Direct-Mail Fundraising

fundraising.rds and *future_fundraising.rds* are the datasets used for this case study. You might find the read_rds function in the readr library helpful for reading these data into R.

Background

A national veterans' organization wishes to develop a predictive model to improve the cost-effectiveness of their direct marketing campaign. The organization, with its in-house database of over 13 million donors, is one of the largest direct-mail fundraisers in the United States. According to their recent mailing records, the overall response rate is 5.1%. Out of those who responded (donated), the average donation is $13.00. Each mailing, which includes a gift of personalized address labels and assortments of cards and envelopes, costs $0.68 to produce and send. Using these facts, we take a sample of this dataset to develop a classification model that can effectively capture donors so that the expected net profit is maximized. Weighted sampling was used, under-representing the non-responders so that the sample has equal numbers of donors and non-donors.

Data

The file *Fundraising.csv* contains 3,000 records with approximately 50% donors (`target = Donor`) and 50% non-donors (`target = No Donor`). The descriptions for the 22 are listed below.

Variable	Description
zip	Zip code group (zip codes were grouped into five groups; Yes = the potential donor belongs to this zip group.)
	00000–19999 ⇒ zipconvert1
	20000–39999 ⇒ zipconvert2
	40000–59999 ⇒ zipconvert3
	60000–79999 ⇒ zipconvert4
	80000–99999 ⇒ zipconvert5
homeowner	Yes = homeowner, No = not a homeowner
num_child	Number of children
income	Household income
female	No = male, Yes = female
wealth	Wealth rating uses median family income and population statistics from each area to index relative wealth within each state. The segments are denoted 0 to 9, with 9 being the highest-wealth group and zero the lowest. Each rating has a different meaning within each state
home_value	Average home value in potential donor's neighborhood in hundreds of dollars
med_fam_inc	Median family income in potential donor's neighborhood in hundreds of dollars
avg_fam_inc	Average family income in potential donor's neighborhood in hundreds
pct_lt15k	Percent earning less than $15K in potential donor's neighborhood
num_prom	Lifetime number of promotions received to date
lifetime_gifts	Dollar amount of lifetime gifts to date
largest_gift	Dollar amount of largest gift to date
last_gift	Dollar amount of most recent gift
months_since_donate	Number of months from last donation to July 2018
time_lag	Number of months between first and second gift
avg_gift	Average dollar amount of gifts to date
target	Outcome variable: binary indicator for response Yes = donor, No = non-donor

Direct-Mail Fundraising

Assignment

Step 1: **Partitioning**. You might think about how to estimate the out of sample error. Either partition the dataset into 80% training and 20% validation or use cross validation (set the seed to 12345).

Step 2: **Model Building**. Follow the following steps to build, evaluate, and choose a model.

1. *Exploratory data analysis.* Examine the predictors and evaluate their association with the response variable. Which might be good candidate predictors? Are any collinear with each other?
2. *Select classification tool and parameters*. Run at least two classification models of your choosing. Describe the two models that you chose, with sufficient detail (method, parameters, variables, etc.) so that it can be reproduced.
3. *Classification under asymmetric response and cost*. Comment on the reasoning behind using weighted sampling to produce a training set with equal numbers of donors and non-donors? Why not use a simple random sample from the original dataset?
4. *Evaluate the fit*. Examine the out of sample error for your models. Use tables or graphs to display your results. Is there a model that dominates?
5. *Select best model*. From your answer in (4), what do you think is the "best" model?

Step 3: **Testing**. The file *FutureFundraising.csv* contains the attributes for future mailing candidates.

6. Using your "best" model from Step 2 (number 4), which of these candidates do you predict as donors and non-donors? Use your best model and predict whether the candidate will be a donor or not. Upload your prediction to the leaderboard and comment on the result.
7. *Submission File.* For each row in the test set, you must predict whether or not the candidate is a donor or not. The .csv file should contain a header and have the following format:

```
value
Donor
Donor
No Donor
Donor
No Donor
No Donor
. . .
etc.
```

You might find the write_csv function helpful

Direct-Mail Fundraising

Document Methodology and Models

Document the entire process and findings for audit requirements and future work. Ideally, you should document the following:

- *Business Objectives and Goals*. Successful businesses are based on both goals and objectives, as they clarify the purpose of the business and help identify necessary actions. Goals are general statements of desired achievement, while objectives are the specific steps or actions you take to reach your goal.
- *Data Sources and Data used.* Readers need to know how the data was obtained because the method you choose affects the results and, by extension, how you likely interpreted those results. Here's where you include your discussion on the reasoning behind using weighted sampling to produce a training set with equal numbers of donors and non-donors? Why not use a simple random sample from the original dataset?
- *Type of Analysis performed: what, why, findings*. Methodology is crucial for any branch of scholarship because an unreliable method produces unreliable results and it misappropriates interpretations of findings. In most cases, there are a variety of different methods you can choose to investigate a research problem. This section should make clear the reasons why you chose a particular method or procedure.
- *Exclusions*. Excluding a class from prediction or an observation is done if its precision or coverage statistics don't meet your threshold of usefulness. For example, exclude it if you don't want the model to predict a particular output field value. Train the solution definition whose output field values you want to exclude.
- *Variable transformations.* Variable transformation is a way to make the data work better in your model. Data variables can have two types of form: numeric variable and categorical variable, and their transformation should have different approaches. If you made a transformation, talk about it. If you created a new variable, talk about it.
- *Business inputs.* Resources such as people, raw materials, energy, information, or finance that are put into a system (such as an economy, manufacturing plant, computer system) to obtain a desired output.
- *Methodology used, background, benefits.* Research methodology is the specific procedures or techniques used to identify, select, process, and analyze information about a topic. In a research paper, the methodology section allows the reader to critically evaluate a study's overall validity and reliability. Discuss any alternative methodologies you tried.
- *Model performance and Validation Results.* Evaluating a model is a core part of building an effective machine learning model. There are several evaluation metrics, like confusion matrix, cross-validation, AUC-ROC curve, etc. Evaluation metrics explain the performance of a model. An important aspect of evaluation metrics is their capability to discriminate among model results.
- *Cut-Off Analysis*. While the ROC curve and corresponding AUC give an overall picture of the behavior of a diagnostic test across all cutoff values, there remains a practical need to determine the specific cutoff value that should be used for individuals requiring labeling. The optimal cutoff value is the one that minimizes cost.
- *Recommendations.* Example of recommendations can be defined as a critical suggestion regarding the best course of action in a certain situation. The whole idea of a recommendation is to provide a beneficial guide that will not only resolve certain issues, but result in a beneficial outcome. What percentage of your data would you recommend for a mailing campaign?
- *Pseudo codes for implementation.* Give me your R code.