HW4 Proposal

RJ Cass

1. Understanding the Problem

Background:

We have a dataset of various client data from previous marketing campaigns, and the result of those efforts (whether or not the client opened a new account). We want to use this dataset to help our client (the bank) understand what is most useful in making their marketing efforts more effective.

Goals:

In this analysis, we want to address several key questions:

i. What characteristics of customers are more likely to take out a new credit card?

- To answer this we will identify which of the explanatory variables is significant in explaining the likelihood of a client taking out a new credit card. Of those, we will look at which combination of elements gives the highest likelihood of opening a new account.

ii. Is there evidence that social media vs. personal contact is more effective in marketing?

- To answer this we will identify a) whether both of these explanatory factors appear to be significant and b) which of them appears to have a larger impact (if any) on the likelihood of opening a new account.

iii. Does repeated contacting seem to increase the likelihood of a person opening an account?

- To answer this we will examine whether repeated contacting appears to have a significant effect on a client opening a new account. If so, we can look at the sign of the affect (positive or negative) to see if it increases or decreases the likleihood of a person opening an account.

2. Exploratory Data Analysis

The first thing to consider in this dataset is that our output variable is binary (yes/no). As such, we will need to create our model based on the likelihood of the output. We also need to consider if all the continuous variables are monotonic in relation to the likelihood of the outcome. Initial exploration indicates this might not be the case for some of the variables, so we will need to explore this further. We also determined that not all categorical variables have an even count/distribution of samples for each value.

3. Desired Attributes

Model attributes required from the research questions

The research questions require that the model be able to indicate which combination of factors yields the highest likelihood of opening an account. They also require that, if possible, we be able to identify which of 2 factors is more impactful in determining the likelihood of opening a new account. They also require that we be able to identify whether a given factor has a positive or negative effect on the likelihood.

Model attributes required from the data

The data require that our model account for a binary outcome variable.

Any other anticipated problems

In our initial investigation we identified that there may be issues with our monotinicity assumption for our numerical explanatory factors. We will also need to account for uneven distribution of sampling among factor levels.

What goes wrong if the above are not accounted for?

If all of our assumptions are not met, the resultant model will not correctly account for each factor, meaning the significance of each factor will be incorrect and their impact on the likelihood may be wrong as well. It will also lack predictive power for values outside our given dataset.

4. Proposed Method

Appropriate models

For this analysis, a binary model will be appropriate, where we create a model relating the explanatory factors to the log-likelihood of a client opening an account.

Specific model proposal

We propose using a model of the form:

$$\ell(p_i) = \log(\frac{p_i}{1-p_i}) = \mathbf{x}_i'\beta$$

where $\mathbf{x}_i'$ is the matrix representing the explanatory variables, and β is the matrix representing the coefficients for each factor.

Method strengths/weaknesses

i. How this method accounts for issues in the dataset

This method allow us to build a model of the log-likelihood of an event occuring given the provided factors. It enforces a value between 0 and 1 which is necessary to get a value for p.

ii. How this method accomplishes research/analysis goals and yields appropriate estimators

This model will allow us to determine which factor levels will maximize the likelihood. It will also allow us to compare relative effects of different factors to determine relative importance. We can also examine the signs of the coefficients for each variable to see if they increase or decrease the likelihood of a client opening an account.

iii. How this method will answer the research questions

We can view the p-values of each factor in the model to see if it is significant, and what value is needed to maximize the likelihood. We can compare p-values across variables. We can view the signs (negative vs. positive) of each coefficient to see if the increase or decrease the likelihood of a client opening an account.

iv. What assumptions are needed to use the model adequately? Are they reasonable to assume and explained well?

This model assumes that each event is indepenent. It also assumes that each factor is monotonic (always increasing or decreasing) in relation to the likelihood of a client opening an account. If these assumptions are not met, the model will not accurately capture the effect of each factor and the significance of each variable will be wrong.