Targeted Marketing Proposal

Camilla McKinnon and Patric Platts

2024-10-28

Introduction

An effective marketing campaign to the right audience is important for companies trying to extend their reach or grow a product. This analysis looks at a specific bank's campaign. The bank uses a personalized marketing method to deliver individualized products to recipients. If the bank can better understand their customers, they can better target future marketing campaigns. This data set contains several variables on client characteristics (like age, job, education, loan status), method of contact, and when the contact happened. The response variable is whether or not the client opened a new bank account. One of the issues with the dataset is that the response rate is a binary variable (measured in 'yes' and 'no'). Traditional standard regression won't work since it would predict continuous values that would fall outside the 0-1 binary classification range. Additionally, many of our variables are categorical, which means they will need to be treated as factors in our model. Some of the factor levels have 'rare occurrences', meaning there are not many observations. Not accounting for that would lead to increased variance and unstable estimates. The following plot shows one of the variables, age, plotted against whether or not the client opened a bank account.

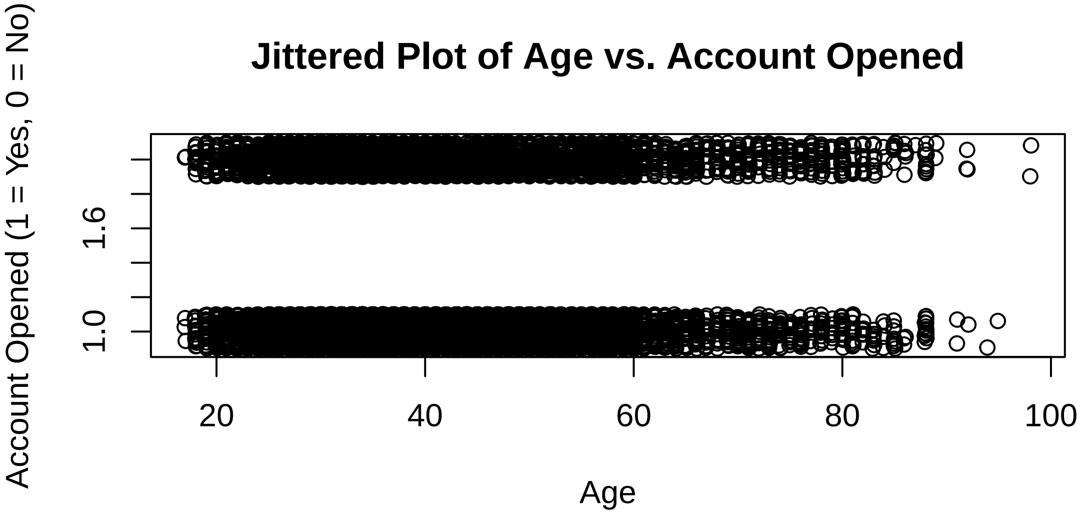

Figure 1: Scatterplot of Age and Account Opening Status with Jittering Applied. Age is plotted along the x-axis, while the binary response variable (Account Opened: 1 = Yes, 0 = No) is shown on the y-axis, with jittering added to reduce overplotting and improve visibility.

After creating a sufficient model, this analysis will explore the following questions: - What characteristics of customers are more likely to take out a new credit card? - Is there evidence that social media vs. personal contact is more effective in marketing? - Does repeated contacting seem to increase the likelihood of a person taking out an account?

Proposed Methods 1 & 2

OLS

One of the models considered for this analysis was ordinary least squares (OLS) logistic regression. OLS is great for classification scenarios. With an OLS logistic regression model, we can look at coefficients and odds ratios to answer the research questions. A positive coefficient for a characteristic means that as that factor increases, or for a certain level, the likelihood of opening an account increases. Additionally, we can compare odds ratios of social media and direct contact to see which is more effective. The "both ways" variable selection method was used to trim down and improve the performance of the model. Assumptions required for this model are that there is independence between observations, linearity, normality and equal variance of residuals.

Lasso

The second model we consider is a logistic regression with LASSO regularization. Logistic regression is appropriate due to the binary nature of the response variable, while LASSO helps manage the large number of predictor variables, many of which have multiple factor levels. Including these factor levels substantially increases the number of predictors in the model, and adding interactions further expands this count to potentially astronomical levels, which could lead to overfitting. LASSO regularization addresses this by shrinking less informative coefficients to zero, effectively performing feature selection and helping prevent overfitting. LASSO with logistic regression helps identify the characteristics and factors associated with the likelihood of opening a new account by selecting only the most relevant predictors. In using LASSO, important assumptions to consider include independence of observations, normality of predictor distributions, and equal variance of residuals to ensure reliable and interpretable results.

Model Evaluation

Each model was evaluated in-sample and out-of-sample using the Area Under the Curve (AUC) metric to assess classification performance. An AUC of 0.5 indicates that the model performs no better than random guessing, while an AUC greater than 0.5 shows improvement over random predictions. An AUC of 1.0 represents perfect classification across all thresholds, so a higher AUC reflects better performance. Out-of-sample AUC was calculated using 10-fold cross-validation, allowing each model to be trained 10 times to provide more reliable performance estimates. Table 1 presents the in-sample and out-of-sample AUC values for each model.

Table 1: Comparison of In-Sample and Out-of-Sample AUC for Basic and LASSO Logistic Regression Models

Model	In.Sample.AUC	Out.of.Sample.AUC
OLS Model	0.7589	0.7553
LASSO Model	0.7677	0.7600

In Table 1, the AUC values for each model appear comparable, with the OLS model at 0.7553 and the LASSO model at 0.7600, showing only a minor difference. Given that both models performed similarly, the OLS model was selected for its simplicity and interpretability over the LASSO model.

Logistic Regression Model:

$$Y_i \overset{ind}{\sim} \text{Bern}(p_i) \qquad \log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i'\boldsymbol{\beta}$$

Table 2: Variables in the Logistic Regression Model (Vector x'_i)

Variable	Description
age	Age of the individual
job	Job type of the individual
marital	Marital status of the individual
education	Education level of the individual
default	Whether the individual has credit in default
contact	Type of communication contact
month	Month of last contact
day_of_week	Day of the week of last contact
campaign	Number of contacts performed during this campaign
cat_pdays	Categorical version of pdays (days since last contact)
pdays	Number of days since the individual was last contacted
previous	Number of contacts before this campaign
poutcome	Outcome of the previous marketing campaign
cat_pdays:previous	Interaction between categorical pdays and previous contacts
job:education	Interaction between job type and education level

A logistic regression model relies on several assumptions, including independence of observations, a monotonic relationship with predictor variables, a binary outcome, and no multicollinearity among predictors. In this analysis, the outcome variable—whether a client has opened a new account—is binary (yes or no). Based on the data collection method, we assume independence between observations. Exploratory analysis also suggests that the predictor variables have monotonic relationships with the outcome, indicating no apparent non-linear relationships.

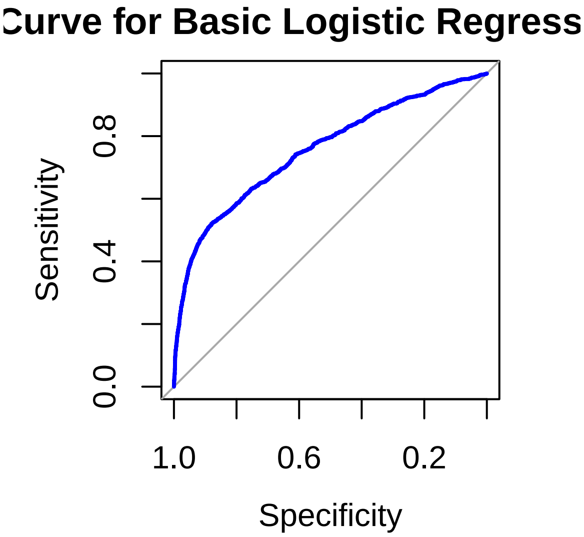

Figure 2: ROC Curve illustrating the model's predictive performance compared to random guessing, with AUC demonstrating the model's accuracy.