Predicting house hold income using two different full model approaches

For sake of simplicity, I tried a simple model of house hold income based on variables at the ESS study; and choose to study immigration status impact on household income. I was able to confirm before starting that this is highly linked to the house-hold level of education (Wilcox's test for the whole EU study is W = 67192000, p-value = 9.588e-09), thus we know that the house-hold level of education is linked with immigration status.

We also conclude that the household income is related to the level of education. As current methods we have seen don't provide a cohesive measurement for non-discrimination, we explore what is the effect on the overall prediction outcomes only; similar method was used through different formulations (such as comparing the probabilities or by comparing the effects). As starting point, we argue that we observe highly discriminatory environment (comparing the income to immigration status with Wilcon's test leads to W = 44226000, p-value < 2.2e-16), thus arguing that there is a difference based on the number of immigrants.

Running the full model, exploring immigration status, education level, having children, age and political leaning reviled that immigration seems to increase income, and the result is statically significant. We conduct the average replace fix in the estimating model and re-computing the model based on the values.

Using this method, we were able to reduce the effect on the outcome variable, however it still remains at significant level. Thus, this fix alone was not able to reduce all discrimination, however it clearly limits the effect of it (W = 307120, p-value = 0.002833).

I also used method kotivated by Caldres & Verwer (2010), I apply two support vector machines, trained to different cases of the discriminatory variable. I can not use the original idea of Naïve Bayes, as our dependable variable has several classes, and binaryfying it would lose the specificity of the model.

We again conduct the machine learning, dividing the classification task to different support vectors. By conducting this process, we can use the predicting with the test datasets for both measurements, and based on those and the immigrant status, compute Wilcon's test to compare the means. We conclude that the method was not effective in removing discrimination (W = 25714, p-value < 2.2e-16).