Fair Machine Learning Problem

1 Fair Machine Learning Problem

1.1 Fairness and Causality

The head of the loan department of your bank has given you an anonimyzed dataset of records of previous customers. This dataset reports as features some demographic and financial informations: *ethnic group* (a binary value denoting whether the customer belongs to the group A or B), *wage* (expressed in thousand of dollar per month), and an *education degree* (expressed as a normalized real value). Moreover, for each customer the bank has collected a *repayment* binary value recording whether the customer promptly repaid the loan he or she had taken from the bank.

x_1 (ethnic group)	x_2 (wage)	x_3 (education degree)	y (repayment)
0	2.63	-1.59	0
0	2.03	7.93	0
1	4.28	-1.88	1
...	...	...	...

The head of the loan department would like to develop an automatic tool using these data to decide whether new customer asking for a loan should be given or refused the loan; the criterion would be to offer a loan only to customers who are likely to repay the loan promptly.

A machine learning team in your group decided to develop a tool based on logistic regression (using L1 normalization). The decision is based on linear regression:

$$\theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3,$$

where θ_i is the parameter associated with feature x_i.

The tool achieves very good result on its test test, so it is submitted to production. However, some days later you receive a notification from the ethical committee of the bank, highlighting a disturbing connection between the decision of the model and the outcome of the tool.

The ethical committee asks for your advice. You decide to introspect the tool looking at the value of the learned parameters:

$$\theta_1 = 7.06$$
$$\theta_2 = 0$$
$$\theta_3 = 0$$

What can you conclude from this analysis? Do you find anything worrying about it?

The leader of the machine learning team defends its tool saying that their machine learning tool just capture a *causal relation* between some specific features of a customer and probability of repayment.

Do you agree with the argument of th machine learning team?

The machine learning team suggests an alternative version of the same logistic model using only feature x_2 and x_3.

You analyze this second version of the tool, and discover that you obtain identical results on the test set. Moreover you observe the following values for the parameters of the tool:

$$\theta_2 = 3.19$$
$$\theta_3 = 0$$

How would you interpret these results? Would you recommend the adoption of this tool?