Discussion Assignment 1: How to Fix Feature Bias?

The article discusses 5 strategies to deal with the problem of feature bias. The particular article is the third in a series of three articles, with all three related to different aspects of the feature bias problem.

The first one counters the conventional idea of using the fairness metrics as means to detect bias or stereotypes in a model. Stereotypes can creep into a model due to the inclusion of a sensitive feature or obvious proxies of sensitive features. It then tries to prove the inability of fairness metrics to detect the underlying stereotyping in a model's decision. It does so by experimenting on 2 models (one with inherent bias) & shows similar fairness metrics for both models, hence the inability to detect the bias.

The second article counters another conventional idea that introducing a sensitive feature does reduce feature bias. The conventional idea is based on the fact that powerful non-linear models can capture the complex interactions between different features and hence incorporate those interactions in their decision making. Thus correct the bias. But the article counters the idea by experimenting on two models (one with all features & the second one with an extra 'female' feature). Using ALE, inspecting Trees decision rules & shapely values, it shows that adding sensitive features only produces a limited correlation (under-correction) but increasing the risk of stereotyping in a model (i.e. main-effect-like of a sensitive feature). A risk worth not taking.

The particular referred article (third in the series) mentions 5 strategies to mitigate feature bias:

1) Modifying the data: This deals with the basic idea to modify the source of the problem in itself (i.e. Data). It might sound like a difficult task in some cases. One can delete a biased feature, change the source of data, change the way data is collected, go for a re-survey, etc just to get less biased data.

2) Inclusion of sensitive features in the model: This same idea that the author explained in the 'second article'. In the second article, the author argues that the conventional idea of using sensitive data to mitigate bias usually isn't so effective & increases the risk of stereotyping in a model.

3) Using a different model: This idea deals with the applicability of different models in different use cases. The author gives the example of the experiment she carried out. In the mentioned experiment the author used two different models ('Random Forest' & 'XGBoost') to compare the results concerning the feature bias. The 'Mean Shapely value' shows that 'XGBoost' performs better in mitigating the effect of a sensitive feature in comparison to 'Random Forest' (i.e. Bias correction). Here the author gives a hypothesis that the observed results might have happened because 'XGBoost' tend to rely more upon few strong predictors in comparison to 'Random Forest' that uses a combination of strong & weak predictors. Hence 'XGBoost' is less affected with correlated features & bias. However, the risk of stereotyping in a model remains, with

‘XGBoost’ ALE plots showing significant ‘main-effect-like’ interactions for ‘female’ & ‘annual Income’ features.

4) Create an explicit interaction term: The idea is to create a new feature that captures the interaction between correlated features. The idea is to explicitly input the particular interaction into the model so that the sensitive feature does not produce a main-effect-like contribution, which would in turn induce stereotyping in the model. However, this method only produces limited improvement in the case of ‘Random Forest’ & almost no effect in the case of ‘XGBoost’.

5) Separate models for each group: This idea is about using different models for different groups to mitigate bias. However, the author points that using different models is a very complex task & can produce shocking results. In the mentioned example, the author used comparable representations of both male & female. However, in many cases, the different groups might have different representations, thus producing peculiar results on different models & groups. The author also noted that the example she used had a simple co-relation only between ‘Female’ & ‘Annual income’. However, in most real cases, the causal effect is not so clear & the co-relations are more complex. Hence, making a separate model approach more difficult.

In the end, the author points to the title of her second article ‘no free lunch for feature bias’ & summarizes that there is no perfect fix or solution for the ‘feature bias’ problem. The solution depends on the many factors, questions, use cases. But each method discussed above helps to deal with ‘feature bias’ in a particular scenario or use case. In my opinion, in sync with the ideas presented in the article, the problem of ‘feature bias’ is rooted in the underlying source data. If we can counter the data at the source only (using various methods beyond the scope of this article), we can then use the ideas discussed in the article much more effectively in comparison to using one idea alone in a silo.