Fair Classifiers with Voluntary, Additional Features

Christian Eberle and Mathias Hega

Abstract— **In some classification scenarios such as credit scoring, a person can choose whether or not to give information regarding a protected feature. However, the classifier is not permitted to treat the missingness of a voluntary feature unfairly. This classification setting is one that is barely touched upon by current fairness research. Our contribution in this paper is twofold: First, we train a logistic regression and a decision tree classifier on a modified version of the german credit dataset that contains a voluntary feature. The experiments show that the classifier is unfair in regard to the missingness of the voluntary feature whenever missingness correlates with bad credibility. We thereby establish the need for fairness constraints in this scenario. Second, we provide a formal definition of this setting and discuss the implications for commonly used fairness definitions.**

I. INTRODUCTION

Our society is continuously seeing more and more situations in which critical decisions are handed off to data-driven algorithms developed using the insights of modern machine learning. These decisions involve, for example, determining whether an individual is fit for receiving a loan, is suitable for a certain job or if they are the one who should receive a specific type of medication that could potentially be life-saving (”classification scenarios”).

Before the world of computational automation and artificial intelligence (”AI”), all these processes could exclusively be handled by other humans who were bound by regulations that prohibited special treatment of other humans based on sensitive attributes like origin, gender or skin color.
Developed countries try to guarantee their citizens unbiased treatment via fairness laws such as the US Civil Rights Act [1]. Yet due to the nature of the software development industry, companies are rarely if ever required to reveal the source code of their algorithms to the public, making third party validation all but impossible and the machine-evaluated results nontransparent. On the other hand, proponents of AI-based decision-making argue that machines are inherently unbiased and thus, algorithms evaluate everyone fairly and equally [2].

A. Motivation

In this paper, we want to showcase that algorithms used for abovementioned classification scenarios can, even when designed with no ill intent, learn to make biased decisions purely based on raw data and thus discriminate against humans. We will look at a real world example in which a person’s eligibility for a bank loan is determined by a machine learning algorithm that operates on information directly given by the loan applicant. More specifically, we will analyze the impact that voluntary answers in surveys can have on the algorithm’s learning process and we will discuss possible approaches on how to deal with missing information in datasets from a fairness standpoint.

To make the term ”fairness” more tangible, a selection of common fairness definitions will be introduced and explained in more detail.
Finally, the paper attempts to outline how a classifier that is supposed to be as fair as possible should implement voluntary features, and explore different types of missingness values.

B. Formal Notation

For this paper, we will use the following notation:

- V: Voluntary feature, with subsets V_m (instances with missing voluntary feature) and V_g (instances with voluntary feature given).
- M: Binary dummy feature encoding missingness. Equals 1 if instance $\in V_m$, and 0 if instance $\in V_g$.
- X: All other features.
- Y: Result of classification.
- $\hat{Y}$: Ground truth label.

C. Paper Structure

In Section two, we introduce the dataset we use for our classification experiments, as well as the way in which the data was preprocessed. Additionally, we provide an overview over important fairness criterions. For a selection of fairness criteria, we also apply the definitions to our specific problem setting.
The third section presents the results of our experiments, followed by the fourh section discussing our findings. To conclude the paper, we give a brief summary in section five.

II. METHODS

A. Dataset

The original German Credit Dataset [3] is a staple in fairness literature and consists of 1000 entries with each entry representing the information given by a single interviewee that applied for a bank loan. In total, the dataset makes use of 20 attributes (”features”), 7 of which are numerical and 14 of which are qualitative. For a complete list, view Table 1.
An additional attribute which serves the purpose of describing whether a loan applicant has good or bad credit score is also included. This attribute is referred to as the ground truth label. It also needs to be mentioned that, for reasons unknown to us, the German Credit Score dataset does not contain any instances of a single female individual.

Feature	Variable type	Num. of classes
Status of existing checking account (voluntary)	qualitative	4+1
Duration in months	numerical	-
Credit history	qualitative	5
Purpose	qualitative	11
Credit amount	numerical	-
Savings accounts/bonds	qualitative	5
Present employment since	qualitative	5
Installment rate in percentage of disposable income	numerical	-
Sex	qualitative	2
Other debtors/guarantors	qualitative	3
Present residence since	numerical	-
Property	qualitative	4
Age (in years)	numerical	-
Other installment plans	qualitative	3
Housing	qualitative	3
Number of existing credits at this bank	numerical	-
Job	qualitative	4
Number of people being liable to provide maintenance for	numerical	-
Telephone	qualitative	2
Foreign worker	qualitative	2
Missingness	qualitative	2

TABLE I: Our modified version of the German Credit dataset. The feature ”Status of existing checking account” (”Account Balance”) has one additional class, for instances missing this feature. Furthermore, the feature ”Sex and Marital Status” was converted into ”Sex” and the new feature ”Missingness” was added.

B. Overview of Commonly Used Fairness Criteria

Definitions based on Predicted Outcome

(1) **Group Fairness/Statistical Parity**
See C. i) below

(2) **Conditional Statistical Parity**
See C. ii) below

Definitions Based on Predicted and Actual Outcome

(3) **Predictive Parity/Outcome Test**
See C. iii) below

(4) **False Positive Error Rate Balance**
A loan applicant with a missing voluntary feature (”mvf.”) and a bad credit score should have the same chance to incorrectly receive a good score as an applicant with no missing voluntary feature.

(5) **False Negative Error Rate Balance**
A loan applicant with a mvf. and a good credit score should have the same chance to incorrectly receive a bad score as an applicant with no mvf.

(6) **Equalised Odds/Disparate Mistreatment**
Loan applicants with a good credit score and applicant with a bad credit score should be classified similarly regardless of whether their voluntary feature is missing or present.

(7) **Conditional Use Accuracy Equality**
The probability of a loan applicant with a bad predicted credit score to have a good credit score **and**
the probability of a loan applicant with a good predicted credit score to have a bad credit should not differ dependent on whether mvf. are present or not

(8) **Overall Accuracy Equality**
The probability for an applicant with a good credit score to be assigned a good predicted credit score **and**
the probability for an applicant with a bad credit score to be assigned a bad predicted credit score is the same for applicants with mvf. and for applicants without mvf.

(9) **Treatment Equality**
The ratio of false positive to false negatives is equal between applicants with mvf. and applicants without mvf.

Definition Based on Predicted probabilites and Actual Outcome

(10) **Calibration/Test-fairness**
For any given predicted probability score s in an interval $[0, 1]$, the probability of having a good credit score should not differ for applicants with mvf. and applicants without mvf.

(11) **Well Calibration** Well Calibrations is an extension of Calibration. In addition to all criteria from Calibration, Well Calibration requires that for any predicted probability score s, applicants should not only have an equal probability to receive a positive score independently from mvf., but the probability itself should also be equal to the predicted probability score.

(12) **Balance for Positive Class**
The expected value of probability assigned by a classifier to applicants with mvf. and applicants without mvf. who both have good credit score should be equal.

(13) **Balance for Negative Class**
A reversed version of Balance for Positive Class. The expected value of probability assigned by a classifier to applicants with mvf. and applicants without mvf. who both have a bad credit score should be equal.

Similarity-based Measures

(14) **Causal Discrimination**
Applicants with and without mvf. who otherwise have the same features will either both receive a good credit score or both receive a bad one.

(15) **Fairness through Unawareness**
See C. (iv) below

(16) **Fairness through Awareness**
Applicants with and without mvf. who are similar with respect to a particular attribute should be classified similarly. [4]

C. Our Main Fairness Criteria

We have singled out 4 specific fairness criteria that we consider for our task and will explain them below. For more potentially applicable fairness criteria, see [5].

(i) **Group Fairness/Statistical Parity**
This definition is satisfied by a classifier if subjects in both protected and non-protected groups have an equal probability of getting a positive classification outcome. In the context of our classification example,

the protected group is V_m, the group of people who are missing their voluntary feature. This means that as an applicant for a loan, if you refuse to provide your voluntary feature you still have the same chance of being deemed credible for a loan as an applicant who provided the voluntary feature.
To put this more formally, the classification result $Y = 1$ is independent of the missingness feature M:

$$P(Y = 1|M = 0) = P(Y = 1|M = 1)$$

(ii) **Conditional Statistical Parity**
This criterion extends criterion (i) by allowing a specific set of legitimate features L to affect the classification outcome. The legitimate features include all other features X plus the voluntary feature V if provided, otherwise it includes only X. The intuition behind this fairness criterion is the same as for (i), only here we also take into consideration the legitimate features L.
Formally, the classification result $Y = 1$ is independent of the missingness feature M, given that the legitimate features L are the same:

$$P(Y = 1|M = 0, L = l) = P(Y = 1|M = 1, L = l)$$

$$\iff \quad P(Y = 1|M = 0, X = x, V = v)$$
$$= P(Y = 1|M = 1, X = x)$$

This criterion comes with the practical challenge that the dimensionality of the classifier input changes for $M = 0$ and $M = 1$. Most classifiers don't permit the usage of missing values in the input, and a dummy value has to be found. This is a difficult task, and different options for replacing missing values are discussed later.

(iii) **Predictive Parity/Outcome Test** According to this criterion, a classifier is fair if it has the same predicted positive value and false discovery rate for all groups. The intuition here is that the classifier is equally good at predicting outcome for all groups. In formal terms, the ground truth label $\hat{Y}$ is independent of the missingness feature M, given the classification result $Y = 1$.

$$P(\hat{Y} = 1|Y = 1, M = 0) = P(\hat{Y} = 1|Y = 1, M = 1),$$
$$P(\hat{Y} = 0|Y = 1, M = 0) = P(\hat{Y} = 0|Y = 1, M = 1)$$

(iv) **Fairness through Unawareness** This classifier is satisfied if sensitive attributes are omitted from the decision-making process. Similarly as for (ii), this poses the challenge that the input dimensionality is different depending on M. Removing explicit information on the missingness is done easily by dropping the dummy feature. This makes "Fairness by Unawareness" very easy to implement. However, the information on missingness is still implicitly contained in the voluntary feature. A sufficiently complex classifier is then able to infer this information and include it into its decision-making. Therefore, we think that this criterion is not a good choice for our scenario.

D. Preprocessing

By default, the feature "Sex & Marital Status" is encoded asymmetrically in regard to sex. Out of the four classes contained in the dataset, three are male categories and only one is female. To avoid bias due to more classes representing males than females, we changed this feature to "Sex" by merging all male classes.

To simulate the partial missingness of a voluntary feature, we randomly picked instances and overwrote their voluntary feature value with a given missingness value. This missingness value was chosen to be the either the mean of all feature instances (data mean), the mean of all feature classes (class mean), or one of the feature classes chosen at random (random class). Additionally, we created a new, binary feature "Missingness" that encodes whether or not a person is missing its voluntary feature.

Furthermore, a vector of class frequencies was given as input to determine the relative portion of missingness per class. For example, the vector $[0.8, 0.2, 0.1, 0.0]^T$ applied to "Account Balance" would result in 80% of the instances in the first class missing this feature, 20% of instances in the second class, 10% of instances in the third class, and no instances in the fourth class.

For classification, we implemented off-the-self logistic regression and decision tree models in Python [6]. We chose these models because they are both inherently explainable, allowing for better analysis of whether or not the model discriminates against missingness of the voluntary feature.

To measure the influence of missingness on the classification result, we examined the coefficients for the logistic regression models and the feature importances for decision tree models. For our logistic regression models, we used the default parameters. Notably, this means that our models included L2 regularization. We also ran the experiments using L1 or no regularization, which led to very similar results. Our decision trees used Gini impurity to measure the quality of a split and had a maximum depth of 3. Matching the training procedure of [5], we applied ten-fold-cross-validation with 90% of the data as training data and the other 10% being test data.

For the logistic regression models, we normalized the data using min-max scaling. This was done to ensure interpretability of the regression coefficients. All of the code that led to our results is publicly available and can be found here.

III. Results

Initially, we planned on turning the feature "Sex" into a voluntary feature since other papers use this as a protected feature for their fairness experiments [5]. However, early test runs showed that the logistic regression coefficients of the feature ”Sex” was one of the lowest and therefore of little importance . To obtain more meaningful results, we chose the feature corresponding to the highest regression coefficient instead. Thus, the categorical feature ”Account Balance” was used as voluntary feature.

A. *Logistic Regression*

To establish a baseline, we first trained a logistic regression classifier on the full dataset without any voluntary features. The highest regression coefficient of this model was for the feature ”Account Balance”, with 1.68. We then trained another model on a modified dataset where ”Account Balance” was removed. After removing ”Account Balance” the test accuracy dropped from 76.1% to 73.8%. This shows that ”Account Balance” contains important information for credit score prediction.

Next, we simulated the partial missingness of ”Account Balance” to turn it into a voluntary feature. In the following, we will compare different choices of missingness class frequencies as well as missingness values.

First, we tested different frequency values for the worst class, while from all other classes 20% of instances were missing. The results are visualized in Fig. 1. With 0% of the worst class missing its voluntary feature, the regression coefficient is 0.3. The higher the missing rate of instances from the worst class, the more the coefficient for ”Missingness” decreases.

Varying the missing rate of the best class has the opposite effect: The higher the missing rate of instances from the best class, the more the coefficient increases. This shows that whether or not missingness of the voluntary feature is beneficial for the predicted credit score depends on the class distribution underlying the ”Missingness” feature.

Before normalization, the classes of Account Balance have numerical values $[1, 2, 3, 4]^T$. The resulting class mean is 2.5, which is only slightly different from the data mean at 2.577. However, even this small difference already leads to a visible decrease of the missingness coefficient values when using the data mean, compared to using the class mean. Whenever ”Missingness” has value 1, using the data mean will lead to a slightly better credit score prediction because the important feature ”Account Balance” is 0.77 higher. Presumably, decreasing the coefficient for Missingness is a way of compensating for this difference.

B. *Decision Trees*

In comparison to the logistic regression models, our decision tree models generally performed worse. The baseline model trained on the full dataset achieved a test accuracy of 73.1%. Training on the modified dataset without the Account Balance feature lowered the test accuracy to 71.1%. Account Balance is again an important feature as it occurs in the first node of the baseline model. The decision threshold of that node is at 2.5, thereby separating classes 1 and 2 from classes 3 and 4.

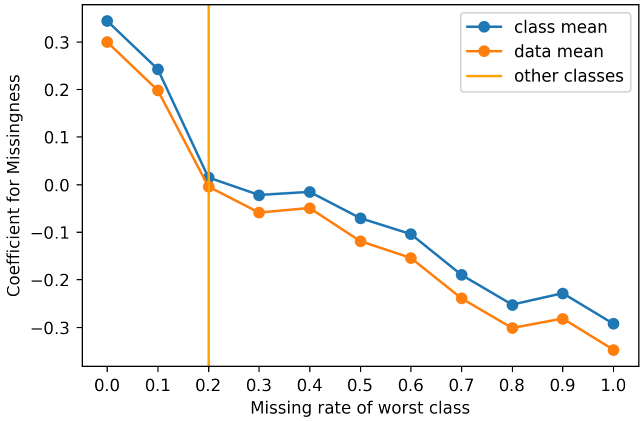

Fig. 1: Logistic regression coefficients depending on the missing rate of the worst class within the feature ”Account Balance”. A high coefficient corresponds to a positive impact on the predicted credit score.

If class mean or data mean were used as missingness value, the feature Missingness had a feature importance close to zero. If the underlying distribution of Missingness is ”bad” and contains many instances from the worst class, the decision threshold at the first node changes to 2.75. By doing so, the decision tree places all missing instances into the same category as the ”bad” classes 1 and 2. Conversely, if the underlying distribution is ”good” the threshold decreases to place all missing instances into the same category as the ”good” classes 3 and 4. As a result, the Missingness feature becomes redundant in the classification process.

To prevent the decision tree from inferring the missingness by using the missingness value, we set the missingness value to be a randomly chosen class. Subsequently, the feature importance of Missingness increased. We tested various missing rates of the worst class to explore the relationship between missing rate and the feature importance of Missingness. The results are shown in Figure 2. Notably, the feature importance of Missingness and Account Balance seem to be symmetrically opposed. If important information is contained within the Missingness feature, this means that Account Balance contains less information. Therefore, whenever the feature importance of Missingness is low, the feature importance of Account Balance is high and vice versa.

IV. DISCUSSION

Implementing a classifier that is fair in regard to the missingness of a voluntary feature comes with many challenges. A naive approach would be to simply train two classifiers separately, one on the subset of data where

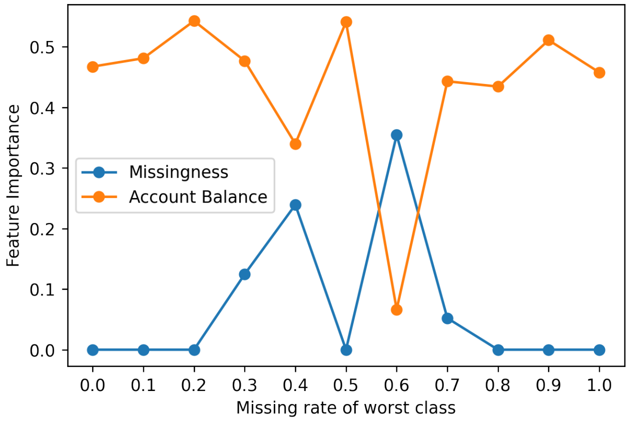

Fig. 2: Feature importances of the decision tree model. While the missing rate of the worst ”Account Balance” class changes, the missing rates of all other classes remain constant at 0.2.

the voluntary feature is missing and another on the rest of the data where the voluntary feature is given. The downside of this approach is that if a negative correlation between missingness and classification result exists, the first classifier will generally give worse predictions than the second classifier. Instead, we propose the usage of a binary dummy variable that encodes missingness. This dummy variable can then be treated as a protected feature, and known methods of implementing fairness constraints can be applied.

In our experiments, we trained a logistic regression and decision tree model on a modified version of the German credit dataset, a commonly used dataset in fairness research [5], [7]–[9]. The test accuracy achieved with the baseline models for logistic regression and decision trees are relatively low, with 76.1% and 73.1% respectively. We assume that this is due to the noisiness of the dataset as well as the simplicity of our models. Other research shows that using more sophisticated methods can improve the test accuracy to be over 80% [7], [10].

However, the main focus of our experiments was not test accuracy but the effect of the Missingness feature on the classification outcome. We showed that for both logistic regression and decision trees, this effect is dependent on the underlying distribution of missingness. If the individuals with missing voluntary feature come predominantly from a ”bad” class that correlates with bad credibility, missingness will have a negative effect on the predicted credibility. This means that by default, classification with voluntary features is not fair in regard to the missingness of the voluntary feature. While our experiments only showed this for logistic regression and decision tree models, we can safely assume that this finding extends to more complex models such as Deep Neural Networks. If a correlation exists between missingness and credibility, a well-working classifier will find and use it. From this comes the need for fairness research specifically for the scenario of voluntary features.

Since most classification models can’t handle sparse input features, not existing values in the voluntary feature have to be substituted by some missingness value. However, it is not obvious how to encode missingness. Depending on the classification model, different substitution methods are required. We propose three possible solutions, each with their own advantages and disadvantages. The first is using the mean of all classes. This suggests itself especially for binary features, such as sex. However, it is not possible at all when the voluntary feature is continuous and takes arbitrarily large values. Moreover, if the classes are not evenly distributed then the class mean might be a poor substitution that does not represent the population average very well. Our second proposed solution tries to tackle this issue by using the data mean of the voluntary feature. This works well if the sampled data is a reasonable representation of the population it is drawn from. In the case of our example dataset, the class mean and data mean are very similar and both seem like a reasonable choice for our logistic regression classifier.

The distinction between data and class mean is only relevant for linear classifiers, such as our logistic regression model. Our decision tree model was able to infer missingness from the missingness value, and adjusted its decision boundary accordingly. To prevent this, we chose one of the class values at random to use as missingness value. This way, the missingness value contained no information and the classifier was forced to infer this information from the Missingness feature. This is important, since otherwise the Missingness feature wouldn’t play a role at all, and applying a fairness criterion to a feature of no to little importance has little effect. A good example for this is fairness through unawareness: If we were to drop the Missingness feature, it wouldn’t change the behavior of our decision tree model at all.

V. CONCLUSION

In this paper, we showed that our classifier discriminates against the missingness of a voluntary feature if the missingness correlates negatively with credibility. We can assume this to be the case in many real-life scenarios, where discrimination might be immoral or illegal. Therefore, we need to research appropriate fairness criteria for the use of voluntary features. Most fairness research focuses on applying different fairness definitions in regard to a protected feature. However, there currently exists no research on how to apply these definitions to a feature where discrimination is permitted against some, but not all instances of the feature. We propose to solve this by creating an additional dummy feature that encodes missingness in a binary fashion.

REFERENCES

[1] "Civil Rights Act of 1964 Title VII equal employment opportunities," https://www.eeoc.gov/statutes/title-vii-civil-rights-act-1964, accessed: 2022-07-07.

[2] "Can Algorithms Be Fairer Than People?" https://lab.cccb.org/en/can-algorithms-be-fairer-than-people/, accessed: 2022-07-09.

[3] "German Credit Dataset," https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data), accessed: 2022-07-09.

[4] T. P. Cynthia Dwork, Moritz Hardt, O. Reingold, and R. Zemel, "Fairness through awareness," *arXiv:1104.3913*, 2011.

[5] S. Verma and J. Rubin, "Fairness definitions explained," in *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 2018, pp. 1–7.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[7] S. Verma, M. Ernst, and R. Just, "Removing biased data to improve fairness and accuracy," *arXiv preprint arXiv:2102.03054*, 2021.

[8] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International conference on machine learning*. PMLR, 2013, pp. 325–333.

[9] F. Buet-Golfouse and I. Utyagulov, "Towards fair unsupervised learning," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1399–1409.

[10] A. Khashman, "Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes," *Expert Systems with Applications*, vol. 37, no. 9, pp. 6233–6239, 2010.

[11] "Civil Rights Laws in the US can be violated by algorithms," https://www.aclu.org/blog/privacy-technology/surveillance-technologies/big-data-can-be-used-violate-civil-rights-laws-and, accessed: 2022-07-07.