duly till now
              A33 : delay in paying off in the past
            A34 : critical account/
                other credits existing (not at this bank)

Attribute 4:  (qualitative)
            Purpose
            A40 : car (new)
            A41 : car (used)
            A42 : furniture/equipment
            A43 : radio/television
            A44 : domestic appliances
            A45 : repairs
            A46 : education
            A47 : (vacation - does not exist?)
            A48 : retraining
            A49 : business
            A410 : others

Attribute 5:  (numerical)
            Credit amount

Attibute 6:  (qualitative)
            Savings account/bonds
            A61 :          ... <  100 DM
            A62 :   100 <= ... <  500 DM
            A63 :   500 <= ... < 1000 DM
            A64 :          .. >= 1000 DM
              A65 :   unknown/ no savings account

Attribute 7:  (qualitative)
            Present employment since
            A71 : unemployed
            A72 :       ... < 1 year
            A73 : 1  <= ... < 4 years
            A74 : 4  <= ... < 7 years
            A75 :       .. >= 7 years

Attribute 8:  (numerical)
            Installment rate in percentage of disposable income

Attribute 9:  (qualitative)
            Personal status and sex
            A91 : male   : divorced/separated
            A92 : female : divorced/separated/married
              A93 : male   : single
            A94 : male   : married/widowed
            A95 : female : single
```

```
Attribute 10: (qualitative)
            Other debtors / guarantors
            A101 : none
            A102 : co-applicant
            A103 : guarantor

Attribute 11: (numerical)
            Present residence since

Attribute 12: (qualitative)
            Property
            A121 : real estate
            A122 : if not A121 : building society savings
agreement/
                           life insurance
              A123 : if not A121/A122 : car or other, not in
attribute 6
            A124 : unknown / no property

Attribute 13: (numerical)
            Age in years

Attribute 14: (qualitative)
            Other installment plans
            A141 : bank
            A142 : stores
            A143 : none

Attribute 15: (qualitative)
            Housing
            A151 : rent
            A152 : own
            A153 : for free

Attribute 16: (numerical)
               Number of existing credits at this bank

Attribute 17: (qualitative)
            Job
            A171 : unemployed/ unskilled  - non-resident
            A172 : unskilled - resident
            A173 : skilled employee / official
            A174 : management/ self-employed/
                 highly qualified employee/ officer

Attribute 18: (numerical)
            Number of people being liable to provide maintenance
for

Attribute 19: (qualitative)
            Telephone
            A191 : none
            A192 : yes, registered under the customers name
```

```
Attribute 20: (qualitative)
            foreign worker
            A201 : yes
            A202 : no

8.  Cost Matrix

This dataset requires use of a cost matrix (see below)


      1        2
----------------------------
  1   0        1
-----------------------
  2   5        0

(1 = Good,  2 = Bad)

the rows represent the actual classification and the columns
the predicted classification.

It is worse to class a customer as good when they are bad (5),
than it is to class a customer as bad when they are good (1).
```

Retrieved from: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data) on 06-04-2021

Description of the Default of Credit Card Clients Data

```
Source:
Name: I-Cheng Yeh
email addresses:
(1) icyeh '@' chu.edu.tw (2) 140910 '@' mail.tku.edu.tw
institutions: (1) Department of Information Management, Chung Hua
University, Taiwan. (2) Department of Civil Engineering, Tamkang
University, Taiwan.
other contact information: 886-2-26215656 ext. 3181


Data Set Information:
This research aimed at the case of customersâ€™ default payments
in Taiwan and compares the predictive accuracy of probability of
default among six data mining methods. From the perspective of
risk management, the result of predictive accuracy of the
estimated probability of default will be more valuable than the
binary result of classification - credible or not credible
clients. Because the real probability of default is unknown, this
study presented the novel â€œSorting Smoothing Methodâ€ to
estimate the real probability of default. With the real
probability of default as the response variable (Y), and the
predictive probability of default as the independent variable
```

(X), the simple linear regression result (Y = A + BX) shows that the forecasting model produced by artificial neural network has the highest coefficient of determination; its regression intercept (A) is close to zero, and regression coefficient (B) to one. Therefore, among the six data mining techniques, artificial neural network is the only one that can accurately estimate the real probability of default.

Attribute Information:
This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable.

This study reviewed the literature and used the following 23 variables as explanatory variables:

X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
X2: Gender (1 = male; 2 = female).
X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
X4: Marital status (1 = married; 2 = single; 3 = others).
X5: Age (year).
X6 - X11: History of past payment.

We tracked the past monthly payment records (from April to September, 2005) as follows:

X6 = the repayment status in September, 2005;

X7 = the repayment status in August, 2005; . . .;
X11 = the repayment status in April, 2005.
The measurement scale for the repayment status is:
-1 = pay duly;
1 = payment delay for one month;
2 = payment delay for two months; . . .;
8 = payment delay for eight months;
9 = payment delay for nine months and above.
X12-X17: Amount of bill statement (NT dollar).

X12 = amount of bill statement in September, 2005;
X13 = amount of bill statement in August, 2005; . . .;
X17 = amount of bill statement in April, 2005.

X18-X23: Amount of previous payment (NT dollar).
X18 = amount paid in September, 2005;
X19 = amount paid in August, 2005; . . .;
X23 = amount paid in April, 2005.

Retrieved from: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients on 06-04-2021

Appendix B

Detailed results from research

Table 1
Detailed performance results of the different fairness models as well as the baseline model on the German Credit Dataset by sensitive attribute.

German Credit Dataset										
model	**accuracy**		**balanced accuracy**		**precision**		**recall**		**specficity**	
sensitive attribute	***Sex***	***Age***	***Sex***	***Age***	***Sex***	***Age***	***Sex***	***Age***	***Sex***	***Age***
Logistic Regression	0.75	0.75	0.73	0.73	0.85	0.85	0.77	0.77	0.69	0.69
Pre-processing models										
Suppression	0.75	0.75	0.73	0.74	0.85	0.86	0.77	0.77	0.69	0.71
Massaging	0.73	0.68	0.7	0.64	0.83	0.79	0.77	0.73	0.63	0.55
Preferential Sampling	0.78	0.76	0.75	0.74	0.86	0.86	0.81	0.79	0.7	0.7
Post-processing models										
Equalized Odds	0.74	0.76	0.7	0.73	0.82	0.85	0.8	0.79	0.6	0.67
Calibrated Equalized Odds	0.78	0.77	0.72	0.71	0.83	0.82	0.86	0.85	0.58	0.57
Reject Option Classification	0.71	0.76	0.74	0.65	0.9	0.77	0.67	0.93	0.82	0.37

Table 2
Detailed performance results of the different fairness models as well as the baseline model on the German Credit Dataset by sensitive attribute.

Default of Credit Card Clients Data										
model	**accuracy**		**balanced accuracy**		**precision**		**recall**		**specficity**	
sensitive attribute	***Sex***	***Age***	***Sex***	***Age***	***Sex***	***Age***	***Sex***	***Age***	***Sex***	***Age***
Logistic Regression	0.69	0.69	0.58	0.58	0.81	0.81	0.79	0.79	0.36	0.36
Pre-processing models										
Suppression	0.69	0.7	0.57	0.58	0.81	0.82	0.79	0.78	0.35	0.38
Massaging	0.69	0.69	0.57	0.56	0.81	0.81	0.79	0.79	0.35	0.34
Preferential Sampling	0.69	0.69	0.58	0.58	0.81	0.81	0.72	0.79	0.36	0.36
Post-processing models										
Equalized Odds	0.6	0.71	0.61	0.61	0.84	0.83	0.6	0.79	0.61	0.43
Calibrated Equalized Odds	0.69	0.69	0.65	0.63	0.86	0.85	0.72	0.73	0.58	0.54
Reject Option Classification	0.72	0.74	0.65	0.65	0.85	0.85	0.77	0.81	0.53	0.5

Table 3
Detailed fairness results of the different fairness models as well as the baseline model on the German Credit Dataset by sensitive attribute. The measures are from left to right statistical parity difference, disparate impact, average odds difference, and equal opportunity difference.

German Dataset								
model	**Stat Parity diff**		**Disparate Impact**		**Average Odds Diff**		**Equal Opp Diff**	
sensitive attribute	***Sex***	***Age***	***Sex***	***Age***	***Sex***	***Age***	***Sex***	***Age***
Logistic Regression	-0.04	0.14	0.94	1.22	-0.01	0.06	0.04	0.21
Pre-processing models								
Suppression	-0.12	0.03	0.83	1.05	-0.07	-0.05	-0.08	0.11
Massaging	-0.23	-0.69	0.68	0.06	-0.2	-0.64	-0.13	-0.78
Preferential Sampling	0.75	0.17	0.87	1.28	-0.06	-0.04	0.05	0.25
Post-processing models								
Equalized Odds	-0.07	0.06	0.9	1.09	-0.01	-0.04	-0.06	0.06
Calibrated Equalized Odds	-0.06	0.31	0.92	1.45	-0.03	0.4	0.06	0.17
Reject Option Classification	-0.01	0.04	0.98	1.05	0.03	-0.08	0.11	0.08

Table 4
Detailed fairness results of the different fairness models as well as the baseline model on the German Credit Dataset by sensitive attribute. The measures are from left to right statistical parity difference, disparate impact, average odds difference, and equal opportunity difference.

Default of Credit Card Clients Data								
model	**Stat Parity diff**		**Disparate Impact**		**Average Odds Diff**		**Equal Opp Diff**	
sensitive attribute	***Sex***	***Age***	***Sex***	***Age***	***Sex***	***Age***	***Sex***	***Age***
Logistic Regression	0.02	-0.05	1.03	0.93	0	-0.03	0.01	-0.06
Pre-processing models								
Suppression	0.01	-0.05	1.02	0.93	0	-0.03	0.01	-0.06
Massaging	0.02	-0.05	1.03	0.94	0	-0.03	0.01	-0.06
Preferential Sampling	0.02	0.02	1.03	1.03	0.01	0.01	0.01	0.01
Post-processing models								
Equalized Odds	-0.01	0.01	0.98	1.01	-0.01	-0.01	0.01	0
Calibrated Equalized Odds	0.28	0.36	1.58	1.57	0.25	0.45	0.27	0.3
Reject Option Classification	0.01	0.02	1.01	1.03	0.03	0.01	0.02	0.01