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 retired people or younger students, who are single, 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.

We also identified that the variable 'age' is not monotonic (see Figure 2: the likelihood that a client opens an account does not always increase/decrease with age). As such, we decided to bucket age into groups: ¡18, 18-25, 26-35, 36-50, 51-70, and 70+. This will allow us to avoid needing to meet any assumptions for this variable when building our models.

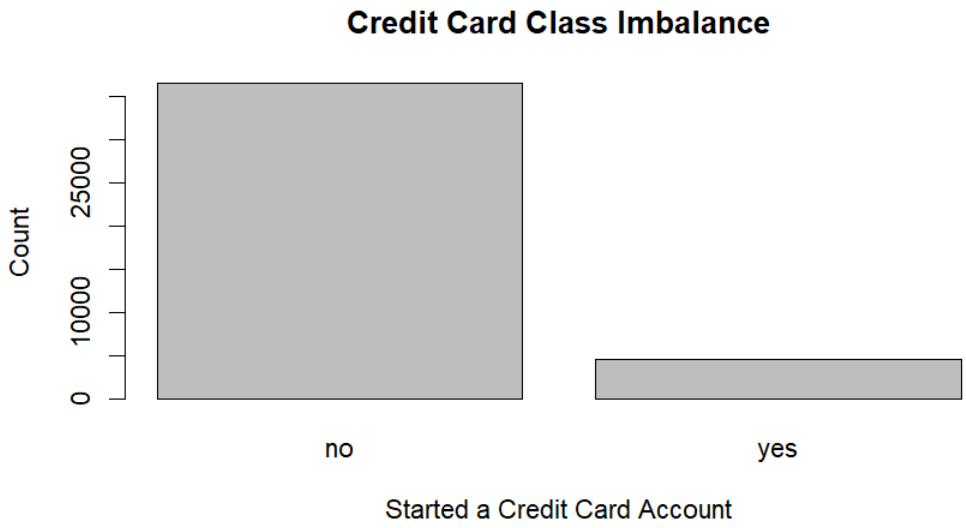

Figure 1: Bar plot showing class imbalance in the target variable.

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 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:

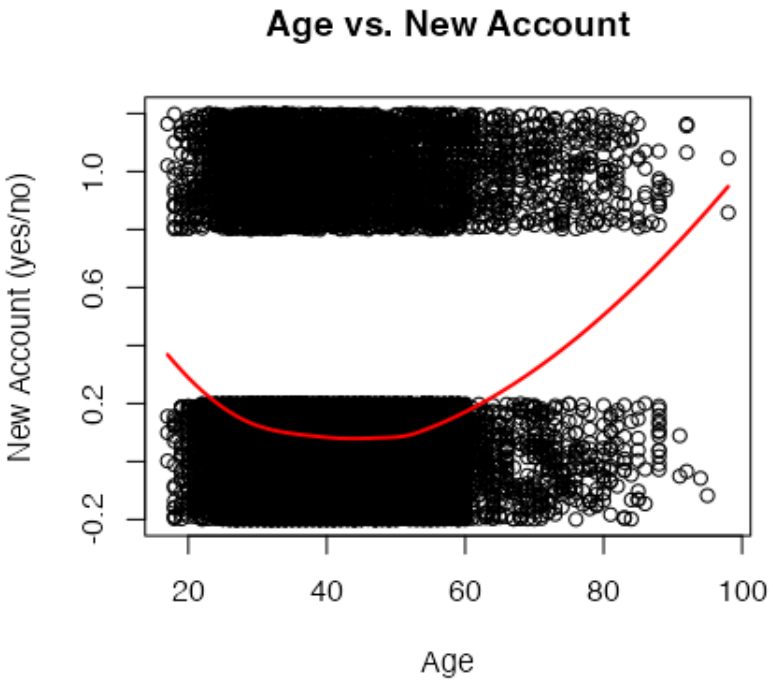

Figure 2: Plot showing that age is non-monotonic

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

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

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.

Table 1: Model AUC Comparison

Model	AUC
1	.7656
2	.7643

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. For our continuous variables, we need to confirm monotonicity. As discussed previously, the only continuous variable that was not monotonic was age, which we bucketed, meaning we meet our assumptions.

As an example of how to interpret the coefficients, using the variable contact:SocialMedia - a coefficient of .9242 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 $e^{.9242} = 2.52$. In other words, making contact with a client via Social Media vs. Personal Contact results in a 152% 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**: 70+, also 18-25. **Day of Week**: Wednesday. **Default**: No. **Education**: Illiterate (but also less education in general increases likelihood). **Job**: Retired, also students. **Marital Status**: Single (or unknown). **Month**: March. **Previous Marketing Outcome**: Success. **Previous**: More previous contacts increases the likelihood.
In short, the categories with the highest likelihood are single young students or single older retired people.

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 older retired people or young students, 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: Table of Coefficients and Uncertainties

Variable	coefs	Lower Bound: 2.5%	Upper Bound: 97.5%
(Intercept)	-3.4430112	-5.8452170	-1.1749729
age18-25	1.0377263	-1.2124410	3.4214913
age26-35	0.7832616	-1.4677327	3.1677876
age36-50	0.6420770	-1.6097763	3.0272799
age51-70	0.9228758	-1.3302147	3.3090079
age70+	1.5926161	-0.6747804	3.9908982
campaign	-0.0710059	-0.0893821	-0.0533546
contactsocialMedia	0.9242107	0.8234829	1.0258098
day_of_weekmon	-0.1764014	-0.2867117	-0.0661218
day_of_weekthu	0.0334614	-0.0726262	0.1397618
day_of_weektue	0.0583493	-0.0505299	0.1673349
day_of_weekwed	0.1035781	-0.0046573	0.2119593
defaultunknown	-0.5126595	-0.6238546	-0.4035804
defaultyes	-9.3219093	NA	5.5102911
educationbasic.6y	0.1426770	-0.0598313	0.3425006
educationbasic.9y	-0.0404023	-0.1994935	0.1195336
educationhigh.school	0.0584307	-0.0958630	0.2140720
educationilliterate	0.9825538	-0.3906690	2.1116548
educationprofessional.course	0.0898131	-0.0808744	0.2610964
educationuniversity.degree	0.1758358	0.0212479	0.3319538
educationunknown	0.2318158	0.0288115	0.4329803
jobblue-collar	-0.2295188	-0.3610445	-0.0984067
jobentrepreneur	-0.1639497	-0.3730833	0.0372592
jobhousemaid	-0.1041341	-0.3510259	0.1333591
jobmanagement	-0.1162118	-0.2600015	0.0251626
jobretired	0.2648357	0.0757781	0.4516608
jobself-employed	-0.1047968	-0.3019019	0.0858554
jobservices	-0.2196004	-0.3646968	-0.0766886
jobstudent	0.4448398	0.2402245	0.6472153
jobtechnician	-0.1272589	-0.2448140	-0.0105247
jobunemployed	0.0839462	-0.1317957	0.2929904
jobunknown	-0.1459557	-0.5693694	0.2491038
maritalmarried	0.0806542	-0.0333834	0.1966859
maritalsingle	0.1551773	0.0261730	0.2856381
maritalunknown	0.2593303	-0.5026975	0.9268543
monthaug	-0.8214416	-0.9569126	-0.6857108
monthdec	0.8254237	0.4776089	1.1705519
monthjul	-0.7350491	-0.8686863	-0.6010735
monthjun	-0.0279573	-0.1797773	0.1240129
monthmar	1.1064919	0.8980625	1.3149341
monthmay	-0.7763564	-0.9041593	-0.6479052
monthnov	-0.8876971	-1.0358741	-0.7402037
monthoct	0.6791084	0.4827941	0.8747656
monthsep	0.4790891	0.2605954	0.6963053
poutcomenonexistent	0.3778367	0.2092798	0.5488923
poutcomesuccess	2.1194824	7 1.9666866	2.2734558
previous	0.3273313	0.2222961	0.4334967