Code for America Data Science Exercise

Background

Data science at Code for America often involves identifying barriers our users encounter, hypothesizing about potential changes to reduce those barriers, and then experimenting to test those hypotheses. This data exercise is based on roughly 2,000 applications and outcomes from our GetCalFresh.org service. CalFresh is the name for SNAP (Supplemental Nutrition Assistance Program, or food stamps) in California. SNAP is funded as an entitlement program: anyone who can demonstrate that they meet the requirements should receive benefits. Income eligibility for SNAP is based on the Federal poverty line, which depends on household size. There are no adjustments made for cost of living.

To be approved, applicants must also complete a phone or in-person interview and submit documents to verify their identity and circumstances. The data in this exercise reflect what we collect through GetCalFresh.org - it does not reflect all the information that a case worker uses to make a decision. For example, simply because a person says they didn't have the interview when we text them doesn't mean that they didn't do it later. Similarly, applicants can submit documents to their county in ways that we do not capture in our database (say, dropping them off or mailing them).

All of the data for the exercise are from San Diego County, which decides as a county how to administer its program.

Download the data here

Instructions

The primary task in this exercise is to **tell a story about what factors are positively and negatively associated with CalFresh approval**. Our goal is to learn about your approach to addressing quantitative research questions. It's okay to speculate or go out on a limb - we aren't going to hold you to any of your numbers or conclusions :-).

- You can provide your answers in any format you'd like (a Google Doc is fine).
- Please include your code for the data exercise at the end of the document you submit.
- In addition to explaining what you found, please include at least one table showing your main results (in response to Question 1).
- Figures or graphs are not required.
- You have up to 8 hours to submit a response, but our goal is for this to only take 4-5 hours. The data exercise does not include any intentional anomalies, data quality problems, or gotchas - so please don't spend time worrying about these kinds of issues.

Questions

1. What factors are most strongly associated with CalFresh approval? (2-4 paragraphs, briefly mention the method(s) you used, any variable selection / model refinement you did, and include an output table).

2. Based on your results, where would you look next in terms of identifying potential improvements to the application? (this is just an exercise and guessing is fine - 1-2 paragraphs max)

3. The *had_interview* variable has missing values for all the individuals who did not respond to our SMS survey. What can we learn from this variable? (1-2 paragraphs max)

4. Let's assume that San Diego administers its program consistently throughout the county. In your model, do you find zip codes to be informative with respect to approval? If so, what might it mean? (1-2 paragraphs max)

Table 1: Variables in the Data Set

Variable	Type	Range	Source	Definition
income	Continuous	0 - 9999	Application	Total household income in the last 30 days (with a small random value added)
household size	Count	1 - 12	Application	Number of people who are applying for benefits
had_interview	Logical	TRUE, FALSE [note: NAs are common]	Text message survey	Applicant's response to an SMS question from GetCalFresh – TRUE indicates that they had the interview.
docs_with_app	Count	0 – 25	Application	Number of verification documents uploaded with the GetCalFresh.org application
docs_after_app	Count	0 – 29	Later Docs (a GetCalFresh website)	Number of verification documents uploaded after the application through GetCalFresh Later Docs (often, after the interview)
stable_housing	Logical	TRUE, FALSE	Application	TRUE if the applicant rents or owns the place where they sleep
under_18_n	Count	0-10	Application	Number of children age 17 and younger on the application
over_59_n	Count	0-10	Application	Number of adults age 60 and above on the application
zip	Categorical	5 digit zip code	Application	Zip for address where the person lives / stays

completion_time_mins	Continuous	0 - 8552	Application metadata	Minutes elapsed between the start and end of the application
approved	Logical	TRUE, FALSE	County-provided	Outcomes exported by the county, TRUE means approved for CalFresh