# In search for the decent person: imposing intersectional group fairness to a credit decision model

December 11, 2023

## Introduction

## Context

The imagined setting studied in this project is as follows. A credit institution is building their first model for prediction whom to give credit to. In light of the new artificial intelligence legistlation they want to make sure that none of the regulations are compromised when introducing algorithmic decision-making. Additionally, the company managers are interested in if the new rules can be exploited even to their financial advantage: if imposing fairness could on one hand make sure that the model is lawful, but on the other hand also possibly increase company profits. A team of analysts are tasked to investigate these matters.

## Technical setting: a random forest for credit default prediction

The dataset used is the German credit dataset [1]. Notably, the dataset is from 1994, which makes its predictive power for the future less certain, especially with respect to fairness: equality has greatly advanced between now and the date of the dataset. Of the features available in the data, some are additionally considered discriminatory as features in decision-making in the present: marital status, sex, age, and immigrant status. Lastly, a significant constraint from the data is its size: there are only 1000 rows.

The dataset was preprocessed by encoding categorical features into one-hot encoded dummy variables or numerical variables depending on the ordinality of each variable. A summary of encodings can be found in table 1 [1]. For hyperparameter optimization, a simple grid search was performed, and chosen parameters can be found in table 2 [1]. Most notable of these are add most notable params here

The model type used to predict credit defaults was a random forest, since the dataset description named it as the method with the best baseline accuracy, which is in line with the common knowledge that it is one of the best out-of-the-box model types [1]. The accuracy listed as baseline, 78%, was achieved for the initial model. For testing fairness levels, leave-one-out cross-validation was done to obtain a test prediction for each data point. This was opted for instead of a separate test set because of the small dataset size: we assumed that slight deviation of the model because of leaving out one data point is less bad than only measuring fairness on a very small data set and not using all data for training.

Specifically, we used the sklearn v. version here RandomForest [1] for classification, and cross-val-predict for obtaining test predictions for all data points in the dataset.

## The fairness condition: intersectional group fairness

To motivate the choice of fairness constraint, we trace back to the foundational goal of credit prediction. The goal of a credit model is to output a probability that expresses one's likelihood of paying it back. Thus, the company objective is to find the decent people: those who duly take care of their resposibilities.

The assumption of credit modeling is that a person's financial status correlates with one's decentness. However, the problem with the use of old data is that the data encodes stereotypes: since no ground truth on default rates is given, one can presuppose that the data penalizes disadvantaged groups in credit grant decisions because of blunt discrimination. To bring this point further, one can even assume that the world encodes stereotypes: it is a known fact, for example, that eqully qualified females tend to earn less than men. This means that the mapping from earnings to decentness between sexes is not equal, and that our foundational assumption is broken: the mapping of nonsensitive features to default rates is not necessarily uniform across different demographic groups.

The hypothesis for correcting this fault is the following. We assume that within each group, grouped by level of societal discrimination, the top percentile of those modeled as most creditworthy is, in reality, equally creditworthy. For the grouping, we assume a phenomenon of compounding: having one disadvantageous trait is better than having two, and the relation of financial status to creditworthiness is unequal also between these groups.

To measure this, we introduce the metric of intersectional group fairness:

definition: intersectional group fairness between all combinations of disadvantage, the true default rate in percentile i is equal for all i.

goal for profit: find the most decent people

assumption of modeling: you can use data to get the target. problem with old data: encodes nontrue stereotypes, such as women default, young default

so: assumption may be broken?

hypothesis: order people by decentness within stereotypical groups. cutoff, above that, grant credit -¿ you get the most decent people key insight: this should also maximize profit grouping by how much stereotypical discrimination intersectionality: more disadvantages compound to more discrimination

additional +'s: profit, affirmative action, law

also, affirmative action. additional +, brand prestige we dont know if this is true but we can check business metrics after implementing the model. question: stereotyped groups are more decent than is assumed by data and nonfair models

also, law law: requires individual fairness. one with similar nonsensitives should have equal decision group fairness causes discrimination in individual fairness sense to advantaged groups allowed in law for affirmative action

## Pre-remedy fairness level

The level of group fairness was measured by splitting the data set in two, one being the disadvantaged group and one for all other subjects, Then, credit grant rates were compared between the groups. This was conducted for all combinations of disadvantages. The results for these experiments are listed in table 3. [1]

The most significant insights we can see from the table are the comparative impact of the disadvantages, and compounding. The results show that age seems to cause the most unfair treatment, whereas sex has the least impact on creditworthiness, with immigrant status between these. As for joint effects, we see the assumed pattern of compounding: having more disadvantages puts one in an even worse position, with the combination of all disadvantages being the worst outcome of all.

## Imposing fairness: correction mechanism here

remedying fairness technique - rationale for using - results for a few techniques, find the best one

## Conclusion

conclusion: suggestions for the institution which corrections made sense? did they remedy the problem? did the problem make sense?

see how business metrics evolve

change fairness constraint if necessary: make disadvanteged even more or less represented, according to their true default rates

# notes

technicalities

dataset we used, most significant characteristics - genders and marital statuses - 1000 rows - old -¿ imposing equality for relevance

the model we used random forest reason: dataset stated that it has the best baseline accuracy of about 78which we also achieved hyperparameter optimization: simple grid search params obtained and short description of their significance class encoding: one hot / self reasoned numericals, discussion on eg if being in unpaid housing is better creditwise than renting

sklearn random forest, sklearn loo cv

# References


[1] S. Thrun, M. Montemerlo, H. Dahlkamp, D. Stavens, A. Aron, J. Diebel, P. Fong, J. Gale, M. Halpenny, G. Hoffmann, K. Lau, C. Oakley, M. Palatucci, V. Pratt, P. Stang, S. Strohband, C. Dupont, L.-E. Jendrossek, C. Koelen, C. Markey, C. Rummel, J. van Niekerk, E. Jensen, P. Alessandrini, G. Bradski, B. Davies, S. Ettinger, A. Kaehler, A. Nefian, and P. Mahoney. "Stanley: The robot that won the DARPA Grand Challenge". In: *Journal of Field Robotics* 23.9 (2006), pp. 661–692. DOI: https://doi.org/10.1002/rob.20147. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/rob.20147. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.20147.