I posted this to Stackoverflow recently:

This is a great dataset because it is messy, has different input data types (and LOTS of candidate inputs), and, as you mention, has two target variables, TARGET\_B and TARGET\_D. It's a classic censored regression problem where the regression part (TARGET\_D--predicting how much a lapsed donor will give in a re-capture campaign) is "censored"--we only know how much people who gave (TARGET\_B = 1) actually did give. That may seem obvious, but think about what we are trying to do: we are really trying to measure the potential amount a lapsed donor will give so we can prioritize who to send the mail piece to. Since we don't know the potential of the non-responders (TARGET\_B = 0), we are limited by what we know and are blind to that TARGET\_B = 0 population's potential (which is obviously not $0...they gave in the past, after all!)

So folks have taken a wide variety of approaches to this data set, one of which is to build two models: a selection model (TARGET\_B) followed by an estimation model (TARGET\_D). One could then treat these as independent estimates (which they are certainly not, but let's roll with it) and multiply the two to together to get an overall $$ estimate (think of it like this: a high likelihood to respond and a high $$ amount should get a higher score than a high likelihood to respond but in a lower $$ amount, etc.)

You don't need to resample to ovoid overfitting. I've written on this topic for years and can give you references if you can't find any. Logistic Regression for example builds an identical model (within the SE of coefficient values) whether you resample (stratify) or not. Avoiding overfit is a good idea, but do that by checking the training and testing error to make sure they are consistent.

Finally, remember that your best model is one that maximizes the cumulative net revenue, not R^2 or PCC or Precision/Recall, etc. So pick a model that does that: maximizes the cumulative net revenue on held out data. Net Revenue is TARGET\_D - 0.68 (0.68 is the cost per mail piece). So think "rank ordered" methods for comparing models, like lift, gain, or best of all, ROI/Profit assessment of model accuracy.

A few final thoughts:

1) the data is surprisingly linear, so linear methods do quite well (not the best, but even the most nonlinear models are "mostly" linear

2) TARGET\_B models are always pretty bad. But that's not necessarily a bad thing. It must means that in the single mailing the PVA used to create this data set, it is hard to tell who will give/respond and who won't. The good news therefore is that if TARGET\_B is confused, the TARGET\_B = 1 population and the TARGET\_B = 0 population are similar, then the TARGET\_D model, built only on TARGET\_B = 1 donors, is also representative of the TARGET\_B = 0 population. So building a TARGET\_D only model is not, in then end, a bad idea. Smaller data, fewer models, and still does pretty well.