536 Homework 4 Proposal

Sam Lee, Gavin Hatch

October 29, 2024

1 Introduction

Companies are interested in knowing the best marketing strategies for selling their products to individuals. In this report, we help a bank find a way to determine which customers are interested in their credit card so they can better target their ads to interested individuals. We will do this by finding out what characteristics of customers are more likely to take out a new credit card, finding whether social media or personal contact is more effective in marketing, and finding out how repeated contacting affects the likelihood of a person taking out an account.

1.1 Data

We were provided a dataset with 15 columns and over 40,000 entries. Our goal is to predict whether a client has opened a new account. We describe the data in the Appendix.

We first acknowledge the disparity between the amount of people who opened up new accounts and people who did not (See Figure 1), where only about 11.27% of individuals within the dataset opened up an account. Assuming our data are independent (i.e. randomly sampled), we will account for this using the F_1 score to evaluate our models[1]. Doing so will allow us to adjust for the skewness in the response variable while also allowing us to adequately evaluate our models.

The data contains several categorical variables. This will prevent us from using standard regression for the analysis. There seems to be some association between **pdays**, **previous**, and **poutcome** as well as between **housing** and **loan**.

As seen in Figure 2, for the *pdays* variable, most of the values are censored at 999 (or no previous contact). This is something to keep in mind when interpreting that variable in the data. Hence, we model the data in a piecewise logistic

[1]The F_1 score can be calculated as, $\frac{2tp}{2tp+fp+fn}$, where tp, fp, and fn are the number of true positives, false positives, and false negatives our estimated model predicts, respectively.

Figure 1: disparity in the response variable: The majority of the customers in the data set chose not to open up an account.

regression process[2], splitting up the censored and non-censored *pday* data:

$$\begin{aligned} &\text{Let } contacted_i = \mathbb{I}\left(pdays_i < 999\right) \\ \Pr\left(y_i = \text{“yes”}\right) = ln\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \gamma_1 pdays_i \times contacted_i + \\ &\gamma_2 \mathbb{I}\left(1 - contacted_i\right) + x_i'\beta + \varepsilon_i \end{aligned} \tag{1}$$

2 Proposed Methods

- Model 1 - Probit Regression
  - Using a Probit regression model will help us deal with binary classification as well as probability estimation. It is also easy to interpret! One weakness of this model is the fact that when converting the categorical variables to dummy variables, there will be a lot to interpret. Other than that, this model will help us answer the questions by providing a significance for each variable in the data.
  - Assumptions
    - Independence: There is nothing in the prompt suggesting that the observations in the data are not independent, so this assumption is met.
    - Monotonicity in X's: There is a linear relationship between each of our explanatory variables and the probit link.
    - General Equation: Probit differs from the logistic regression model in the link function. Rather than using the logit link function, it uses the probit link function which is based off of

[2]Note that the regression coefficient vector, β, consists of all the coefficients for the covariates described in the data section, with indicator functions, $\mathbb{I}(\cdot)$, used for each level of the categorical variables. Interestingly, a preliminary estimation of Equation 3 zeros-out γ_1 but keeps the coefficient that estimates an intercept-like coefficient on the censored data, that is, an estimate for γ_2.

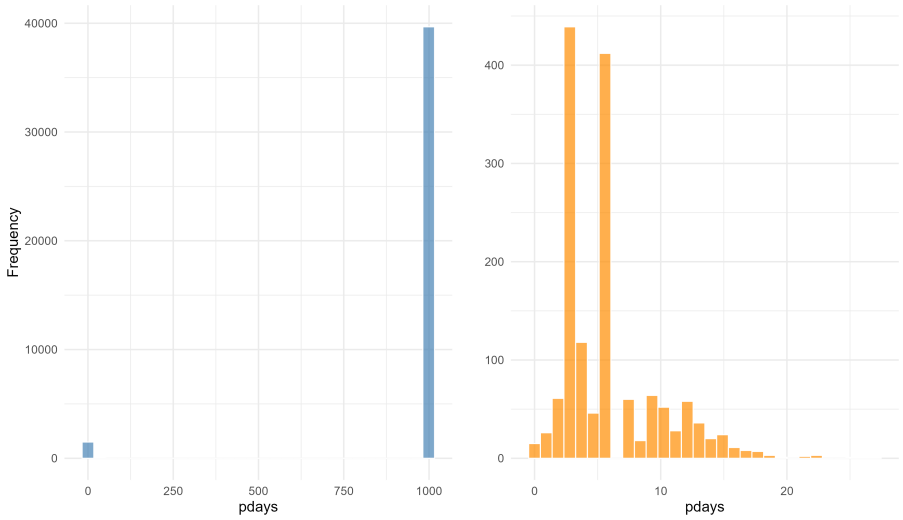

Figure 2: Distribution of *pdays*—dissected by the censored (left) and non-censored (right) components

the Cumulative Distribution Function for the standard normal distribution.

$$\Phi^{-1}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{2}$$

- Model 2 - LASSO Logistic Regression[3]

  We propose the following logistic regression model to predict outcomes for Y. To reduce dimensionality, given the large number of covariates induced by the categorical factors, we combine LASSO regression with the general logistic regression model to achieve a more parsimonious fit. The general equation is outlined in Equation 3.

$$\hat{\beta} = \arg\min_{b} \left(-\sum_{i=1}^{n} \left[y_i \, log f\left(X_i; b\right) + \left(1 - y_i\right) log f\left(X_i; b\right) \right] + \lambda \sum_{j=1}^{p} |\beta_j| \right). \tag{3}$$

  We operate under the assumption of linearity: There is a linear relationship between each of our explanatory variables the log-odds and similar

[3]By way of logistic regression, we let the link function, $f\left(X_i, b\right)$ be defined as, $f\left(X_i, b\right) = \frac{1}{1+\exp\left(-\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)\right)}$.

to Model 1, we assume independence between observations[4]. The LASSO regression framework also allows us to estimate confidence intervals on the *most relevant factors* influencing whether a costumer will open a new account through bootstrap estimation.

[4]In other words, we assume that $Pr(Y|X) = \prod_{i=1}^{n} Pr\left(y_i = 1|X_i\right)$.

3 Appendix

3.1 Data

Here we provide an in depth description of the covariates used in this analysis and the data types these variables take on in the data set.

- **RESPONSE VARIABLE**: **y** - has the client opened a new account? (binary: "yes","no")
- **age** (numeric)
- **job**: type of job (categorical: "admin.", "blue-collar", "entrepreneur", "housemaid", "management", "retired", "self-employed", "services", "student", "technician", "unemployed", "unknown")
- **marital**: status (categorical: "divorced","married","single","unknown"; note: "divorced" means divorced or widowed)
- **education** (categorical: "basic.4y", "basic.6y", "basic.9y", "high.school", "illiterate", "professional.course", "university.degree", "unknown")
- **default**: has credit in default? (categorical: "no","yes","unknown")
- **housing**: has housing loan? (categorical: "no","yes","unknown")
- **loan**: has personal loan? (categorical: "no","yes","unknown")
- **contact**: contact communication type (categorical: "social","direct")
- **month**: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
- **day of week**: last contact day of the week (categorical: "mon", "tue", "wed", "thu", "fri")
- **campaign**: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- **pdays**: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
- **previous**: number of contacts performed before this campaign and for this client (numeric)
- **poutcome**: outcome of the previous marketing campaign (categorical: "failure","nonexistent","success")