Unit 4 Report

RJ Cass and Jarom Asher

October 2025

Abstract

Successful marketing campaigns are those that target individuals most likely to respond positively and purchase/engage with the product. We wanted to figure out what type of people tend to open new credit cards through this bank. Using a logit model, we found that older, single students/retired people who we have contacted before are the most likely to sign up for this credit card. We also identified that Social Media yields a higher likelihood of clients opening a new account, as does increased previous contacts.

1 Introduction

Marketing campaigns are strategic, organized efforts designed to promote a specific company goal or product. In today's digital landscape, marketing can reach consumers through a variety of channels, including social media, mobile apps, and text messaging. However, the effectiveness of these campaigns ultimately depends on their ability to connect with individuals who are genuinely interested in the offering.

In an effort to improve our bank's future credit card marketing strategies, we aim to answer the following questions:

1. What characteristics of customers are more likely to take out a new credit card?
2. Is there evidence that social media vs. personal contact is more effective in marketing?
3. Does repeated contacting seem to increase the likelihood of a person taking out an account?

Exploratory data analysis highlights several important considerations. First, our outcome variable is binary (whether or not a customer opened a credit card account), which means a simple linear regression model is inappropriate; we will instead use a classification model suited for binary data. Second, the dataset exhibits class imbalance—there are many more 'no' responses than 'yes' responses (Figure 1). This imbalance can potentially bias predictions and inflate performance metrics (e.g., high accuracy simply by predicting "no" for all cases). We will address this issue carefully, applying appropriate resampling or weighting techniques if needed to ensure a fair and effective model.

2 Methodology

2.1 Model 1: Logit

The first model considered is a logit model where the variables were selected using hybrid stepwise selection (both forward and backward). Basically, we fit a model using all the

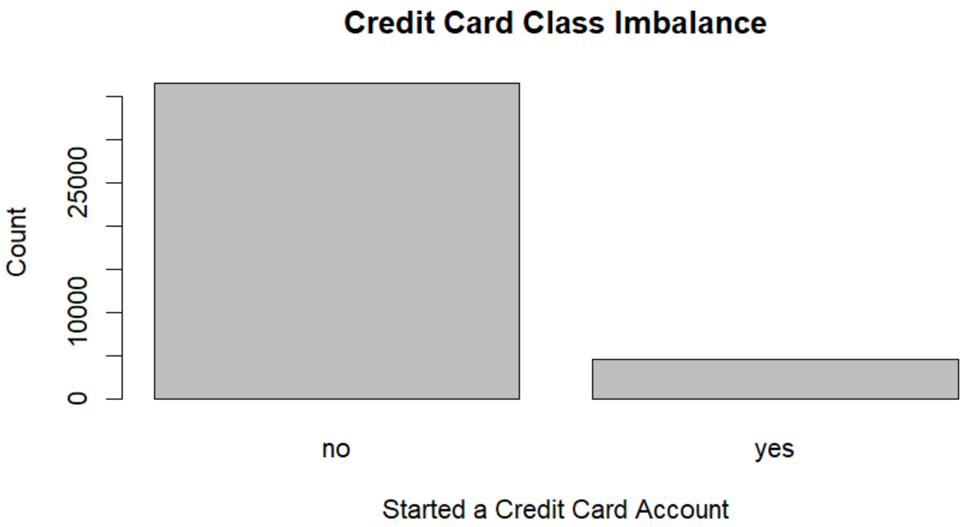

Figure 1: Bar plot showing class imbalance in the target variable.

possible variables, and removed/added variables until the model was the most efficient. Using this method we identified that the 'Housing' and 'Loan' variables are not significant in explaining the likelihood of opening an account. We also found that 39,673 of 41,188 (96%) of people had **not** previously been contacted. This creates an imbalance as explained previously, so we decided to remove 'pdays' from the model.

The equation for this model is as follows:

$$Y_i = ln(\frac{p}{1-p}) = \boldsymbol{\beta x}' + \epsilon, \epsilon \sim N(0, \sigma^2)$$

where β is the matrix of coefficients, and x' is the matrix of variables (note that each level of each factor, such as month = October, has its own coefficient).

2.2 Model 2: Probit

Our second model is a probit model. Like the logit model, it is designed specifically for classifying binary outcomes, just with a different link function. Backward selection similarly suggested removing just the 'Housing' and 'Loan' variables, and we also decided to remove pdays from this model because it adds very little to our model.

The equation for this model is as follows:

$$\begin{aligned} &Y_i \overset{\text{ind}}{\sim} \text{Bern}(p_i), \\ &p_i = \Phi(x'\boldsymbol{\beta}), \\ &\Phi: \text{ standard normal CDF.} \end{aligned}$$

where β is the matrix of coefficients, and x' is the matrix of variables.

Table 1: Model AUC Comparison

Model	AUC
1	.7634
2	.7643

2.3 Model Selection

To compare the two models we examined the Area Under the Curve (AUC) of each graph: basically, a representation of how well each model explains the variability of the data, as shown in Table 1.

Both models perform almost equally in capturing the behavior of the data. As such, we decided to use the logit model to answer the research questions due to being more interpretable. The calculated coefficients for each factor/level are provided in the appendix.

A logit model has a few assumptions that need to be met. In the cases of our categorical variables, we really only need to meet the assumption that they are independent, which we are assuming to be the case. In practice, this may not be the case as it's likely that some variables such as age and education, or job, are correlated. For our continuous variables, we need to confirm monotonicity. Viewing Figure 2, we can see that with the exception of age, each of the continuous variables appears to be monotonic with the log-odds, so our assumptions are met. In the case of age, as identified later, both older clients and students have a higher likelihood of opening an account, so this assumption is not adequately met, which may impact the predictive powers of our model.

As an example of how to interpret the coefficients, using the variable contact:SocialMedia - a coefficient of .9431 means that, holding everything else constant, the odds of a client opening a new account when contacted via Social Media as opposed to Personal Contact are $e^{.9431} = 2.57$. In other words, making contact with a client via Social Media vs. Personal Contact results in a 157% increase in the odds of them opening an account.

3 Results

We can use the model selected above to answer our research questions.

3.1 Customer Characteristics

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

Using the provided coefficients in Table 2 we can identify which factors increase the likelihood of a client opening a new account. For each variable, we identify the most impactful category: **Age**: older increases likelihood. **Day of Week**: Wednesday. **Default**: No. **Education**: Illiterate (but also less education in general increases likelihood). **Job**: Retired or students. **Marital Status**: Single (or unknown). **Month**: March. **Previous Marketing Outcome**: Success. **Previous**: More previous contacts increases the likelihood.

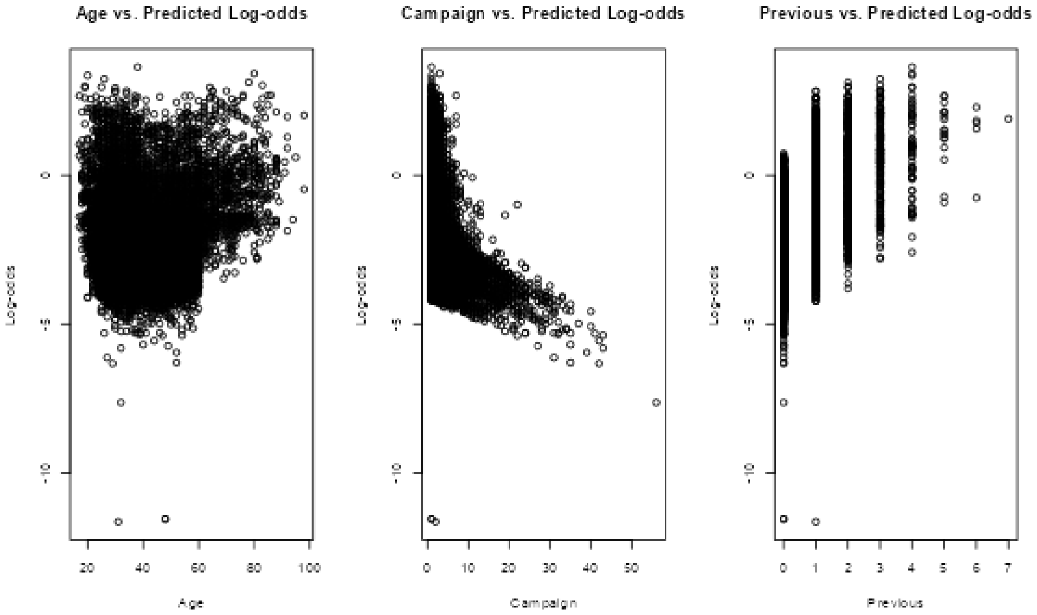

Figure 2: Scatterplots showing the relationships of continuous variables vs. the predicted log-odds

3.2 Social Media vs Personal Contact

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

Yes, as demonstrated in the example provided in Section 2.3, there is evidence that making contact via Social Media vs. Calling increases the likelihood of the client opening an account. It's important to note that we cannot assume causation from this analysis (ie. we cannot say that contacting via Social Media is the reason they decided to open an account), but there is a strong correlation.

3.3 Repeated Contact

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

Yes, because the 'previous' variable has a positive coefficient (.3351), this means that an increasing number of previous contacts results in an increasing likelihood of the client opening a new account. Again, we need to be careful about assuming causation.

4 Conclusion

We used the provided marketing data to develop a logit model to identify which factors most influence the likelihood of a client opening a new account. We identified that the characteristics of customers most likely to open a new account are those with less education, that are retired or are students, are older, are single, that have previously been contacted (the more the better!), and that have previously opened a card. In terms of timing, March was the peak month, with Wednesdays being the peak day. We also identified that contacting via Social Media versus Calling increases the likelihood of opening a new account, as does a larger number of previous contacts.

Moving forward, to gain a better understanding of what factors most influence the likelihood of clients opening a new account, we suggest finding ways to account for the imbalance of factors (as described earlier, some of the factors are heavily imbalanced towards one level, including the outcome). Also, we suggest identifying industry standard metrics and including those in the analysis, which may include things such as income.

5 Teamwork

Jarom: Abstract, Introduction, Model 2; **RJ**: Results, Conclusion, Model 1 **Both**: Revision

6 Appendix

Table 2: Model Coefficients

Variable	Estimate	Lower Bound: 2.5%	Upper Bound: 97.5%
(Intercept)	-2.8966	-3.2466	-2.5487
age	0.0052	0.0011	0.0092
campaign	-0.0715	-0.0899	-0.0539
contactsocialMedia	0.9431	0.8423	1.0447
day_of_weekmon	-0.1793	-0.2895	-0.0692
day_of_weekthu	0.0386	-0.0673	0.1448
day_of_weektue	0.0592	-0.0495	0.1679
day_of_weekwed	0.1079	-0.0002	0.2162
defaultunknown	-0.5419	-0.6526	-0.4333
defaultyes	-9.3723	NA	5.4636
educationbasic.6y	0.0849	-0.1161	0.2831
educationbasic.9y	-0.0663	-0.2242	0.0922
educationhigh.school	0.0225	-0.1303	0.1766
educationilliterate	0.9949	-0.3514	2.1076
educationprofessional.course	0.0482	-0.1209	0.2177
educationuniversity.degree	0.1392	-0.0138	0.2936
educationunknown	0.2026	0.0013	0.4020
jobblue-collar	-0.2375	-0.3689	-0.1066
jobentrepreneur	-0.1825	-0.3918	0.0188
jobhousemaid	-0.0887	-0.3345	0.1478
jobmanagement	-0.1245	-0.2682	0.0168
jobretired	0.4927	0.3152	0.6688
jobself-employed	-0.1014	-0.2983	0.0891
jobservices	-0.2049	-0.3498	-0.0621
jobstudent	0.6210	0.4275	0.8123
jobtechnician	-0.1264	-0.2440	-0.0097
jobunemployed	0.0765	-0.1397	0.2859
jobunknown	-0.1194	-0.5412	0.2741
maritalmarried	0.0791	-0.0343	0.1946
maritalsingle	0.2199	0.0913	0.3500
maritalunknown	0.2644	-0.4980	0.9332
monthaug	-0.8277	-0.9628	-0.6924
monthdec	0.8556	0.5094	1.1991
monthjul	-0.7194	-0.8527	-0.5858
monthjun	-0.0130	-0.1647	0.1389
monthmar	1.1381	0.9303	1.3459
monthmay	-0.7699	-0.8974	-0.6418
monthnov	-0.8952	-1.0428	-0.7482
monthoct	0.7102	0.5147	0.9050
monthsep	0.5165	0.2987	0.7329
poutcomenonexistent	0.3825	0.2141	0.5534
poutcomesuccess	2.1222	1.9696	2.2759
previous	0.3351	0.2301	0.4412