Reject Inference: Amplifying Bias or Achieving Fairness in Credit Scoring?

When training credit scoring models exclusively on approved applications, the creditor risks introducing bias as the model has not observed an important part of the population to which it will be applied. This is known as the reject inference problem. The absence of outcome information for rejected applicants creates a scenario in which the credit scorer can opt to either accept that bias and do nothing about it or make certain assumptions about the behaviour of the rejected applicants. This choice has sparked debate in the field, with some arguing that any potential benefit is due to pure chance, and others arguing that it is a good practice that can yield moderate performance increases.

A key problem recognised by existing work is that training on the accepted population can consolidate unreasonable biases, which has potential implications for classifier performance, especially close to the acceptance threshold. However, little consideration has been given to the fairness implications of reject inference. With regards to the latter, though, we face the potential problem that our reject population may have an unfair bias encoded into it (e.g. if protected groups are overrepresented in the rejected subpopulation), which reject inference methods might then amplify.

In this paper, using Lending Club data, we study the effect of different reject inference methods, such as extrapolation, hard cut-off augmentation, fuzzy augmentation and parcelling, on the (inferred) fairness of the final classifier. In addition, we will propose fairness-enhanced reject inference methods and study what impact they have. To do so, we will use as benchmarks standard reject inference methods, choosing not to apply reject inference, and labelling methods that rely on simple machine learning systems.

In addition, we will present some follow up work that will consist of applying said methods to an industry-provided dataset that contains a true sensitive attribute and a missing not at random reject population.