Targeted Marketing Proposal

Camilla McKinnon and Patric Platts

2024-10-28

Abstract

This analysis examines the bank's marketing campaign aimed at encouraging customers to open new credit card accounts. Using logistic regression, the model predicts the likelihood of a positive response based on customer demographics, contact methods, and campaign history. Results indicate that job type, education, and marital status are significant predictors, with students and social media contacts most likely to open accounts. Additionally, repeated contact across campaigns slightly increases the likelihood of a response, guiding future marketing strategies.

Introduction

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 challenges 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, as an example only three individuals responded 'yes' to having credit in default, which is much less than those in the other categories. Not accounting for that would lead to increased variance and unstable estimates. The Figure 1 shows one of the variables, age, plotted against whether or not the client opened a bank account.

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.

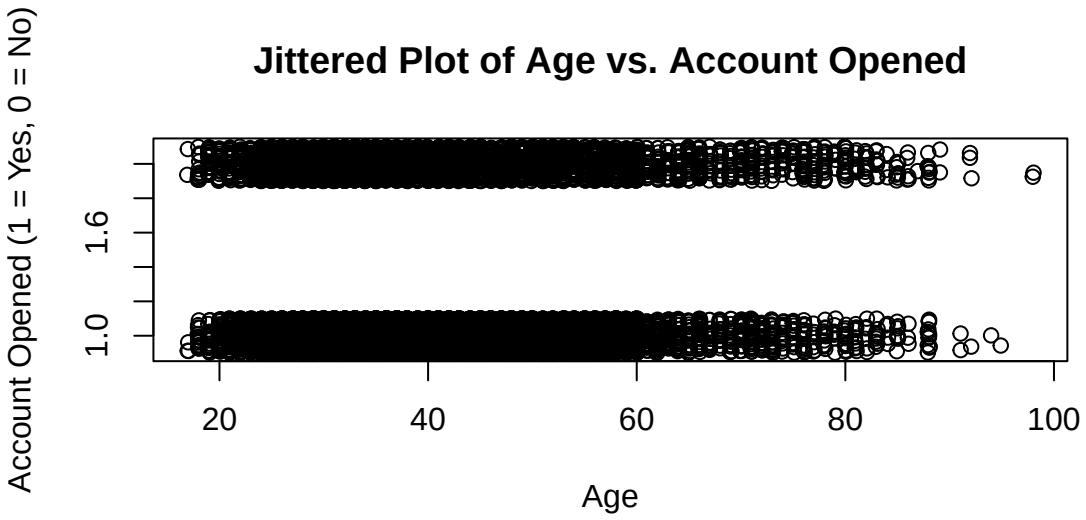

Figure 1: Scatterplot of Age and Account Opening Status with Jittering Applied. Age is plotted along the x-axis, while the binary response variable (Account Opened: 1 = Yes, 0 = No) is shown on the y-axis, with jittering added to reduce overplotting and improve visibility.

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

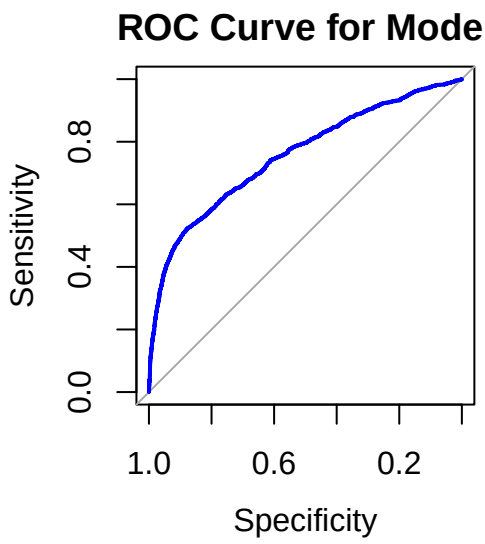

Figure 2: ROC Curve illustrating the model's predictive performance compared to random guessing.

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. Figure 2 visualizes the sensitivity versus specificity of the model across various thresholds, illustrating its predictive accuracy at correctly identifying outcomes.

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 (X matrix)

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

To address issues with rare events in the data, certain factors were combined into broader categories to improve model performance. For example, the 'illiterate' category in education was merged with 'basic.4yr' to form a single category, 'basic.4yr,' encompassing individuals with four years of education or less. This adjustment was made based on the similarity between basic four-year education and illiterate categories. A bidirectional selection process was applied to the OLS logistic model to identify predictors significant in determining the outcome. This process removed 'housing' and 'loan' variables, as they did not contribute to improving prediction accuracy.

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. For example see Figure 3 for the monotonic relationship between age versus account opened.

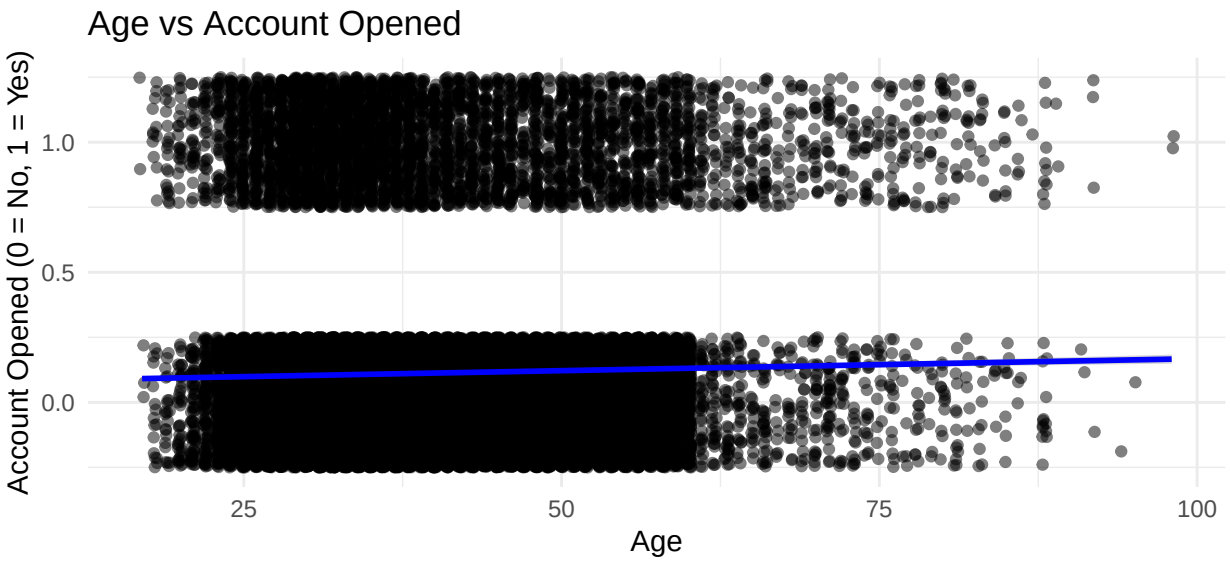

Figure 3: Monotonic relationship between age and account being opened.

Results

Table 3: Odds Estimates and 95% Confidence Intervals for Selected Variables

Term	Estimate	Lower 95% CI	Upper 95% CI
age	1.005355	1.0004407	1.0102936
jobstudent	2.080028	1.4796734	2.9239682
maritalsingle	1.148625	1.0366697	1.2726703
educationprofessional.course	1.109569	0.7301360	1.6861829
educationuniversity.degree	1.125703	0.9486087	1.3358600
contactsocialMedia	2.597810	2.3028218	2.9305863
campaign	0.923279	0.9031872	0.9438177
previous	1.116422	0.8606437	1.4482154

The odds coefficients reveal characteristics associated with a higher likelihood of applying for a credit card. Variable selection indicates that 'housing' and 'loan' status are not significant predictors of application outcomes, narrowing the focus to other influential factors.

Among job categories, students exhibit the highest likelihood of opening a new credit card, particularly compared to admins and managers. Education level also impacts the likelihood: individuals with less than a high school education are less likely to apply, while those with professional courses are 1.39 times more likely than high school graduates. In contrast, those with a university degree are 0.857 times less likely than high school graduates to open a card.

Marital status further affects the odds, with divorced individuals are less likely to apply than married ones, while single individuals have 1.103 times higher odds. Additionally, each additional year of age slightly increases the likelihood of applying for a credit card by a factor of 1.005.

The bank is also examining the effectiveness of social media versus personal contact in marketing. The model indicates that the odds of opening a credit card are 2.544 times higher for individuals contacted through social media compared to those approached directly.

As number of contacts occurring during a campaign increases for a specific customer the likelihood of them opening a credit card decreases. However, if contacts were performed in a previous campaign, the odds of them opening a card increase by 1.017, suggesting a slight potential increased interest in opening cards if it's the second campaign they've been in.

Conclusion

This analysis looked at the bank's credit card marketing campaign. To identify important factors and trends, we built a logistic regression model. The model indicated that customers who are students, have a high school education, or who are single are most likely to apply for a credit card. Customers with who obtained a university degree afterwards become less likely to open a card, though customers who went on to obtain a professional course are more likely to open a card. Social media contacting was more effective than direct contacting. Also if someone was contacted in a previous campaign, and then again in the current one, they were more likely to open a card as well. The potential exists that our model did not account for some correlation between the predictor variables. However, since the proposed model that would have accounted for that performed comparably, we went with the standard logistic regression. Further analysis could explore additional factors affecting credit card openings, such as income, credit score, and customer rated service quality.

Teamwork

Each of us looked at one of the proposed models, looking at the assumptions and various selection methods associated. We then split up the writing into equal parts, coming together for the final draft of the project.