Mitigating Digital Discrimination in the German Credit Dataset by Fair Pre-Processing to predict Fair Credit Risks

An application of fair pre-processing and supervised machine learning techniques

Aleyna Kartal

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

STUDENT NUMBER

2077892

THESIS COMMITEE

Dr Peter Hendrix
Dr. Raquel Alhama

LOCATION

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

DATE

December 2, 2022

WORD COUNT

8743

ACKNOWLEDGEMENT

I want to take a moment to deeply and honestly thank Dr. Peter Hendrix for his guidance, support and encouraging words during my thesis period. I can say that I am grateful to the time Dr. Peter Hendrix dedicated in guiding my study. In addition, I would like to thank my grandparents, parents, little brother and fiancée for always supporting me and believing in me. My biggest role model in life, my grandmother, sadly passed away. I feel honored to fulfill her dream and one of the promises I made to her of finishing my masters. Erdinç Saçan, my old teacher and now colleague, has inspired me with his book 'Inclusive Artificial Intelligence' to do my thesis about such an important ethical topic. I want to thank Erdinç for always inspiring me. Lastly, I would like to thank Lina (my cat) for her unconditional support and distracting moments so I take healthy breaks from my thesis.

Table of contents

Abstract

Due to the rise of automated decision-making in the banking world, it is essential that decisions made by algorithms are digital discrimination free. Personal data attributes like gender must be dealt with to avoid this type of bias in the training data. The drive to make fairness-aware ML algorithms has increased and there are a few tools that provide a framework for detecting and mitigating digital discrimination.

This study focuses on illustrating the effect of pre-processing techniques on the fairness and model performance in a financial banking dataset for various classifiers: Support Vector Machine, Random Forest, eXtreme Gradient Boosting, Logistic Regression and Neural Network. The tool used within the study is the AI Fairness 360 toolkit. Reweighing and Disparate Impact Remover are the pre-processing techniques applied to mitigate bias. The fairness metrics are Disparate Impact, Statical Parity Difference, Average Odds Difference and Equal Opportunity Difference.

Based on the steps above, the following research question will be answered: "In the German Credit dataset, to what extent can digital discrimination in assigning loan approval be mitigated by fair pre-processing techniques while preserving model performance?" Previous research focused on different classifiers with different baselines and, in some cases, with other tools. Furthermore, this study added the protected attribute of foreign workers as it is a personal data attribute.

This study concluded that there is a trade-off between fairness and model performance. However, this differs per classifier and pre-processing technique. In some cases, pre-processing techniques can yield higher performances than the baseline model or even decrease fairness.

Keywords: *Digital discrimination, pre-processing, fairness metrics, AIF360, ethics, discrimination-aware ML, AI Fairness 360 toolkit*

1. Introduction

This chapter contains the project definition, elaboration on the scientific and societal relevance followed by the research strategy in which the research questions are elaborated.

1.1 Project definition

Decades ago, futurists anticipated machines relieving professionals and managers of making decisions (Achieng, Majuto, Aseka & Astiaya, 2019). Nowadays, the promise of automated decision-making systems is used worldwide by companies and organizations to make efficient data-driven decisions. With the rise of automated decision-making, it is essential to make fair machine learning (ML) predictions (Biswas & Rajan, 2021). The automated decision-making is booming in systems for banking and hiring, etc.

Concerns are raised about automated decisions and them inheriting historical bias and discrimination from training data. Incidents about ML models exhibiting discrimination among people based on personal data like age, gender, race, etc. have been reported in the recent years (Friedler et al., 2019). Companies can be exposed to different risks such as reputational and operational as a result of unintentional bias.

The drive to make fairness-aware ML algorithms has increased (Friedler et al., 2019). Digital discrimination refers to the unethical, unfair or just different treatment of users based on personal data by an automated decision-making algorithm. ML algorithms can produce discriminatory outcomes based on personal data attributes. Those attributes are also called protected characteristics/attributes. Decisions cannot be based on those attributes.

There are various debiasing techniques or bias mitigation techniques: pre-processing, in-processing and post-processing. Where pre-processing aims on the training data, in-processing is focused on the algorithm itself and post-processing on correcting the predictions by making them more fair. Furthermore, an open-source toolkit, AI fairness 360, has been developed by a research community and contains various techniques to unmask and mitigate bias in ML models (Zhang et al., 2021).

This study focuses on detecting and mitigating digital discrimination in a financial banking dataset with pre-processing techniques and AI fairness 360 metrics. The effect of the pre-processing techniques will be illustrated along with the results of the AI fairness metrics.

1.2 Scientific and Societal Relevance

Along with the rise of automated decision-making, there is an increase in concerns regarding digital discrimination. The automated decision-making model should produce fair non-biased outcomes which are non-discriminatory. Digital discrimination can lead to serious problems due to unethical/unfair treatment of people based on personal data.

The addressed problem has a societal as well as a scientific relevance. From the societal point of view, on the one hand it is important as people should be treated equally and fair, on the other hand the machine should give the impression of trust to the people in being treated fair. This trust can be reached by showing that data scientists, in fact, can have control on mitigating bias in automated decision-making algorithms.

From the scientific point of view, the outcome of the thesis contributes to scientific research regarding ethical considerations in automated decision-making. The study shows how discrimination bias can be detected and mitigated. Data scientists get

an insight on the impact of (not) mitigating the bias on the decision outcome, making them strive to mitigate it in their work field as well. Lastly, it will contribute to scientific research as the effect of the fair pre-processing will be applied on different supervised ML models to see the effect on their model performance.

1.3 Research strategy

This study contains one main research question and two research subquestions.

Research question: ***"In the German Credit dataset, to what extent can digital discrimination in assigning loan approval be mitigated by fair pre-processing techniques while preserving model performance?"***

SQ1: "How do the pre-processing techniques for solving imbalances/non-normal distributions in the German Credit Dataset affect the outcomes on the AI fairness 360 metrics?"

The processing techniques that are used to solve imbalances/non-normal distributions in the German Credit (GC) Dataset are: 1) reweighing as there are two classes from which the bad-risk contains significantly less observations than the good-risk (Chen, 2018) and, 2) Disparate Impact Remover (DIR) for improving group fairness by editing feature values.

The effect of pre-processing is illustrated with a comparison of the AI fairness 360 metrics: Disparate Impact (DI) and Average Odds Difference (AOD) on the original- and transformed datasets after training the classifiers. Measuring the fairness scores with all four metrics before and after pre-processing is necessary to have an insight in the bias mitigation (Bellamy et al., 2018) (Hufthammer et al., 2020). The fairness scores of the Statically Parity Difference (SPD) and Equal Opportunity Difference (EOD) are to be found in appendix C (see elaboration in paragraph 3.1).

SQ2: ***"How does the performance of different ML models compare before and after fair pre-processing?"***

This study applies five different ML models: Support Vector Machine (SVM), Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Logistic Regression (LR) and Neural Network (NN) (see elaboration in paragraph 3.3).

After training the classifiers and measuring the fairness on the original- and transformed datasets, the performance for each model is calculated with 2 error metrics: F-measure (for imbalanced data) and AUC (suitable after rebalancing the data) (Salunkhe & Mali, 2018). The elaboration of use for these metrics can be found in paragraph 3.4.

As last step, the model performance for each classifier and pre-processed dataset is analyzed. This to compare the effect of the pre-processing techniques on the model performance.

2. Literature Review

Previous research with regards to credit risk models, pre-processing, fairness metrics, and ML models are elaborated within this chapter. The contribution of this study is also presented.

2.1 Introduction credit risks and fairness

In the world of finance, ML algorithms are applied to make value-based automated decisions regarding financial data. Tasks such as fraud detection or loan approval have been automated and standardized. This to boost the efficiency of services and to cut down costs. Due to those benefits, financial companies are keen on adopting ML and/or AI. The algorithms made by data scientists contain data of the companies' customers, increasing the demand to be aware of and to mitigate data bias. Occasionally, that bias is the root origin of the accidental biases. (Zhang & Zhou, 2019)

As bias and imbalances often occur in financial data, the chance of misclassifying targets heightens (Zhang & Zhou, 2019). This can also be seen in the study of Zhang & Zhou (2019) in which a credit card dataset was analyzed. Not only did it contain biases in age and race but also imbalances for gender as the classes where unequally represented with female as the unprivileged group. Within the study, synthesis cases have been created with the Synthetic Minority Oversampling Technique (SMOTE) to oversample the minority class due to the large imbalance in the grouped age variable. SMOTE transitioned the imbalanced data to balanced. After fitting the balanced data with LightGBM, it showed a significant improvement on the performance metrics. In addition, the false negatives rate dropped from 0.62 to 0.23. As the imbalance in `Gender` is not as strong as in the `Age` variable, the bias mitigation technique reweighing was applied to modify the weights in the training samples. After the reweighing, there has been a minor difference in the resulting outcome. The study concluded that by rising consideration in the fairness of algorithms, financial companies can achieve further confidence from customers in being fairly assigned a credit risk.

The outcome of a biased algorithm could lead to harmful decisions for the bank, as well as for the customer. A biased algorithm can make decisions that will be a loss of business, or financial loss to the bank while for the customers it is important to be treated fair as it affects the loan (Pappadà & Pauli, 2022). The false recognition of good applicants can for example lead to a huge economic loss for the bank.

Although making the ML model fair is morally desirable, often a trade-off between fairness vs. model performance is present (Zelaya, 2019). Adjusting the performance can make it less desirable from the financial point of view as bad risks can be chosen as good risks due to the fairness. Despite the probability of this trade-off, there are also sources which show that there is just a little to none effect on model performance (Bellamy et al., 2018) (Stevens et al., 2020). In short, it is possible that after applying pre-processing techniques, the model performance drops.

2.2 Pre-Processing

As presented in paragraph 2.1, fairness is an essential concern in today's automated decision-making applications. The AI fairness 360 toolkit contains the AIF360 library which carries fairness metrics and bias mitigations algorithms for models and datasets. Some of the bias mitigation algorithms contain models for pre-processing.

Using protected attributes in making decisions lead to direct discrimination, also called disparate treatment. Nowadays, DI is more pervasive. In this case the protected attributes were not made use of, but the dependent correlated attributes caused significantly diverse results for every group (Calmon et al., 2017).

To rebalance training data, pre-processing needs to be applied. Either by undersampling (removing instances from recognized majority class), or oversampling (introducing instances from the new minority class). In general, oversampling is preferred as with undersampling relevant instances may be removed (Cordón et al., 2018).

Bellamy et al. (2018) set up an experiment on three datasets, one of them being the GC dataset. Within the pre-processing stage, the AIF360 library was used for bias mitigation algorithms: reweighing and DIR. The mean effect on fairness of the bias mitigation algorithms on the classifiers LR and RF are illustrated vs. the balanced model performance. For both reweighing and DIR, the fairness increases on both classifiers while the model performance stays approximately equal. The DIR has a higher effect on the RF classifier than the LR classifier with regard to improving fairness.

The GC dataset contains the protected attribute gender which is unbalanced (Zelaya, 2019). Minority groups can suffer from unfair decisions when a classifier gets trained over such an imbalanced dataset. The attribute must be balanced first. The study of Zelaya (2019) balanced the attributes label (creditworthiness) and gender. The second attribute has a skewed distribution and is, therefore unrepresentative of the real-world population. The experiment carried out undersampling and oversampling with SMOTE on both attributes and trained a RF on the training sets. The model performance for the label variable had a small increase in model performance from 0.76 to 0.77. With the techniques under- and oversampling, the model performance of the label and gender variable decreased.

The study Salunkhe & Mali (2018), regarding a credit scoring system, has shown a significant improvement of performance in terms of F-measure and Area under Curve (AUC) in classifying good/bad risks after pre-processing. For the pre-processing, the imbalance between classes was reduced by combining dimensionality reduction and random re-sampling. The classification algorithms LR and SVM have shown a small increase in the AUC and the F-measure on the re-sampled GC dataset. In contrary to LR and SVM, Naïve Bayes and Random subspace have shown a decrease in as well as the AUC as the F-measure.

2.3 AI Fairness 360 Metrics

In some studies four fairness metrics were used on protected attributes in a binary classification task (Bellamy et al., 2018) (Hufthammer et al., 2020). The evaluation was executed with: DI, SPD, AOD and EOD fairness metrics were used to evaluate the fairness (Zelaya, 2019). While the optimal value for fairness of the DI is 1, the other metrics strive to a value of 0.

The study of Bellamy et al. (2018) set up an experiment on the GC dataset. The fairness metrics SPD and DI were applied on the pre-processing technique reweighing. As a result of reweighing, the SPD for 'Sex' decreased from -0.09 to -0.025. On the DI, the fairness has improved from 0.88 to 0.96.

An elaboration on the metrics can be found in the paper Zhang & Zhou (2019). The probability of assigning a positive outcome to an unprivileged group as-compared-with a privileged group can be measured with DI. The disparity between the probability of an unprivileged- and privileged group receiving a positive outcome is measured by SPD. In case of an equal true- and false positive rate for the unprivileged and privileged group, AOD is satisfied. In case of a negative value, the unprivileged group has a disadvantage. Lastly, EOD illustrated the disparity among a privileged- and unprivileged group regarding true positive rates.

2.4 ML Models on Financial datasets

For several decades, the usage of automated decision-making within the financial industry is carried out to support risk management. LR on the GC dataset has been applied and the results show that it is in fact possible to correct fairness to some extent, without losing a strong model performance in the prediction of the model. In short, the aim is to exclude the protected attributes while preserving a reasonable performance. For future work the authors want to investigate the effect of the fairness on other ML models such as SVM, RF or NN. (Szepannek & Lübke, 2021)

Banks have personal data of clients which are used for automated decision-making systems like the credit analysis system (Chen, 2018). The paper contains an approach of tree-based ML models: Decision Tree (DT), AdaBoost, Bagging and RF. The tree-based models have been reweighted due to the imbalanced data, which increased their model performance. Of all four models, the bagging model outperformed the other models with the model performance of predicting at least 70% of as well as the good as the bad risks. Followed by the bagging model is RF, which outperforms the adaboost model and DT with a model performance of at least 63%.

Another study that applied RF, is the study of Bellamy et al. (2018). Within the study the ML Models LR and NN were also applied after pre-, in- or post-processing. The results show that even though the pre-processing has slightly lowered the model performance of RF, it still has the highest model performance of at least 75%. After RF, LR is the model with the highest model performance (~ 73%) even though the pre-processing techniques also slightly lowered the model performance of this model. On the NN, only post-processing has taken place and has the lowest model performance (66,5%).

A study has analyzed various scientific works and found that several techniques have been applied to credit scoring applications: DT, Artificial NN (ANN), SVM, LR, Discriminant Analysis (DA), among others (Rangel-Díaz-de-la-Vega et al., 2020). While LR with a model performance of 89,3% got outperformed by DT's (93,2%), both had better results than the DA (87.5%) (Leung, Cheong & Cheong, 2008). The same study has shown accuracies for the GC dataset: k-NN (78,0%), SAIS (75,4%), Naïve Bayes (74,7%). As a performance measure, the ROC has been used in the study. The baseline SAIS has a R^2 of 0.227 and is suggested to be an appropriate classifier as the ROC point approximates the point (0,1). One of the studies applied ANN to a credit approval task and argued that ANN has a predictive model performance of 87% and indicates strong reliability (Ilgun, Mekic & Mekic, 2014).

For generating fair credit scores, two different approaches were applied: DT and the non-parametric learning method K-Nearest Neighbor (KNN). Both approaches have been evaluated with the AUC and F-measure metrics. (Pandey & Bandhu, 2022). DT shows a F-measure of 0.413 and an AUC of 0.7. KNN showed a F-measure of 0.464 paired with an AUC of 0.75.

2.5 Contribution

The essential need for fair ML models has been the main driver for this study to detect and mitigate bias. In various current studies, pre-processing techniques like SMOTE have been applied. Within this study, the focus is on the pre-processing techniques that are available in the AIF360 library. This to illustrate the power of the tools within the library to detect and mitigate bias.

Firstly, various studies used different models with different baselines. Within this study the focus is on the five selected models. Prior research did not have a study with those 5 classifiers in one study and is therefore missing in prior research. The models without pre-processing are set as baseline. The classifiers get trained on the

original- and pre-processed datasets. The classifiers will be measured with the fairness metrics as well as the F-measure and AUC (Nguyen, Kiyoaki & Huynh, 2021) (Pandey & Bandhu, 2022).

Secondly, `Foreign worker` is added as a protected attribute as it is a personal data attribute. This study illustrates that the protected attributes should be considered carefully while preparing the data for the model.

Lastly, missing in prior research is the elaboration on the chosen fairness metrics. Each metric has its own strength and weakness. Even though it would be valuable to reach fairness on all fairness metrics, it is important to investigate which metric is best suiting and representing the domain problem at hand (see elaboration in 3.1).

In short, this study illustrates the effect of fair pre-processing on the different supervised ML models and their performance. This is crucial as it contributes to the knowledge of (future) data scientists regarding fair pre-processing in the financial field.

3. Methodology

This section elaborates the following methods that are applied during the study. Firstly, the fairness metrics are elaborated, followed by the pre-processing techniques and the classifiers. The pre-processing techniques and fairness metrics are selected as the scope of this study are the tools from the AIF360 library.

3.1 AI Fairness Metrics

During the study, the elaborated fairness metrics below will be used to assess fairness.

Disparate Impact

DI is the discrepancy in probability of favorable outcomes among privileged- and unprivileged group. In other words, an automated decision making process has DI in case the group are disproportionately advantaged or disadvantaged based on their protected attribute values (Barocas & Selbst, 2016). One of the advantages of DI is that it is used as a tool to protect individuals from harmful and discriminatory outcomes in case of predictive analytics (MacCarthy, 2017). This makes it harder to ignore discriminating practices in for example employment. The DI is widely used in predictive analytics for credit banking as the DI detects and avoids the unintentional discrimination for unprivileged groups (Kallus, Mao & Zhou, 2021).

Crucial is to keep in mind that even when there are no protected attributes, there might still be DI present. Also, the tool can have implications with regard to policy (Petersen, 2005).

While the ideal value for DI is 1, the values between 0.8 and 1.25 represent some level of fairness. The privileged group has an advantage if the value is lower than 1. In case the value is higher than 1, the unprivileged group has an advantage. Given the values C the binary class to be predicted, and X the protected attribute, the DI can be measured as follows:

$$\frac{Pr(C = 1 \,|X = 0)}{Pr(C = 1 \mid X = 1)}$$

Statistical Parity Difference

The discrepancy in probability of the favorable results of the unprivileged- to the privileged group is called SPD. The objective is to reach Statistical Parity (SP), also known as Demographic Parity, in which each group of applicants have an equal probability of being selected (Corbett-Davies et al., 2017). For instance, hiring males and females at equal rates.

Even though SP represents group fairness in which the privileged- and unprivileged groups have equal results, it however does not treat the individuals fair (Dwork et al., 2011). One of the disadvantages is the SPD being prone to historical biases and reinforcing the bias. SPD can be used as a metric in case there is awareness of the historical bias affecting the data quality and being able to set policies in order to support unprivileged groups. Lastly, it can be used when there is a necessity for change in the current situation to improve results for the unprivileged groups.

The ideal value of fairness for SPD is 0, and fairness is accomplished if the value is between -0.1 and 0.1. Values below the 0 represent unfairness against the unprivileged group. The SPD can be measured as follows:

$$P(C = 1 \,|X = 0) - P(C = 1 \mid X = 1)$$

Average Odds Difference

The average discrepancy of False Positive Rate (FPR) and True Positive Rate (TPR) among the unprivileged- and privileged group is called AOD. The best practice to use AOD is when it is crucial to predict the positive outcome accordingly. For example, being able to detect a not-creditworthy applicant. Furthermore, AOD is used when we focus on minimizing False Positives (FP's) that form a risk (Hardt et al., 2016). For example, when we want to reduce loan approval for applicants that would not be able to pay the loan back. Therefore, the EOD is suited for credit modeling.

0 is the ideal fairness value for AOD, and fairness is accomplished as long as the value is between -0.1 and 0.1. Values below 0 represent an advantage for the privileged group, while values higher than 0 represent an advantage for the unprivileged group. The AOD can be measured as follows:

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

Equal Opportunity Difference

EOD is the discrepancy of TPR's among the unprivileged- and privileged groups. The objective is to reach the equal opportunity of individuals or the equal proportion amount of individuals per group to reach the desired results. This metric however does not fully deal with FP's which can result again in DI (Radovanović et al., 2020). The best practice to use EOD is when it is crucial to predict the positive outcome accordingly (Hardt et al., 2016).

As the emphasis is on the TPR's, the EOD should not be used when for example the FP's are a risk to both applicant and the bank.

The ideal fairness value for EOD is 0, and fairness is reached in case the value is between -0.1 and 0.1. Values below 0 represent and advantage for the privileged group and vice versa. The EOD is calculated as follows:

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

Due to the fact that DI is used in credit models and the AOD is suited the best for credit models as it focuses on minimizing risks for the bank and the applicant, it is decided to focus on the results of those 2 metrics. Within this study, the AOD is seen as the primary metric with regard to fairness as it is best suited for credit models. The results of the SPD and EOD are presented in appendix C.

3.2 Pre-processing

In order to increase fairness, the fairness pre-processing techniques reweighing and DIR are applied to the training data to detect and mitigate bias.

Reweighing

Reweighing allows one to add weight to all (group, label) combinations instead of changing the target labels in the training set (Kamiran & Calders, 2012). In other words, each tuple is assigned a weight and neither the label nor feature values are modified.

Reweighing has an independency constraint in which protected attributed when it comes to making predictions. The independence constraint refers to the absence of the dependency among protected and output features. By using weights, the dependency can be removed.

Given the values P_{exp} the expected probability of representing a specific protected group, P_{act} the actual probability of representing a specific group and $d \wedge +$ the probability of being a specific class. The weights are calculated by using the

following formula (Calders, Kamiran, & Pechenizkiy, 2009):

$$W(D = d \mid x(Class) = +) = \frac{P_{\text{exp}}(d \wedge +)}{P_{\text{act}}(d \wedge +)}$$

Disparate Impact Remover

The DIR is focused on improving group fairness by modifying the feature values and keeping the rank order in the groups. Eliminating the ability to tell the difference among groups is the objective of DIR (Feldman et al., 2015).

Given the values C the binary class to be predicted, X the protected attribute and Y the remaining attributes: $D = (X, Y, C)$. Y is used to predict X. A repaired version of D is noted as $\overline{D}$ in which the features are changed to reach fairness. To avoid DI of $\overline{D}$ on X, the following criteria needs to be reached (Feldman et al., 2015): $\overline{D} = (X, \overline{Y}, C)$. The technique creates $\bar{y}$, $\bar{y} = F_A^{-1}(F_x(y))$, in which A defines a distribution and $y \in Y_x$: modifications take place to all Y to predict X, while also keeping the distribution of Y to be able to predict Y.

3.3 Supervised Binary Classification Algorithms

As a result of the literature study in chapter 2, this study uses five supervised binary classification algorithms (see table 1). The selected algorithms are used on the same or other relevant datasets in the financial domain. The baseline models are the models that are trained on the not pre-processed datasets.

A comparison of these commonly used algorithms in the context of fairness based research is very limited, while it is important as not all algorithms necessarily respond in the same way to the pre-processing steps taken.

There are several reasons for why the models above have been chosen. The main reason is that the models have been used in different papers, but not have been applied in one study.

Another reason for implementing LR, SVM, RF and NN is the future work section of the study Szepannek & Lübke (2021). The authors mentioned that after investigating the effect of LR, the effect of fairness on the other models is yet to be carried out. In the study of Chen (2018) different tree-based models were implemented. It showed that the bagging model outperformed the RF model. As AdaBoost was already implemented, XGBoost is applied to contribute to research.

The paper of Bellamy et al. (2018) researched the effect of pre-processing on the RF and LR. However, the NN is only post-processed.

Table 1 *Selected classifiers and sources*

Classifier	Source
SVM	Szepannek & Lübke, 2021; Rangel-Díaz-de-la-Vega et al., 2020
RF	Bellamy et al., 2018; Szepannek & Lübke, 2021; Chen, 2018
XGBoost	Instead of using the AdaBoost in Chen (2018), the extension XGBoost on the RF will be used to expand research.
LR	Bellamy et al., 2018; Szepannek & Lübke, 2021; Rangel-Díaz-de-la-Vega et al., 2008
NN	Szepannek & Lübke, 2021; Rangel-Díaz-de-la-Vega et al., 2020

3.4 Evaluation Metrics

After training the models, the unseen test set is used to make the predictions on unseen data and to evaluate the model.

F-measure

The F-measure is a performance evaluation metric and is widely used for imbalanced class datasets (Salunkhe & Mali, 2018). While the AUC uses the thresholds between TPR and FPR, the F1 score represents the average value of the precision and recall. A higher score represents a better performing model with regards to prediction. The ideal value is 1 (good prediction), while the worst case value is 0 (no good predictions). The F-measure is chosen as the dataset in this study is imbalanced and this metric performs well on imbalanced datasets avoiding misleading results (Pandey & Bandhu, 2022).

AUC

The AUC is used in various studies for classification tasks (Salunkhe & Mali, 2018) (Pandey & Bandhu, 2022). After pre-processing and re-sampling the data, the AUC is used on a credit risk dataset. It measures to what degree the model can separate the classes. The ideal value is, while the worst case value is 0. The AUC metric is used in this study and seen as the primary metric with regard to model performance due to the fact it has been used in several studies in the domain of credit risk modeling.

4. Experimental Setup

4.1 Data

The study used the GC dataset which contains 1000 loan applications of a bank. Furthermore, the original dataset contains 20 independent variables with demographic and socio-economic data about the applicant and 1 dependent variable that classifies the applicant as creditworthy or not-creditworthy (see table 2). Even though there is no specific data available about whether the loan ended up being paid, there is a variable that shows how the applicants in general dealt with previous loans if any were requested. The modifications made on the dataset are elaborated in paragraph 4.3.1.

Derived from the applicant's profile, the loan managers decide whether to accept or decline the loan application. To maximize the profit and minimize the credit risk, the bank needs an automated decision rule regarding approving or disapproving the loan (STAT 897D, sd).

Within the project, one of the objectives is to mitigate digital discrimination. To mitigate digital discrimination, the protected attributes need to be detected. The GC dataset contains three protected attributes: `Age`, `Gender` and `Foreign worker`. In addition to the protected attribute gender, research has shown that ML algorithms increase gender segregation. An example: male holders have more elite business credit cards than female holders (Zhang & Zhou, 2019).

Table 2 *Attributes description GC Dataset*

Variable	Description
Creditability	Creditworthiness of the applicant
Account_balance	The applicants' current balance on the account
Duration_credit	The duration of the credit in months
Payment_status_ previous_credit	Some status possibilities of the applicant regarding the previous credits
Purpose_credit	Possible reasons for the applicants' credit request
Credit_amount	The applicants' total credit in Deutsche Mark
Value_savings_stocks	The value of the applicants' savings and/or stocks
Length_current_employment	The length of the applicants' employment by the current employer
Instalment_percent	The instalment rate of the applicant's income
Gender	The applicants' gender
Guarantors	Other debtors, if any, with the applicant
Duration_current_address	The duration of the applicants' stay in the current household
Most_valuable_available_asset	The applicants' most valuable assets
Age	The applicants' age expressed in years
Concurrent_credits	Other ongoing credits of the applicant
Apartment_type	The apartment type of the applicant
Number_credits_at_ this_bank	Total count of the applicants' ongoing credit at the current bank
Occupation	The applicant's job
Number_dependents	The amount of people that are entitled to maintenance
Telephone	Whether the phone number is registered under the applicants' name
Foreign_worker	Whether the applicant is an immigrant worker

4.2 Software

In order to conduct the study, the programming languages R and Python are used. The bias mitigation algorithms and fairness metrics in the AI Fairness 360 toolkit are applied. Those algorithms and metrics are available in the library AIF360. In addition, the package caret is used for the different models. Other packages used for the classifiers are kernlab, randomForest, Xgboost, plyr, caTools, arm and RSNNS. The pre-processing algorithms are applied in Python, while the classifiers are trained in R.

4.3 Data Science Pipeline

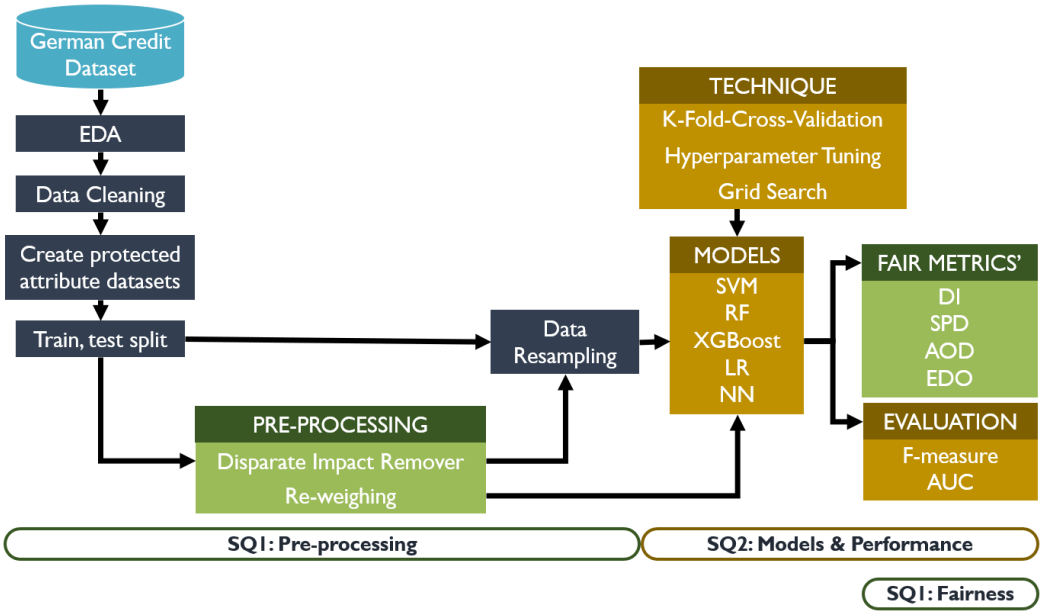
This study contains various stages which are visualized in figure 1. A detailed sketch of the process is provided in appendix A. After loading the GC dataset, the data is explored (EDA) to gain insight in the data distribution and to detect where data cleaning is needed. Even though there is no missing data, some modifications are applied (see paragraph 4.3.1).

The study gives an insight to the effect of different classifiers on the pre-processed datasets, therefore improving the performance is not the primary goal and no feature selection will take place.

Subquestion 1 is focused on the two pre-processing techniques and on the fairness measures. Crucial here is to keep as well as the original datasets as the pre-processed datasets on which pre-processing has not been applied. This way the effect of pre-processing on the fairness scores can be recorded and compared after training the classifiers. The pre-processing techniques that are applied are reweighing and DIR.

Training the classifiers and the model evaluation of the study takes place in subquestion 2. The base line classifier models are the SVM, RF, XGBoost, LR, NN without the pre-processing techniques. Therefore, the models will be trained with- and without the pre-processing techniques to illustrate the effect on the fairness metrics and the model performance. The model performance is evaluated with the F-measure and AUC.

Figure 1 *Process of this study*



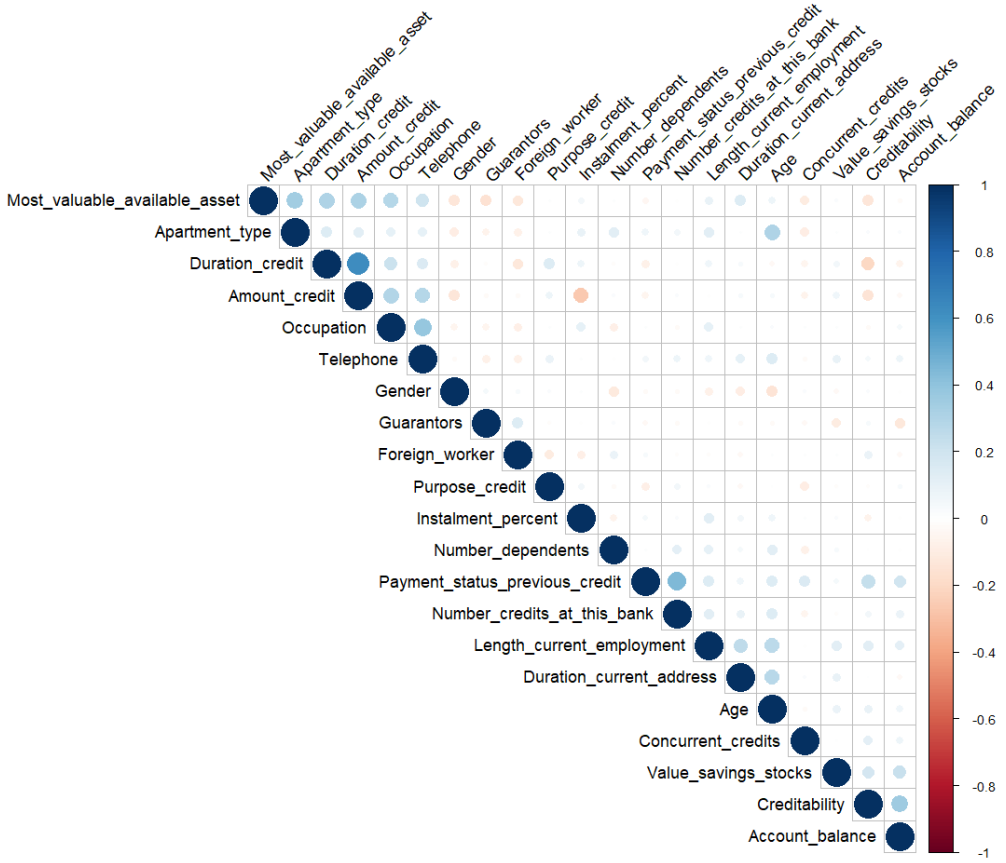
4.3.1 Exploratory Data Analysis and Data Cleaning

Exploratory Data Analysis (EDA) has taken place to get some insight into the data, the needed modifications and the data distribution. The dataset did not contain missing values, therefore handling missing values was not needed.

While exploring the dataset, several modifications have taken place on the original dataset. The attribute `Sex….Marital.Status` mainly contained the marital status for males. Decided is to only focus on the gender of the applicant in this column. Furthermore, to prepare the dataset for the classifiers, copies of the datasets are created and the target variable is converted to the data type factor. For the NN however, the datasets are kept as it only accepts a numeric input.

The correlation of the different attributes are shown in figure 2. The target variable has some correlation with `Payment_status_previous_credit`, `Duration_credit`, etc.

Figure 2 *Correlation attributes GC Dataset*

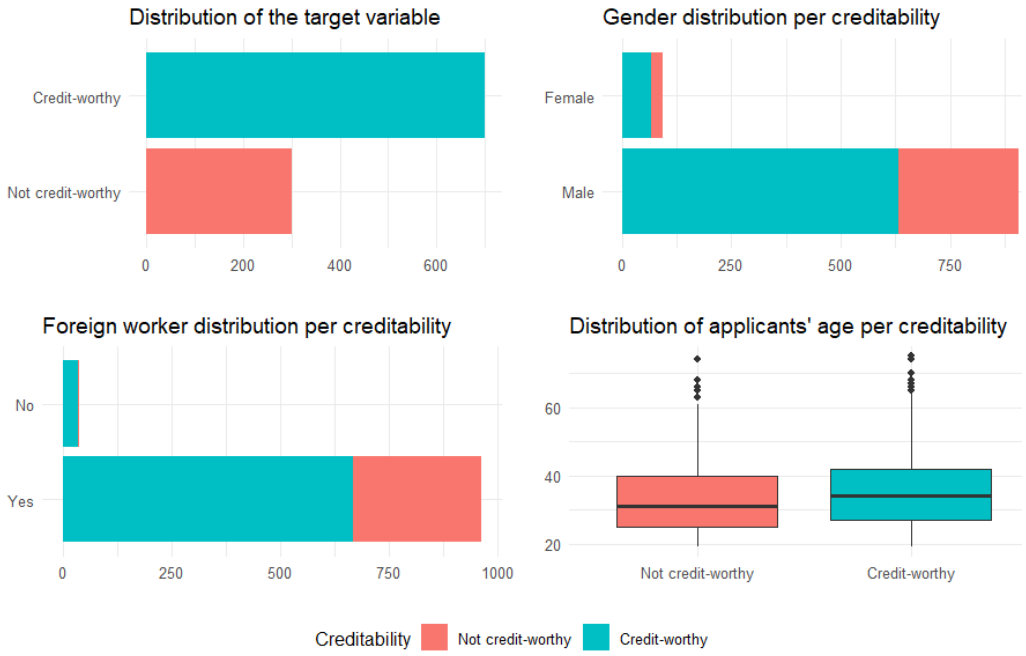


The EDA makes it possible to identify the favorable label and the privileged groups. While the favorable label represents the advantageous outcome, the privileged group represents the group that has a systematic advantage in the corresponding protected attribute.

The target variable `Creditability` is imbalanced (see figure 3). 70% of the applicants are credit-worthy and 30% are not. In terms of favorable labels, the preferred label is credit-worthy. The protected attributes have also been analyzed. As can be seen in figure 3 there is a strong imbalance in the `Gender` variable: there are 908 males and just 92 females, meaning the privileged group is males. Also remarkable is the distribution of the `Foreign worker` variable (see figure 3): 963 applicants are foreign workers while 37 are not foreign, making foreign workers the privileged group. Interesting in the variables gender and foreign worker is that even though there are

less females and less not foreign workers, in most cases their credit requests would be approved. According to the data (see figure 3), there is a right-skewed distribution in the age variable. The boxplot shows that the median age of being creditworthy is slightly higher than not-creditworthy, making the applicants with the age of >25 the privileged group.

Figure 3 *Distribution of the target, gender, foreign worker and age variables*



4.3.2 Feature Scaling

The independent features `Duration Credit`, `Amount credit`, `Number dependents` and many more are in very wide and different ranges: (4-72 months), (250-18424 Deutsche Mark) and (1-2). To scale the range of the independent features to the same range, feature scaling is applied. This is crucial to avoid the dilemma of features, that have a relative superior magnitude, dominating the performance of a trained classifier. For feature scaling the min-max normalization is used to scale the features in the range [0,1].

4.3.3 Protected Attributes Datasets and Train-Test Split

In order to measure the fairness per protected attribute, for each protected attribute a separate dataset is created (see figure 4 and table 3). Dataset 'GC Foreign worker' for example only contains foreign worker 1 of the protected attributes: `Foreign worker`. The other protected attributes are excluded.

Each dataset is converted to datasets that are compatible with the AIF360 library. The library contains different dataset classes. A `StructuredDataset` is created for each dataset in figure 7. As input for the structured dataset, the privileged and unprivileged groups are set beforehand. For `Gender` the privileged group is male and the unprivileged group is female. The privileged group for `Age` is >25 and the privileged group for `Foreign worker` is Yes (0).

The three structured datasets are randomly split into the train- and test set. Due to the fact that a K-Fold-Cross-Validation (k=10) takes place before training the final classifiers, there is no separate validation set (see paragraph 4.3.5).

Figure 4 *Dataset per protected attribute*

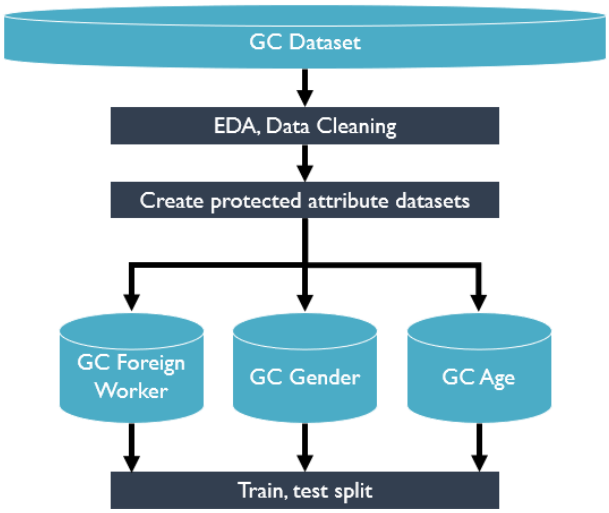


Table 3 *Dataset per protected attribute*

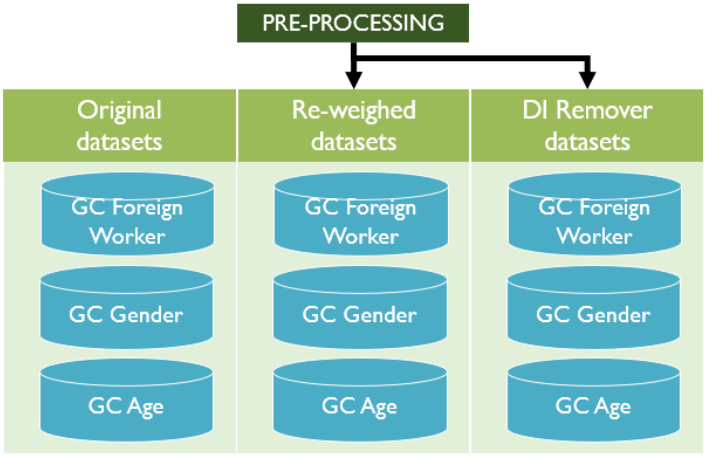
	GC Foreign Worker	GC Gender	GC Age
Domain	Credit	Credit	Credit
Number of attributes	19	19	19
Number of sensitive attributes	1	1	1
Names of the sensitive attributes	Foreign worker	Gender	Age
Favorable sensitive value	Yes	Male	>25
Desired decision label	Creditworthy	Creditworthy	Creditworthy
Undesired decision label	Not-creditworthy	Not-creditworthy	Not-creditworthy
Binary decision label	Yes	Yes	Yes

4.3.4 Pre-processing

As illustrated in figure 5, each dataset is pre-processed with the reweighing and DIR techniques.

Each pre-processing technique is applied on each protected attribute dataset, meaning that for each pre-processing technique there are three separate datasets (see figure 5). As the baseline of this study are the classifiers which are trained on not pre-processing datasets, the original datasets are also kept.

Figure 5 *The original dataset and the pre-processed datasets*



4.3.5 Resampling and K-Fold-Cross-Validation

After the pre-processing stage, upsamling is applied to the original- and pre-processed DIR datasets, because the imbalance in the outcome variable is still present. The reweighed datasets are not resampled as with reweighing weights are already added to each observation without changing the target variables. With an equal number of observations per class, the weights can still differ per class after resampling.

The K-Fold-Cross-Validation (k=10) is applied on each dataset, except for the datasets for NN (k=5) to reduce the high computational load. After shuffling the dataset on random, the training set is split into k groups. This means that each split the groups for testing and training are selected. The model is fitted on the training data and evaluated on the test data. After the K-Fold-Cross-Validation has taken place, the final model performance evaluation is with the unseen test data that is not used in the K-Fold-Cross-Validation.

4.3.6 Supervised Binary Classification Algorithms

Along with training the classifiers, cross-validation took place and the hyperparameters were tuned. Due to the fact that the dataset is not very big, grid search is used to compute the optimal values of each parameter by going through each possible combination. In table 4, the classifiers, their methods and the corresponding hyperparameters that are tuned are shown. In appendix B an overview of the hyperparameter settings and the selected optimal settings are shown.

Table 4 *Hyperparameters for selected classifiers*

Classifier	Method	Hyperparameters
SVM	svmLinear	C
	svmPoly	degree, scale, C
	svmRadial	sigma, C
RF	rf	mtry
XGBoost	xgbTree	nrounds, max_depth, eta, gamma, colsample_bytree, min_child_weight, subsample
LR	glm	-
	logitBoost	nIter
	bayesglm	-
NN	mlpML	layer1, layer2, layer3

5. Results

The fairness scores and model performance of the pre-processing techniques on the protected attributes and various classifiers are presented in this section. For each classifier and protected attribute, as well as the fairness scores as the model performance of the baseline model is compared to the reweighed and DIR models. Lastly, an error analysis on the model with the highest fairness takes place.

5.1 Fairness versus performance

Below, for each classifier the fairness scores of the DI and AOD measures are aligned with the model performance (AUC and F-measure). While the optimal value for the DI is 1, the AOD strives to a value of 0.

Radial SVM

Table 5 shows that in most cases, the pre-processing techniques on the radial SVM models do not improve DI scores. They however do improve the AOD scores, especially reweighing. Noteworthy is that the baseline model for DI has the best fairness for `Foreign worker` and `Age`, while the AOD score is best when the model is reweighed. The reweighed model is still not within the range of fairness.

As expected, the baseline model of the radial SVM has the highest model performance compared to the models on which bias mitigation has taken place. The model performance for the model dropped after pre-processing, but only increased for `Foreign worker` while the AOD stayed constant and the DI score got worse.

Table 5 *The fairness scores of the metrics DI and AOD compared to the model performance with AUC and F-measure as metric for the radial SVM models before and after pre-processing.*

Radial SVM		DI	AOD	AUC	F-measure
Baseline	Foreign worker	**1.012**	0.226	0.88	0.819
	Gender	0.631	-0.078	**0.878**	**0.856**
	Age	**0.775**	-0.386	**0.874**	**0.832**
Reweighed	Foreign worker	1.114	**0.134**	0.872	0.841
	Gender	1.293	0.127	0.871	0.838
	Age	0.761	**-0.218**	0.872	0.832
DIR	Foreign worker	1.524	0.226	**0.883**	**0.848**
	Gender	**0.77**	**0.07**	0.825	0.827
	Age	0.704	-0.443	0.863	0.826

Random Forest

In table 6 the fairness scores show that in general the pre-processing techniques on the RF models do have a positive effect on the fairness as they get closer to the range of (optimal) fairness. Looking at the AOD scores, the best approach for the RF model is by using DIR as a pre-processing technique as it presents results within the fairness threshold.

The baseline model for the RF yields the best performance (at least 92%) as the pre-processing techniques resulted in a drop in model performance. Remarkable is that the best performing model in terms of fairness, is performing the worst looking at the AUC and F-measure. With the baseline model, the fairness scores are not even close to fairness, showing a trade-off between fairness and performance. Even though

a drop in model performance took place with the DIR, we still can get a model that is fair and has considerable predictive power (at least 85%) for all three attributes. Taking into account the SPD and EOD as well (see appendix C, table 2), we see that the pre-processed with DIR dataset shows fair scores for `Gender`.

Table 6 *The fairness scores of the metrics DI and AOD compared to the model performance with AUC and F-measure as metric for the Random Forest models before and after pre-processing.*

Random Forest		DI	AOD	AUC	F-measure
Baseline	Foreign worker	0.773	-0.044	**0.952**	**0.919**
	Gender	0.688	**0,0**	**0.923**	**0.905**
	Age	0.628	-0.212	**0.932**	**0.879**
Reweighed	Foreign worker	0.817	-0.109	0.868	0.865
	Gender	1.159	0.112	0.870	0.835
	Age	**0.924**	-0.028	0.882	0.843
DIR	Foreign worker	**0.857**	**-0.017**	0.865	0.838
	Gender	**0.937**	0.021	0.865	0.841
	Age	0.674	**0.005**	0.852	0.839

XGBoost

The XGBoost models present scattered scores for the DI (see table 7), but show that for the AOD the baseline model ensures the closest scores to fairness followed by the model that is pre-processed with the DIR technique.

The pre-processed DIR model shows a higher performance in terms of AUC (at least 93%). Even though the DIR yields the highest performance, the baseline model also yields a model with predictive power. Interesting to see is that for SPD the results are scattered as well (see appendix C, table 3), while for EOD all the scores are within a range of fairness for the baseline model.

Table 7 *The fairness scores of the metrics DI and AOD compared to the model performance with AUC and F-measure as metric for the XGBoost models before and after pre-processing.*

XGBoost		DI	AOD	AUC	F-measure
Baseline	Foreign worker	0.779	-0.028	0.93	0.87
	Gender	0.618	**-0.078**	0.91	**0.89**
	Age	**0.864**	**-0.133**	**0.936**	**0.886**
Reweighed	Foreign worker	0.848	-0.076	0.882	0.859
	Gender	**1.181**	0.117	0.889	0.844
	Age	0.633	-0.253	0.884	0.835
DIR	Foreign worker	**0.857**	**0.026**	**0.935**	**0.874**
	Gender	0.649	-0.193	**0.928**	0.883
	Age	0.806	-0.113	0.925	0.866

Logistic Regression

In contrast to the tables above, the scores in table 8 show that both pre-processing techniques on the LR models improved the fairness scores for DI and AOD even though they are not always in the range of fairness. The worst performing model is the baseline model in which none of the protected attributes reaches the fairness threshold.

The highest model performance is yield with the reweighed model (at least 86%), while the baseline model shows the worst performance (at least 73%). The fairness scores in are not too far from the fairness threshold, meaning that with the reweighed model robust predictions can be made that are also very close to fairness. Surprisingly for SPD and EOD (see appendix C, table 4), the highest fair scores are exactly aligned with DI and AOD. The scores are within or close to the range of fairness.

Table 8 *The fairness scores of the metrics DI and AOD compared to the model performance with AUC and F-measure as metric for the Logistic Regression models before and after pre-processing.*

Logistic Regression		DI	AOD	AUC	F-measure
Baseline	Foreign worker	2.045	0.16	0.747	0.706
	Gender	0.664	-0.428	0.731	0.728
	Age	0.488	-0.281	0.781	0.735
Reweighed	Foreign worker	1.413	0.156	**0.867**	**0.846**
	Gender	1.223	0.117	**0.859**	**0.821**
	Age	**0.809**	**-0.152**	**0.865**	**0.833**
DIR	Foreign worker	**1.331**	**0.155**	0.754	0.719
	Gender	**0.846**	**-0.039**	0.783	0.74
	Age	0.439	-0.306	0.776	0.738

Neural Network

Table 9 presents that in terms of the AOD metric, the DIR model is the best model for `Foreign worker` and `Age`. For `Gender`, this model is 0.40 score away from the fairness threshold.

The best performing model for the NN is the reweighed model (at least 87%). Comparing this model to the DI fairness metric, it would result in unfair predictions for `Foreign worker` and `Gender`. The SPD is aligned with DI, and the EOD is namely doing well on the baseline model (see appendix C, table 5). The EOD is however not always in the range of fairness.

Table 9 *The fairness scores of the metrics DI and AOD compared to the model performance with AUC and F-measure as metric for the Neural Network (Multilayer Perceptron) models before and after pre-processing.*

Neural Network		DI	AOD	AUC	F-measure
Baseline	Foreign worker	1.492	0.194	0.854	**0.838**
	Gender	0.713	**-0.063**	0.852	**0.838**
	Age	**0.881**	-0.054	0.841	0.808
Reweighed	Foreign worker	1.420	0.183	**0.866**	0.827
	Gender	1.138	0.104	**0.869**	0.823
	Age	0.818	-0.139	**0.873**	**0.818**
DIR	Foreign worker	**1.408**	**0.069**	0.834	0.824
	Gender	**1.119**	0.14	0.868	0.828
	Age	0.693	**-0.048**	0.852	0.809

5.2 Summary of the findings

Concluding from the results, the pre-processing techniques do have an effect on decision making. The techniques effect the outcomes on the fairness scores positively

and negatively. Having a positive effect does not necessary mean that the score is within the fairness threshold, but is at least closer to it. In most cases, DIR present the highest fairness scores followed by the reweighed models. For the XGBoost, the baseline model had higher fairness scores. However, we also saw that the results of fairness for each attribute can (slightly) differ per metric: DI and AOD. But also for SPD and EOD.

Likewise, the performance of the different models can be effected in both directions. While it is expected to yield the best performance without bias mitigation, this study showed the contrary for the models XGBoost, LR and NN which perform better when the model is reweighed or pre-processed with the DIR. Therefore, the effect of the pre-processing techniques differ per model on the attributes and are therefore model specific.

The results show that in most cases there is a trade-off between the fairness score and model performance. However, we also saw that the drop in performance is not very high. The worst performing models are still close to the best performing models. This can be seen in for example the RF that is pre-processed with DIR. The model is within the fairness threshold and contains considerable predictive power (at least 85%). Like mentioned in the literature study, this is what we aim for: excluding the protected attributes while preserving a reasonable performance. The trade-off however differs per model.

Due to the fact that this study focusses on fairness, namely the EOD metric, the most fair model to use in terms of is the RF model that is pre-processed with the DIR technique (see table 7).

5.3 Error Pattern Analysis

As mentioned, the RF model that is pre-processed with the DIR technique yields the highest fairness score. Per protected attribute, the confusion matrices of this model is presented in tables 10, 11 and 12. The scores contain errors, which could be reduced by selecting the most optimal performing model instead of the most fair model. It can be argued that the EOD indeed focuses on reducing the FPR, by keeping equal chances across group to be classified TP or FN. Even though the chances per group cannot be seen, the FPR is relevantly less. The overall matrices below contain both privileged- and unprivileged groups.

Generally the model is better at predicting the TP's and performs less in predicting TN's for all three variables. For `Age` (see table 12) there is more improvement needed as it contains more FN's compared to `Foreign worker` and `Gender`. Also, even after upsampling the minority class, there is still an imbalance for this class.

Table 10 *Confusion matrix for the 'Foreign worker' variable of the Random Forest model that is pre-processed with the DIR technique*

		Reference	
		Creditworthy	Not-creditworthy
Predicted	Creditworthy	158	55
	Not-creditworthy	33	54

Table 11 *Confusion matrix for the 'Gender' variable of the Random Forest model that is pre-processed with the DIR technique*

		Reference	
		Creditworthy	Not-creditworthy
Predicted	Creditworthy	174	41
	Not-creditworthy	35	50

Table 12 *Confusion matrix for the 'Age' variable of the Random Forest model that is pre-processed with the DIR technique*

		Reference	
		Creditworthy	Not-creditworthy
Predicted	Creditworthy	109	44
	Not-creditworthy	102	45

6. Discussion

This study explored fairness aware modeling on various classifiers with pre-processing techniques in the banking domain. The goal of this study was to answer the following research question: *"In the German Credit dataset, to what extent can digital discrimination in assigning loan approval be mitigated by fair pre-processing techniques while preserving model performance?"*. During the study, the pre-processing techniques reweighing and DIR were applied on various classifiers: radial SVM, RF, XGBoost, LR and NN. For each model, the effect on the fairness scores as well as the model performance is measured. To detect the effect of the pre-processing techniques on the classifiers, the models without pre-processing were set as a baseline.

6.1 Summary and discussion of the results

The results of this study show a complex interplay between ML models, pre-processing techniques, fairness scores and model performance. The trade-off between the fairness scores and model performance differs per classifier and pre-processing technique on the protected attributes. The results showed that digital discrimination can be mitigated by fair pre-processing techniques but returns a (small) drop in model performance for the most fair model. The drop in model performance can be explained by the fact that the protected attributes contain predictive values that contribute to the decision making: approving the loan or not. Another reason for the drop in model performance with regard to the DIR was presented in the error analysis. We saw that even though the minority class was upsampled, it still did not represent the minority class well.

Another founding of the complex interplay is that in some cases, the preprocessing techniques (slightly) optimized the model performance for the XGBoost, LR and NN, while decreasing the fairness score. This could be explained by the fact that the protected attributes do not contain predictive values and are pre-processed in a way that boosts the performance. However, the attribute shows its importance as leaving it out, even after pre-processing, makes the model less fair. In short, the pre-processing techniques differ per model and are thus model specific.

In addition, even though the pre-processing techniques increased the fairness, the fairness scores did not always fall within the fairness threshold. This could be due to the fact that some models trained on certain datasets need more techniques like in- and post-processing techniques to reduce the digital discrimination even further. In short, one type of processing technique might not be enough to get within the fairness threshold. However, the results showed that per protected attribute the fairness scores (slightly) differ. Each metric (DI, SPD, AOD, EOD) measures other discrepancies between the privileged- and unprivileged groups. Crucial to realize is that in case not all metrics are within the fairness threshold, research is needed to detect which ones are the most important for the domain of the study. Within this study, the RF that was pre-processed with DIR was chosen as the most fair model based on the AOD. The literature study showed that AOD is the most suitable metric in the banking domain as it focuses on minimizing risks for the bank and the applicant.

6.2 Literature comparison

As described in the literature review, making a ML model is morally desirable but often creates a trade-off between fairness and model performance (Zelaya, 2019). This can also be seen in this study, especially for the RF model preprocessed with the DIR. Contrary to the study of Zelaya (2019), this study showed that the pre-processing improved the model performance in some cases. For example, the fairness results are very close to fairness and have a high model performance for the XGBoost preprocessed with DIR, reweighed LR and reweighed NN models. This means that model performance can be preserved while almost being within the fairness threshold. The reason can be that the protected attributes do not contain predictive values that contribute to the decision making and that the type of classifiers does not reduce the model performance combined with the pre-processing techniques. In some cases, there was a decrease in fairness after applying pre-processing techniques, performing worse which contradicts the prior research in Bellamy et al. (2018) and Pappadà & Pauli (2022), and my own expectations. This could be due to the minority group of the target label making misclassifications, even after upsampling as presented in the error pattern analysis.

6.4 Limitations and future directions

This study contains a few limitations. The NN provided a higher predictive power than the XGBoost, while the contrary is expected. This could be due to the settings of the hyperparameter tuning or the number of layers (see appendix B, table 1) used within the study. As there was a shortage of resources (time and especially computational power) more layers could not be added. By making the NN more complex, without wanting to overfit, the NN could present a higher predictive power. However, as the study is not about optimizing the model performance this is not a big problem in this case. In addition, the baseline model of the XGBoost is expected to have a higher model performance than the RF. This was not the case as there was a time constraint and other implementations with various hyperparameters for both RF and XGBoost could not be applied to see if the performance could be modified. Another limitation was the small dataset. However the same dataset is used in numerous previous scientific studies for credit modelling and is therefore valid to use in this study.

If this study were followed up upon, the minority group suffering from the unfair decisions (even after being upsampled) could be avoided by using dimensionality reduction and SMOTE (Zelaya, 2019). Including the dimensionality reduction and SMOTE, this study can be applied on bigger datasets (in the same and in other domains) to examine if the pre-processing techniques and classifiers have a similar effect as in this study. The focus then can be shifted to how model performance can be preserved while mitigating digital discrimination. For this not only the pre-processing techniques can be studied, but also the in-, and post-processing techniques to see the effect of this combination. The protected attributes can be handled differently as well. In this study we distinguished the protected attributes and the fairness scores were measured per protected attribute. Applicants can be part of various (un)protected groups at once (e.g., being a young unforeign female). In a future study, addressing fairness for individuals being part of the other protected attributes can be taken into account by combining them. Lastly, an addition is contacting a data science ethics department to

ask for their input in deciding what an acceptable model is when it comes to a high trade-off between fairness and model performance.

6.5 Scientific and societal impact

In addition to existing scientific knowledge, this study has shown that pre-processing techniques can impact the fairness scores while slightly dropping the model performance. Contrary to existing knowledge, in some cases, the pre-processing techniques improved the model performance but reduced the fairness scores. With regard to scientific impact this study has shown that the effect of fairness on the model performance can vary per classifier and pre-processing technique. Therefore, a data scientist should do research to the most fair and robust model. Furthermore, in the methods section, we saw that the focus of fairness differs per metric, meaning that data scientists should keep in mind to research the usage of the metrics and decide which metric is the most suitable for the problem at hand.

As a societal impact, society is as well as introduced to the possibility of making fair predictions because data scientists do have the toolkits (e.g., the AI Fairness 360 toolkit) to make the models more fair. This way society can see that even though there is some digital discrimination in some algorithms, the data scientist is working on more fairness-aware ML as the problem is being acknowledged. Not is it only acknowledged, as mentioned before, progress is made as for example the AI Fairness 360 toolkit is set up to help the society in being treated fair. Due to the rise of automated decision-making in the banking world, it is essential that decisions made by algorithms are digitally discrimination free and also give the society a feeling of being treated fair. The society can see that personal data attributes can be left out while making fair predictions. In short, applicants will not only be treated fair but they will also feel like being treated fairly in terms of credit approval as their personal data attributes do not play a role in creditworthiness. Therefore, this study can raise the trust of the society in the algorithms made to mitigate digital discrimination for important decisions like credit risks.

7. Conclusion

As automated decision-making is rising, especially in the banking world, concerns raise about automated decisions containing digital discrimination. It is essential to make fair ML predictions to ensure applicants that they are being treated fair, while also ensuring the bank that the decisions are made accordingly to maximize the profit and minimize the credit risk. For this, more research is needed to the expected trade-off between fairness and model performance. This study contributed to the scientific field by researching this important trade-off and revealing that digital discrimination in a banking dataset can be mitigated by fair pre-processing techniques, but returns a (small) drop in model performance for the most fair model therefore showing a trade-off. This is however was not the case for all models. Diverging from previous works, this study also showed a decrease in fairness after pre-processing. Therefore, this study shows that even though some patterns were found validating previous studies, there are also some diverging results. The study is limited to one dataset and future work implies more research and applying the pre-processing techniques to various datasets to ensure the results could be generalizable. Furthermore, a limitation was using only pre-processing techniques, in a future study the in- and post-processing techniques could be added to see the effect on the fairness and model performance.

Data Source/Code/Ethics Statement

The work on this thesis did not involve collecting data from human participants or animals. The German Credit Dataset has been acquired from the STAT897D website for Applied Data Mining and Statistical Learning (STAT 897D, sd). The original owner of the data used in this thesis retains ownership of the data during and after the completion of this thesis. Therefore I acknowledge that I do not have any legal claim on the data. The code used in this thesis is not publicly available.

References

Achieng, S., Majuto, C., Aseka, P., & Astiaya, E. (2019, 9 1). *Replacing Humans with Machines: Threats and Opportunities.* Retrieved from East African Journal of Business & Economics: https://journals.eanso.org/index.php/eajbe/article/view/80

Barocas, S., & Selbst, A. D. (2016, 6). *Big data's disparate impact.* Retrieved from HeinOnline: https://heinonline-org.tilburguniversity.idm.oclc.org/HOL/Page?handle=hein.journals/calr104&div=25&g_sent=1&casa_token=&collection=journals

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., . . . Zhang, Y. (2018, 10 3). *AI FAIRNESS 360: AN EXTENSIBLE TOOLKIT FOR DETECTING, UNDERSTANDING, AND MITIGATING UNWANTED ALGORITHMIC BIAS.* Retrieved from arXiv: https://arxiv.org/pdf/1810.01943

Biswas, S., & Rajan, H. (2021, 8 18). *Fair Preprocessing: Towards Understanding Compositional.* Retrieved from https://dl.acm.org/doi/abs/10.1145/3468264.3468536

Bonte, C., & Vercauteren, F. (2018, 10 11). *Privacy-preserving logistic regression training.* Retrieved from SpringerLink: https://link.springer.com/article/10.1186/s12920-018-0398-y

Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). *Building Classifiers with Independency Constraints.* Retrieved from IEEE Xplore: https://proceedings.mlr.press/v162/li22p/li22p.pdf

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). *Optimized Pre-Processing for Discrimination Prevention.* Retrieved from GitHub: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjA0qvUtKn6AhWIraQKHZy4DsEQFnoECAQQAQ&url=https%3A%2F%2Fkrvarshney.github.io%2Fpubs%2FCalmonWVRV_nips2017.pdf&usg=AOvVaw1K-11osHJAEprrrYavGwS8

Cervantes, J., Garcia-Lamont, F., Rondríguez-Mazahua, L., & Lopez, A. (2020, 5 8). *A comprehensive survey on support vector machine classification: Applications, challenges and trends.* Retrieved from ScienceDirect: https://www-sciencedirect-com.tilburguniversity.idm.oclc.org/science/article/pii/S0925231220307153?via%3Dihub

Chen, Z. (2018, 11 19). *The Application of Tree-based model to Unbalanced German Credit Data Analysis.* Retrieved from EDP Sciences: https://www.matec-conferences.org/articles/matecconf/abs/2018/91/matecconf_eitce2018_01005/matecconf_eitce2018_01005.html

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017, 6 10). *Algorithmic decision making and the cost of fairness.* Retrieved from arXiv: https://arxiv.org/pdf/1701.08230.pdf

Cordón, I., García, S., Fernández, A., & Herrera, F. (2018, 8 23). *Imbalance: Oversampling algorithms for imbalanced classification in R.* Retrieved from ScienceDirect: https://www-sciencedirect-com.tilburguniversity.idm.oclc.org/science/article/pii/S095070511830385X?casa_token=-6tfgTXk0tAAAAAA:keMVmlDgacWquyq1dAAhHw7iKhd6dqaeUeFa4oHRU_pDOgPNdD9ILW76w8-DA9Yp6ZVEbAM9mOM

Duan, J. (2019, 5). *Financial system modeling using deep neural networks (DNNs) for effective risk assessment and prediction.* Retrieved from ScienceDirect: https://www-sciencedirect-com.tilburguniversity.idm.oclc.org/science/article/pii/S0016003219301462

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011, 11 30). *Fairness Through Awareness.* Retrieved from arXiv: https://arxiv.org/pdf/1104.3913.pdf

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015, 8 10). *Certifying and Removing Disparate Impact.* Retrieved from ACM Digital Library: https://dl-acm-org.tilburguniversity.idm.oclc.org/doi/10.1145/2783258.2783311

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019, 1 29). *A comparative study of fairness-enhancing interventions in machine learning.* Retrieved from Association for Computating Machinery: https://dl.acm.org/doi/abs/10.1145/3287560.3287589

Fritz, M., & Berger, P. D. (2015). *Binary Logistic Regression.* Retrieved from ScienceDirect: https://www-sciencedirect-com.tilburguniversity.idm.oclc.org/topics/computer-science/binary-logistic-regression

Hardt, M., Price, E., & Srebro, N. (2016, 10 11). *Equality of Opportunity in Supervised Learning.* Retrieved from arXiv: https://arxiv.org/pdf/1610.02413.pdf

Hufthammer, K. T., Aasheim, T. H., Ånneland, S., Brynjulfsen, H., & Slavkovik, M. (2020, 11 23). *Bias mitigation with AIF360: A comparative study.* Retrieved from Bergen Open Research Archive: https://ojs.bibsys.no/index.php/NIK/article/view/833

Ilgun, E., Mekic, E., & Mekic, E. (2014). *Application of Ann in Australian Credit Card Approval.* Retrieved from European Researcher: http://www.erjournal.ru/journals_n/1393749061.pdf

Kallus, N., Mao, X., & Zhou, A. (2021, 4 2). *Assessing Algorithmic Fairness with Unobserved Protected ClassUsing Data Combination.* Retrieved from informs PubsOnLine: https://pubsonline-informs-org.tilburguniversity.idm.oclc.org/doi/epdf/10.1287/mnsc.2020.3850

Kamiran, F., & Calders, T. (2012). *Data preprocessing techniques for classification without discrimination.* Retrieved from SpringerLink: https://link-springer-com.tilburguniversity.idm.oclc.org/article/10.1007/s10115-011-0463-8

Kirasich, K., Smith, T., & Sadler, B. (2018). *Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets.* Retrieved from SMU Data Science Review.

Leung, K., Cheong, F., & Cheong, C. (2008, 11 19). *Consumer Credit Scoring using an Artificial Immune System Algorithm.* Retrieved from IEEE Xplore: https://ieeexplore-ieee-org.tilburguniversity.idm.oclc.org/abstract/document/4424908

MacCarthy, M. (2017). *Standards of fairness for disparate impact assessment of big data algorithms.* Retrieved from HEINONLINE: https://heinonline-org.tilburguniversity.idm.oclc.org/HOL/Page?handle=hein.journals/cumlr48&div=8&id=&page=&collection=journals

Nguyen, B.-H., Kiyoaki, S., & Huynh, V.-N. (2021, 2 25). *Topics in Financial Filings and Bankruptcy Prediction with Distributed Representations of Textual Data.* Retrieved from SpringerLink: https://link-springer-com.tilburguniversity.idm.oclc.org/chapter/10.1007/978-3-030-67670-4_19

Nobre, J., & Neves, R. F. (2019). *Combining Principal Component Analysis, Discrete Wavelet Transform and XGBoost to trade in the financial markets.* Retrieved from ScienceDirect: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0261629

Pandey, P., & Bandhu, K. C. (2022, 5 24). *A credit risk assessment on borrowers classification using optimized decision tree and KNN with bayesian optimization.* Retrieved from SpringerLink: https://link-springer-com.tilburguniversity.idm.oclc.org/article/10.1007/s41870-022-00974-1

Pappadà, R., & Pauli, F. (2022, 6 30). *Discrimination in machine learning algorithms.* Retrieved from arXiv: https://arxiv.org/abs/2207.00108

Petersen, T. (2005). *Measurement in Discrimination.* Retrieved from ScienceDirect: https://www-sciencedirect-com.tilburguniversity.idm.oclc.org/topics/computer-science/disparate-impact

Radovanović, S., Petrović, A., Delibašić, B., & Suknović, M. (2020, 7 11). *Enforcing fairness in logistic regression algorithm.* Retrieved from IEEE XPlore: https://ieeexplore-ieee-org.tilburguniversity.idm.oclc.org/stamp/stamp.jsp?tp=&arnumber=9194676

Rangel-Díaz-de-la-Vega, A., Villuendas-Rey, Y., Yáñez-Márquez, C., Camacho-Nieto, O., & López-Yáñez, I. (2020, 4 16). *Impact of imbalanced datasets preprocessing in the performance of associative classifiers.* Retrieved from Tilburg University: https://tilburguniversity.on.worldcat.org/atoztitles/link?sid=google&auinit=Y&aulast=Villuendas-Rey&atitle=Impact+of+imbalanced+datasets+preprocessing+in+the+performance+of+associative+classifiers&id=doi:10.3390/app10082779&title=Applied+Sciences&volume=1

Salunkhe, U. R., & Mali, S. N. (2018, 1 20). *A Hybrid Approach for Preprocessing of Imbalanced Data in Credit Scoring Systems.* Retrieved from SpringerLink: https://link-springer-com.tilburguniversity.idm.oclc.org/chapter/10.1007/978-981-10-7245-1_10

STAT 897D. (n.d.). *Analysis of German Credit Data*. Retrieved from STAT 897D: Applied Data Mining and Statistical Learning: https://online.stat.psu.edu/stat857/node/215/

Stevens, A., Deruyck, P., Van Veldhoven, Z., & Vanthienen, J. (2020, 12 1). *Explainability and Fairness in Machine Learning: Improve Fair End-to-end Lending for Kiva.* Retrieved from IEEE Xplore: https://ieeexplore-ieee-org.tilburguniversity.idm.oclc.org/abstract/document/9308371

Szepannek, G., & Lübke, K. (2021, 10 14). *Facing the Challenges of Developing Fair Risk Scoring Models.* Retrieved from National Library of Medicine: National Center for Biotechnology Information: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8552888/

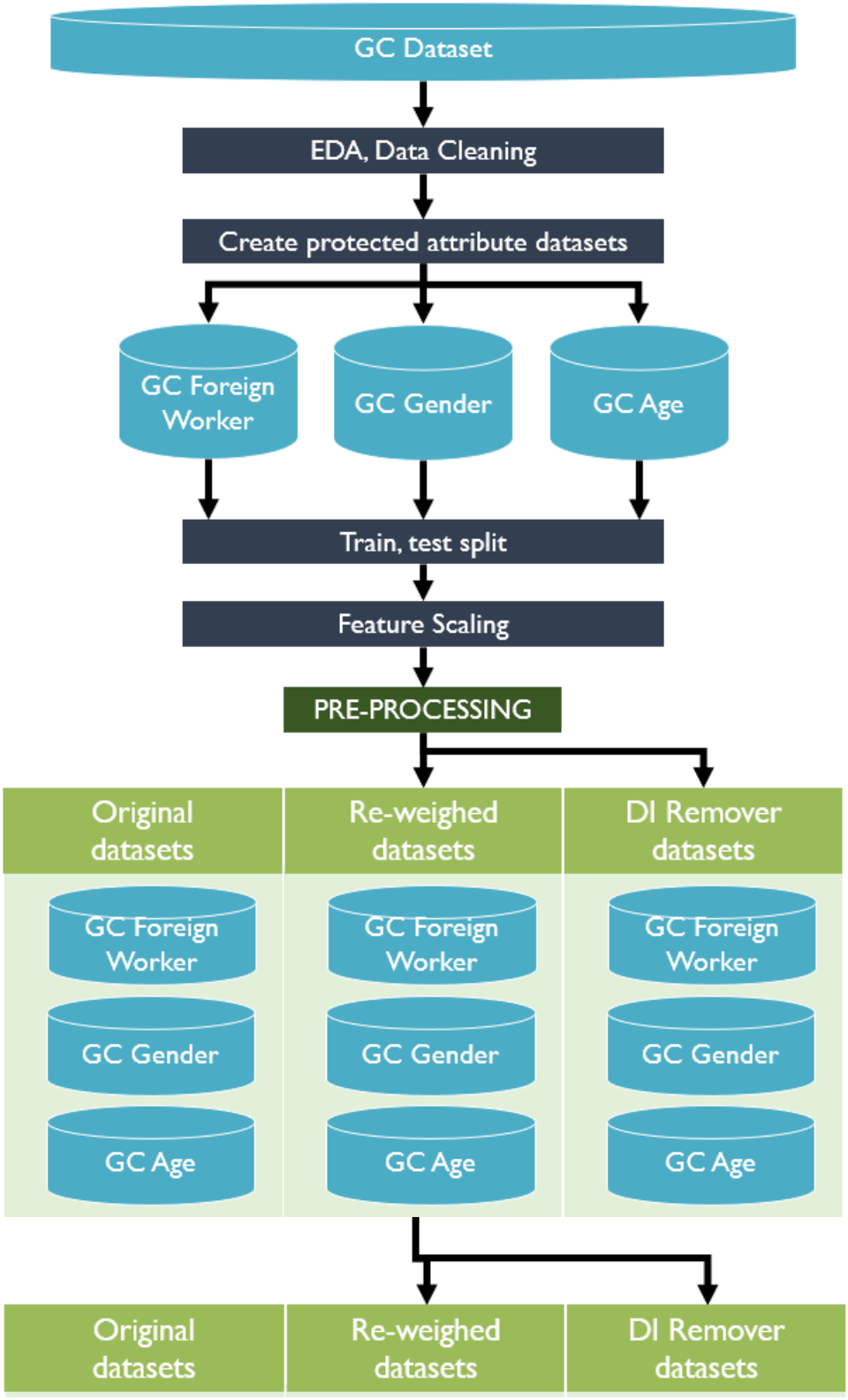
Zelaya, C. V. (2019, 6 6). *Towards Explaining the Effects of Data Preprocessing on Machine Learning.* Retrieved from IEEE Xplore: https://ieeexplore-ieee-org.tilburguniversity.idm.oclc.org/stamp/stamp.jsp?tp=&arnumber=8731532

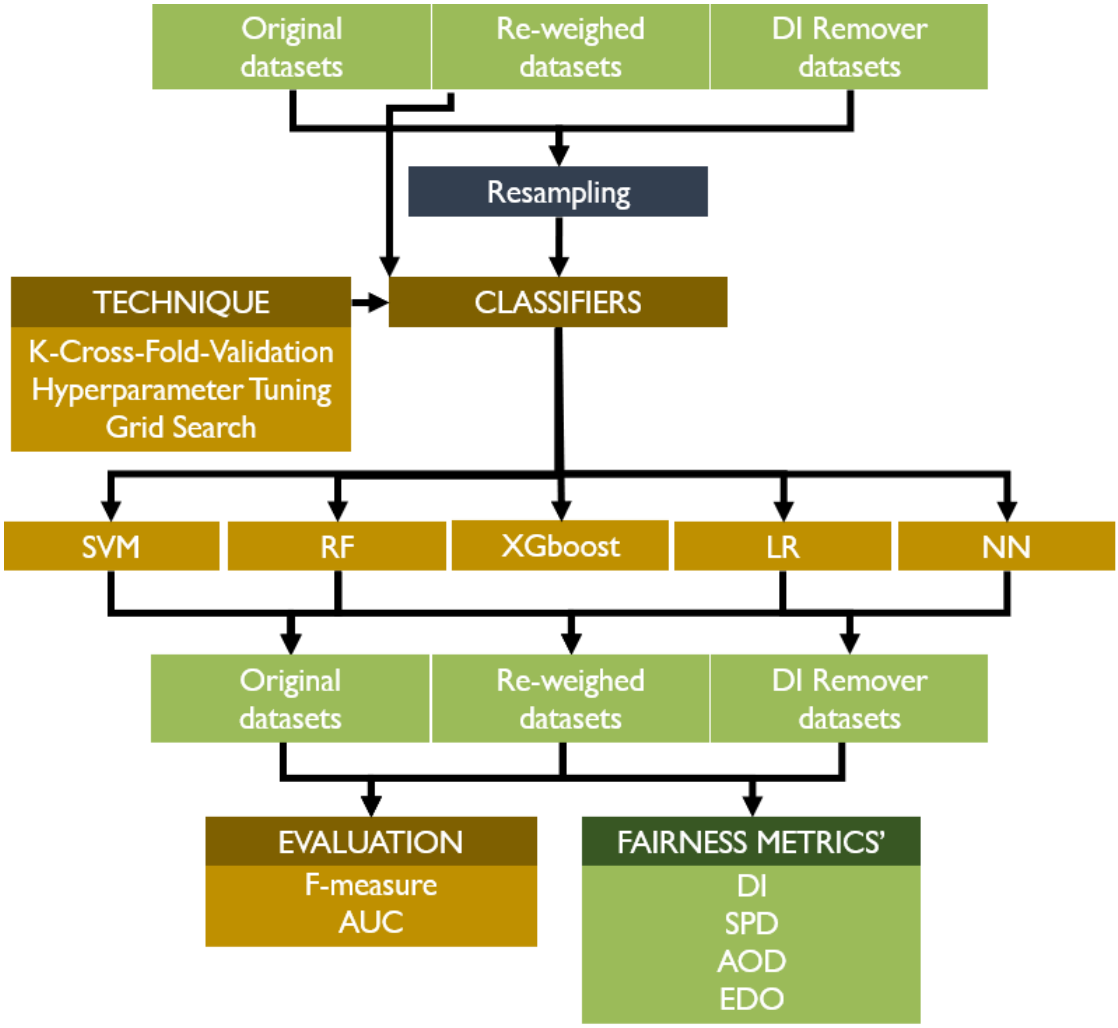
Zhang, Y., & Zhou, L. (2019, 12 16). *Fairness Assessment for Artificial Intelligence in Financial Industry.* Retrieved from arXiv: https://arxiv.org/abs/1912.07211

Zhang, Y., Bellamy, R., Liao, Q. V., & Singh, M. (2021, 5). *Introduction to AI Fairness.* Retrieved from Association for Computing Machinery: https://dl.acm.org/doi/abs/10.1145/3411763.3444998

Appendices and Supplementary Materials

Appendix A: Data Science Pipeline




Original datasets
Re-weighed datasets
DI Remover datasets
Resampling
TECHNIQUE
K-Cross-Fold-Validation
Hyperparameter Tuning
Grid Search
CLASSIFIERS
SVM
RF
XGboost
LR
NN
Original datasets
Re-weighed datasets
DI Remover datasets
EVALUATION
F-measure
AUC
FAIRNESS METRICS'
DI
SPD
AOD
EDO

Appendix B: Hyperparameter settings

Table 1 *Hyperparameter settings for each classifier for the hyperparameter tuning and grid search*

Classifier	Method	Hyperparameters	Settings
SVM	svmPoly	degree	1, 2, **3**
		scale	0.001, 0.01, **0.1**
		C	0.25, 0.5, **1**
RF	rf	mtry	1, 2, **3**, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
XGBoost	xgbTree	nrounds*	5, 10, 15, 20, **25**, 30, **35**, **40**, 45, **50**
		max_depth	3, 5, **7**, 10
		eta	**0.3**
		gamma	**0**
		colsample_bytree	**1**
		min_child_weight	**1**
		subsample	**0.6**
LR	glm	-	-
NN	mlpML	layer1	4, **16**, 32
		layer2	16, 32, **64**
		layer3	**32**, 64, 128

***** Multiple settings are selected due to the fact that the optimal hyperparameter values differ between protected attributes

Table 2 *Hyperparameter settings selected for the radial SVM after hyperparameter tuning and grid search for the original and pre-processed datasets*

Model	**Protected attribute**	**Optimal parameter settings**		
SVM		Degree	Scale	C
Baseline	Foreign worker	3	0.1	1
	Gender	3	0.1	0.25
	Age	3	0.1	1
Reweighed	Foreign worker	3	0.01	0.5
	Gender	3	0.01	1
	Age	3	0.01	0.5
DIR	Foreign worker	3	0.1	1
	Gender	3	0.01	1
	Age	3	0.1	0.25

Table 3 *Hyperparameter settings selected for the Random Forest after hyperparameter tuning and grid search for the original and pre-processed datasets*

Model	**Protected attribute**	**Parameters**
Random Forest		mytry
Baseline	Foreign worker	3
	Gender	4
	Age	8
Reweighed	Foreign worker	3
	Gender	1
	Age	7
DIR	Foreign worker	3
	Gender	7
	Age	2

Table 4 *Hyperparameter settings selected for the XGboost after hyperparameter tuning and grid search for the original and pre-processed datasets*

Model	Protected attribute	Parameters						
XGBoost		nrounds	Max_Depth	Eta	Gamma	colsample_ bytree	min_child_ weight	subsample
Baseline	Foreign worker	40	10	0.3	0	1	1	0.6
	Gender	25	7	0.3	0	1	1	0.6
	Age	50	7	0.3	0	1	1	0.6
Reweighed	Foreign worker	20	3	0.3	0	1	1	0.6
	Gender	35	3	0.3	0	1	1	0.6
	Age	25	7	0.3	0	1	1	0.6
DIR	Foreign worker	35	10	0.3	0	1	1	0.6
	Gender	40	10	0.3	0	1	1	0.6
	Age	50	7	0.3	0	1	1	0.6

Table 5 *Hyperparameter settings selected for the Neural Network (Multilayer Perceptron) after hyperparameter tuning and grid search for the original and pre-processed datasets*

Model	Protected attribute	Parameters		
NN Multilayer Perceptron		layer1	layer2	layer3
Baseline	Foreign worker	16	16	16
	Gender	16	32	64
	Age	16	64	32
Reweighed	Foreign worker	16	64	64
	Gender	4	64	64
	Age	4	64	32
DIR	Foreign worker	16	64	32
	Gender	16	16	32
	Age	16	16	32

Appendix C: Fairness scores vs model performance

Table 1 *Model performance versus the fairness of the SVM Polynomial models for the original and pre-processed datasets*

Model	Protected attribute	Model performance		Fairness			
Polynomial SVM		**AUC**	**F-measure**	**DI**	**SPD**	**AOD**	**EOD**
Baseline	Foreign worker	0.88	0.819	**1.012**	**0.007**	0.226	**-0.012**
	Gender	**0.878**	**0.856**	0.631	-0.189	-0.078	-0.042
	Age	**0.874**	**0.832**	**0.775**	**-0.145**	-0.386	-0.150
Reweighed	Foreign worker	0.872	0.841	1.114	0.085	**0.134**	-0.206
	Gender	0.871	0.838	1.293	0.227	0.127	0.298
	Age	0.872	0.832	0.761	-0.193	**-0.218**	**-0.139**
DIR	Foreign worker	**0.883**	**0.848**	1.524	0.344	0.226	0.343
	Gender	0.825	0.827	**0.77**	**-0.112**	**0.07**	**0.004**
	Age	0.863	0.826	0.704	-0.208	-0.443	-0.221

Table 2 *Model performance versus the fairness of the Random Forest models for the original and pre-processed datasets*

Model	**Protected attribute**	**Model performance**		**Fairness**			
Random Forest		**AUC**	**F-measure**	**DI**	**SPD**	**AOD**	**EOD**
Baseline	Foreign worker	**0.952**	**0.919**	0.773	-0.147	-0.044	**-0.248**
	Gender	**0.923**	**0.905**	0.688	-0.16	**0,0**	**0,0**
	Age	**0.932**	**0.879**	0.628	-0.281	-0.212	-0.227
Reweighed	Foreign worker	0.868	0.865	0.817	-0.150	-0.109	-0.383
	Gender	0.870	0.835	1.159	0.125	0.112	0.190
	Age	0.882	0.843	**0.924**	**-0.062**	-0.028	-0.025
DIR	Foreign worker	0.865	0.838	**0.857**	**-0.095**	**-0.017**	-0.268
	Gender	0.865	0.841	**0.937**	**-0.040**	0.021	0.034
	Age	0.852	0.839	0.674	-0.175	**0.005**	**-0.01**

Table 3 *Model performance versus the fairness of the XGBoost models for the original and pre-processed datasets*

Model	**Protected attribute**	**Model performance**		**Fairness**			
XGBoost		AUC	F-measure	DI	SPD	AOD	EOD
Baseline	Foreign worker	0.93	0.87	0.779	-0.142	-0.028	**-0.243**
	Gender	0.91	**0.89**	0.618	-0.2	**-0.078**	**-0.052**
	Age	**0.936**	**0.886**	**0.864**	**-0.097**	**-0.133**	**-0.066**
Reweighed	Foreign worker	0.882	0.859	0.848	-0.119	-0.076	-0.353
	Gender	0.889	0.844	**1.181**	**0.139**	0.117	0.211
	Age	0.884	0.835	0.633	-0.296	-0.253	-0.254
DIR	Foreign worker	**0.935**	**0.874**	**0.857**	**-0.095**	**0.026**	-0.268
	Gender	**0.928**	0.883	0.649	-0.216	-0.193	-0.157
	Age	0.925	0.866	0.806	-0.140	-0.113	-0.096

Table 4 *Model performance versus the fairness of the LR models for the original and pre-processed datasets*

Model	**Protected attribute**	**Model performance**		**Fairness**			
Generalized Linear Model		AUC	F-measure	DI	SPD	AOD	EOD
Baseline	Foreign worker	0.747	0.706	2.045	0.447	0.16	0.471
	Gender	0.731	0.728	0.664	-0.163	-0.428	-0.154
	Age	0.781	0.735	0.488	-0.323	-0.281	-0.271
Reweighed	Foreign worker	**0.867**	**0.846**	1.413	0.293	0.156	0.342
	Gender	**0.859**	**0.821**	1.223	0.174	0.117	0.243
	Age	**0.865**	**0.833**	**0.809**	**-0.161**	**-0.152**	**-0.129**
DIR	Foreign worker	0.754	0.719	**1.331**	**0.142**	**0.155**	**-0.031**
	Gender	0.783	0.74	**0.846**	**-0.078**	**-0.039**	**0.0**
	Age	0.776	0.738	0.439	-0.332	-0.306	-0.276

Table 5 *Model performance versus the fairness of the NN (Multilayer Perceptron) models for the original and pre-processed datasets*

Model	**Protected attribute**	**Model performance**		**Fairness**			
Multilayer Perceptron		AUC	F-measure	DI	SPD	AOD	EOD
Baseline	Foreign worker	0.854	**0.838**	1.492	0.330	0.194	0.379
	Gender	0.852	**0.838**	0.713	-0.201	**-0.063**	**-0.044**
	Age	0.841	0.808	**0.881**	**-0.078**	-0.054	**-0.038**
Reweighed	Foreign worker	**0.866**	0.827	1.420	0.296	0.183	**0.337**
	Gender	**0.869**	0.823	1.138	0.105	0.104	0.137
	Age	**0.873**	**0.818**	0.818	-0.114	-0.139	-0.059
DIR	Foreign worker	0.834	0.824	**1.408**	**0.241**	**0.069**	0.358
	Gender	0.868	0.828	**1.119**	**0.082**	0.14	0.129
	Age	0.852	0.809	0.693	-0.16	**-0.048**	-0.038