Subject: Predciting Low Birth Weight Prediction with Random Forest in R
Posted by Muhammad Naeem on Thu, 22 Aug 2024 05:46:05 GMT

Hi everyone,

I'm working with PDHS(2017-2018) data from the Pakistan Demographic Health Survey and have a variable indicating low birth weight, which I've categorized as follows:

0: Normal weight (2500 grams or above)
1: Low birth weight (<2500 grams)
In my dataset, the proportions are:

Normal weight: 80%
Low birth weight: 20%
I am using a Random Forest classifier in R to predict low birth weight. However, the model is producing very few correct predictions for the low birth weight category (1).

Given this class imbalance, what strategies or adjustments can I make to improve the model's performance for the minority class? Any advice or suggestions would be greatly appreciated!

Thank you!

File Attachments
1) ml_pak.csv, downloaded 105 times

Subject: Re: Predciting Low Birth Weight Prediction with Random Forest in R
Posted by Janet-DHS on Tue, 27 Aug 2024 18:17:02 GMT

Following is a response from DHS staff member, Tom Pullum:

I suggest that you tabulate the responses (m19 in the KR file). You will see that 82% of the children were either "not weighed" or have a "don't know" code. For children with numerical values, most are heaped at multiples of 500 grams, with a substantial bump right at 2500 grams, the boundary for LBW. There is a lot of omission and measurement error. You are further losing information by collapsing an interval-level variable into a binary variable.

But even if you had no omission or measurement error, it would be very hard to predict (or fit) birthweight, especially using the variables in DHS surveys. For example, the surveys do not include good estimates of gestational age, which has a clear causal relationship with birthweight.

There is a huge literature on this topic. DHS staff cannot help. Perhaps other users can offer suggestions.