IN-STK5000 – Adaptive Methods for Data-Based Decision Making
Project: Credit risk for mortgages - part 3

Luca Attanasio, S M Mamun Ar Rashid

September 2019

1 Introduction

Some fair models have already been discussed and implemented previously: always grant model, random model. The first model is fair because everyone is granted a loan. The second model is fair because everyone is granted a loan with the same probability of $1/2$ of being accepted. However, individual samples of data need to be analyzed from a mathematical perspective to find a better model which grants fairness and possibly enhances the utility. If the utility is enhanced, then the bank has a better income.

2 Applying fairness

In order to apply fairness, we could use a bayesian logistic regressor, since bayesian models are implicitly fair as discussed in [1]. Decisions can be balanced with respect to the feature: *credit history_A31*. Notice that it makes sense to balance our decision based on this feature, because it indicates whether all credits at this bank were duly paid back. Balancing fairness on this feature helps to make fair decisions in this regard.

Notice that around 70% of the dataset pays back their loans. In Table 1, we evaluate metrics on the whole dataset. 71% of those who paid duly back their credits, repaid the loan. 43% of those who did not duly pay back their credits paid back their loan. The deviation is calculated with respect to the amount of people who paid back their loan in the dataset: 70%.

To apply fairness on a policy π, the utility function has to be customized [1] w.r.t to the utility defined in the previous parts of the assignment. The new utility U^f must include a fairness term F:

$$E(U^f|A) = (1-\lambda)E(U|A) + \lambda E(F) \tag{1}$$

where λ is the amount of fairness which we can inflate (set to 0.4 in this paper). F is the fairness term to balance the sensitive variable. In the following equation,

	repaid	tot	prob_repaid	deviation
credit history_A31				
0	679	951	0.713985	0.013985
1	21	49	0.428571	0.271429

Table 1: Calculating the true probability of paying back the loan based on the value of Credit history_A31. The policy should not be fair in order to maximize the utility, otherwise the bank may loose more money.

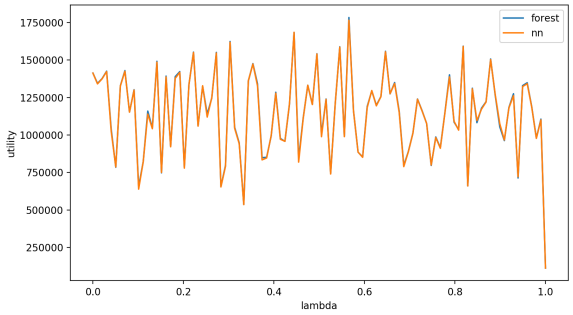

Figure 1: Utility function when changing lambda in a range between 0 and 1.

the action $a = \{A, B\}$ can be chosen between A, granting the loan, B not granting the loan. y is the outcome, z is the sensitive variable:

$$F_\theta^\pi(a, y, z) = |P_\theta^\pi(a|y, z) - P_\theta^\pi(a|y)| \tag{2}$$

Then, U^f must be maximized using the following rule:

$$\max\{E(U^f|A), E(U^f|B)\} \tag{3}$$

The plot in Figure 1 shows how the utility changes when applying fairness with different λ. The utility does not suffer from significant effects when applying more fairness.

2.1 Measuring the fairness of the policy

How can you measure whether your policy is fair?
The policy is fair if it makes fair decisions on the testing set w.r.t. one sensitive feature. To measure the fairness of the policy with respect to a sensitive variable z, the balance is evaluated. The feature z is *balanced* if:

$$P_\theta^\pi(a|y, z) = P_\theta^\pi(a|y) \tag{4}$$

where a is the action chosen (grant/not grant), z is the feature *Credit history_A31* and y is the true outcome. In addition, π is the policy and θ the parameter set of the algorithm.

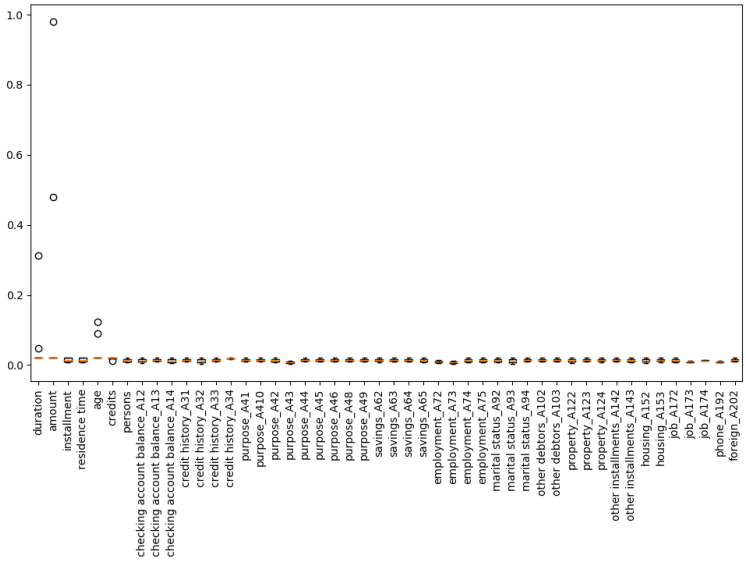

Figure 2: Balance w.r.t. all features in the dataset

Calculating the fairness can be done by obtaining $F_\theta^\pi(a, y, z)$. If this value is low, the policy is making fair decisions.

In Figure 2, the balance w.r.t. to every feature of the dataset z, using the testing set (the parameter a is known), is calculated. The model chosen in the evaluation is the random forest. Since in most cases the balance is really low, the policy is fair.

How does the original training data affect the fairness of your policy?
Statistics on the original data were evaluated. For example, look at Table 1. Regarding the other features of the dataset, some variables are implicitly fair, others are less fair.

• *Identify sensitive variables. Do the original features already imply some bias in data collection?*
One sensitive variable is the *Credit history_A31* feature, the value analyzed in this report.

• *Analyse the data or your decision function with simple statistics such as histograms.*
In our assignment two models were developed: *neural network*, *random forest*. We now focus on trying to understand if the *random forest* is making fair decisions. First of all, as shown in the previous parts of the assignment, the model is almost always granting loans. This states that the model is fair. This assumption is verified by evaluating the amount of loans granted w.r.t. to the total loans on the testing set Figure 3. As a reminder, the sensitive variable is $z = \{Credit\ history\_A31\}$. As a comparison measure, when using the true outcomes on the whole dataset, the following histogram was obtained Figure 4,

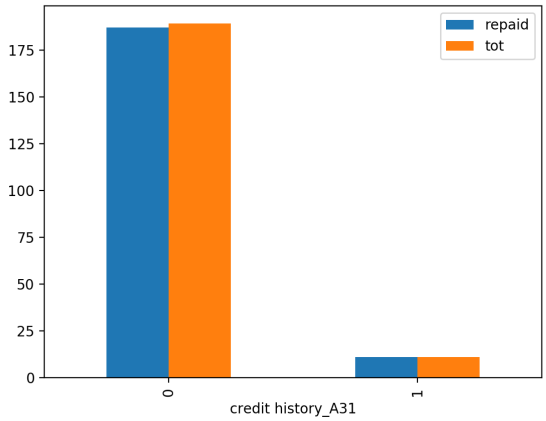

Figure 3: Amount of loans granted (repaid) w.r.t. the total loans (tot).

see also Table 1.

2.2 Measuring the variation of the action for different outcomes when the sensitive variable changes

• For balance/calibration, measure the total variation of the action (or outcome) distribution for different outcomes (or actions) when the **sensitive variable varies**.
The Figure 3 visualises the requested measurement on the sensitive variable.

• *Advanced: What would happen if you were looking at fairness by also taking into account the* ***amount of loan*** *requested?*
If looking at fairness by taking into account the amount of loan requested, the expected utility would change and consequently the decisions would be balanced with respect to the amount of loan requested. Those who ask for less would be granted loans with the same probability of those who ask for more.

3 Additional content

3.1 Measuring fairness

Another way to measure fairness is by calculating the calibration value. A policy π is calibrated for parameter θ with respect to the sensitive variable z if:

$$P_\theta^\pi(y|a,z) = P_\theta^\pi(y|a) \tag{5}$$

for each a and z. Calibration means that $y \perp\!\!\!\perp z|a,\theta,\pi$: y is independent of the sensitive variable z given the action a of granting or not the loan. This means

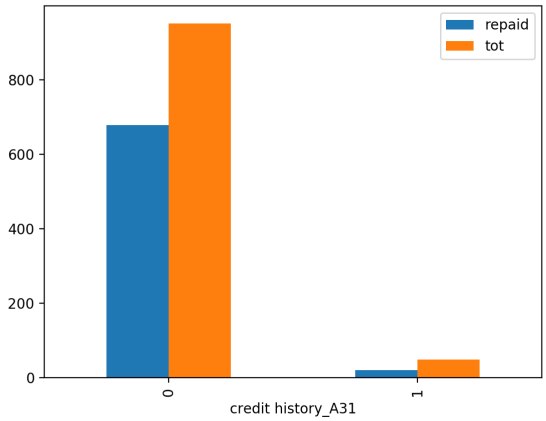

Figure 4: True amount of loans granted (repaid) w.r.t. the total loans (tot).

the distribution of outcomes is the same indipendently of the action and value of z (we are fair w.r.t. to z).

To calculate the first probability in the previous equation, let's suppose our model is depicted in Figure 5.

The problem to be faced is how to get an estimate of $P_\theta^\pi(y|a,z)$, from the dataset. To do so, we can estimate the calibration value using the marginalization property based on our model.

$$P_\theta^\pi(y|a,z) = \sum_x P_\theta^\pi(y|x,a,z) P_\theta^\pi(x|a,z) \tag{6}$$

y is independent of a and z and $P_\theta^\pi(x|a,z)$ can be rewritten using Bayes rule:

$$P_\theta^\pi(y|a,z) = \sum_x P_\theta^\pi(y|x) \frac{P_\theta^\pi(a|x,z) P_\theta^\pi(x|z)}{P_\theta^\pi(a|z)} \tag{7}$$

Since a only depends on x:

$$P_\theta^\pi(y|a,z) = \sum_x P_\theta^\pi(y|x) \frac{\pi(a|x) P_\theta^\pi(x|z)}{\sum_{x'} \pi(a|x') P_\theta^\pi(x'|z)} \tag{8}$$

where $\pi(a|x)$ is the value of the policy for an action given the sample x, $P_\theta^\pi(y|x)$ is the probability, taken from the classifier, that the outcome is y, given a sample x. $P_\theta^\pi(x|z)$ can be obtained with Bayes rule:

$$P_\theta^\pi(x|z) = \frac{P_\theta^\pi(z|x) P_\theta^\pi(x)}{P_\theta^\pi(z)} \tag{9}$$

where $P_\theta^\pi(z|x)$ can be estimated from a classifier and $P_\theta^\pi(z)$ or $P_\theta^\pi(x)$ from the dataset.

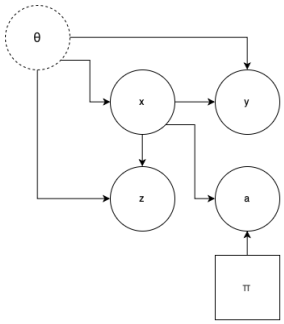

Figure 5: Model for fairness.

3.2 Minimizing calibration value.

To optimize the probability value with respect to the policy, we must calculate it's derivative and set it to 0. The value results in a minimum in the function if it's second derivative order is lower than 0.

References

[1] Y. Liu, G. Radanovic, C. Dimitrakakis, D. Mandal, and D. C. Parkes, "Calibrated fairness in bandits," *arXiv preprint arXiv:1707.01875*, 2017.