Stat 536 Final

Jared D. Fisher

12/8/2021

Medicaid is a government-run health insurance program whose main eligibility requirement is having a low income (threshold adjusted for different factors). In 2008, Oregon decided to expand their Medicaid offerings to about 30,000 more people, but did so at random from about 90,000 applications they received. Those randomly chosen received an invitation to join Medicaid (and thus receive health insurance), and those who were not randomly chosen did not get invited. However, not everyone who was invited ended up joining. Without an invitation, people cannot sign up for Medicaid through this expansion program, but it's possible they can become eligible through the standard route and/or other programs. (We know it's possible as the data contain some people who were not invited but ended up being enrolling in Medicaid when surveyed later).

Audience: present to me as the leader of the Medicaid expansion project, who has advanced degrees in public heath and is comfortable discussing machine learning but not the coding skills/time to do it myself (thus I've hired you to consult!). From the perspective of this leader: we're always trying to help eligible people sign up for health insurance. While being insured by Medicaid is free in terms of money/dollars, it does have a cost in time spent: doing paperwork, navigating the government offices (physically, electronically, structurally), and so forth. Thus, we think that some people may be more likely to join if they have some guidance on and assistance with the process, but this takes employee time (and thus tax dollars). Thus, we want to be able to forecast who is likely to be responsive to these efforts in the current program, and perhaps identify what types of people the current approach does not work for.

I'm looking for two models in the final product of your presentation actually: both an interpretable model and a very-accurate model. It may be the case that these are one and the same. We plan to include the very-accurate model's forecasts in our computer database so that when we have (limited) time to proactively reach out to people we can prioritize contacting those who are most likely to respond and register. Additionally, we have many social workers meeting with people for other reasons, so we would like to have an interpretable model's guidance we can use to train these professionals on what to look for when considering if this current program is right for someone.

The dataset you are provided contains the following variables. Note that we have a huge database of information, so while "more data" is not an acceptable statement of next steps, it is acceptable to request particular variables of interest!

- enrolled: 1 = the individual was enrolled in Medicaid at any future point in the longitudinal study, while 0 = never enrolled in Medicaid (may have another insurance or be uninsured).
- invited: selected in the lottery
- age: age of individual
- female: individual is female
- self.plus1: application includes self up + 1 additional person
- self.plus2: application includes self up + 2 additional people
- hispanic: individual indicated ethnicity to be Hispanic or Latino/a
- race.white: individual indicated race is white
- race.black: individual indicated race is black
- race.nwother: individual indicated race is other non-white race

- english: individual requested English-language materials
- signed.self.up: individual signed him or herself up for the lottery list
- phone: gave a phone number on lottery sign up
- pobox: gave a PO Box as an address
- first.day: signed up for lottery list on first day
- income: household income as percent of federal poverty line

To summarize, we are interested in the following research questions:

1. How accurately can forecast participation in Oregon's Medicaid program? How much accuracy do we sacrifice if use an interpretable model instead of a more-complex model?
2. What does the interpretable model tell us? In other words, what "rules of thumb" do we know about who is likely to participate in this current program?
3. Who does this program not work well for? i.e. who is not joining even when invited, and likely not insured otherwise?

For this final exam/project:

- You must work on your own
- I will help debug coding errors only (your code must throw an error)
- Schedule a time during finals week to come present your final
- Give an answer to each of the research questions
- Follow the course case study rubric, with the exception: the two models criteria is for the "very-accurate" model exploration. The "interpretable" model only counts as one of your "two models" if it's is reasonably accurate. In other words, your in-sample and out-of-sample CV comparisons should include two "very-accurate" models and an interpretable model (which is probably 3 models, but maybe 2).