Customer Long Term Deposit Turnover Analysis

REPORT

Athanasioy Antonis, MSc Business Analytics, P2822102

SUMMARY

A Generalized Linear Model was created in order to identify important factors that affect whether or not a bank depositor chooses to subscribe to a long-term deposit promoted by the bank. After conducting a brief data analysis and analysis of associations, the model was created using stepwise method with BIC as the model selection criterion. The important factors that made it into our model was month, age group, default status, contact device type, previous outcome of similar campaign, call duration, employment variation rate, consumer confidence index, 3 month Euribor and price index. The final model is much than the naive model, whoever it still misses critical components of the underlying client behavior pattern under examination.

INTRODUCTION

The purpose of this report is to identify which factors play a significant role in determining whether or not a customer "buys" long term deposits. Roughly 40k observations were collected regarding phone calls made to depositors. Along with the main variable of interest, demographic details about depositors are available as well as previous contact history with said depositor. Lastly, macroeconomic indicators at the time are also recorded in hopes of adding explanatory power.

EXPLONATORY ANALYSIS & ANALYSIS OF ASSOCIATIONS

Our dataset consists of 39883 phone calls. Out of those 39883 phone calls, 3987 *(10%)* resulted in the depositor buying the long-term deposit and the rest 35896 *(90%)* resulted in a negative result.

For each phone call, we have demographic data about the customer (age, marital status, job and education), client related data such as default status on loans, contact information data such as day of the week and month on which the phone call was made, contact history with said client and various macroeconomic indices (**Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε.**). It is reasonable to assume that many of the above variables affect whether a depositor chooses a long-term deposit plan. A full Description of our data can be seen in Table 1.

The mean age of prospective clients is 40 years old, and the mean duration of each call is 4 minutes (258 seconds) (Table2). The youngest prospective client is 17 years old and the oldest is 98. The typical client has an administrative job, is married and has a university degree (Table 3). Most clients have not defaulted on their loans, but we do have some "unknown" values. Most depositors have also a housing loan, while the majority of them haven't taken a loan (Table 4).

As seen in Table 2, most of numeric variables have outliers, as indicated by the high skewness (for example in duration, campaign etc.)

Variable	Class	First Values
age	numeric	56, 57, 37, 40, 56, 45
job	factor	housemaid, services, services, admin., services, services
marital	factor	married, married, married, married, married, married
education	factor	basic.4y, high.school, high.school, basic.6y, high.school, basic.9y
default	factor	no, unknown, no, no, no, unknown
housing	factor	no, no, yes, no, no, no
loan	factor	no, no, no, no, yes, no
contact	factor	telephone, telephone, telephone, telephone, telephone, telephone
month	factor	may, may, may, may, may, may
day_of_week	factor	mon, mon, mon, mon, mon, mon
duration	numeric	261, 149, 226, 151, 307, 198
campaign	numeric	1, 1, 1, 1, 1, 1
pdays	numeric	999, 999, 999, 999, 999, 999
previous	numeric	0, 0, 0, 0, 0, 0
poutcome	factor	nonexistent, nonexistent, nonexistent, nonexistent, nonexistent, nonexistent
emp.var.rate	numeric	1.1, 1.1, 1.1, 1.1, 1.1, 1.1
cons.price.idx	numeric	93.994, 93.994, 93.994, 93.994, 93.994, 93.994
cons.conf.idx	numeric	-36.4, -36.4, -36.4, -36.4, -36.4, -36.4
euribor3m	numeric	4.857, 4.857, 4.857, 4.857, 4.857, 4.857
SUBSCRIBED	factor	no, no, no, no, no, no

Table 1 Dataset Description

	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m
n	39883.00	39883.00	39883.00	39883.00	39883.00	39883.00	39883.00	39883.00	39883.00
mean	39.98	256.70	2.59	972.80	0.14	0.13	93.55	-40.46	3.71
sd	10.18	258.84	2.80	159.20	0.42	1.57	0.57	4.61	1.69
median	38.00	177.00	2.00	999.00	0.00	1.10	93.44	-41.80	4.86
min	17.00	0.00	1.00	0.00	0.00	-3.40	92.20	-50.00	0.63
max	98.00	4918.00	56.00	999.00	5.00	1.40	94.47	-26.90	5.04
range	81.00	4918.00	55.00	999.00	5.00	4.80	2.26	23.10	4.41
skew	0.73	3.26	4.73	-5.91	3.60	-0.81	-0.21	0.36	-0.81
kurtosis	0.62	20.04	36.31	32.94	17.30	-0.94	-0.84	-0.40	-1.24

Table 2 Descriptive Statistics of our variables

job	marital	education
admin. :10020	divorced: 4469	University degree :11667
blue-collar: 9174	married :24320	High school: 9218
technician: 6537	single :11015	basic.9y: 5971
services: 3894	unknown: 79	Professional course: 5049
management: 2845		basic.4y: 4047
retired: 1543		basic.6y: 2270
(Other): 5870		(Other): 1661

Table 3 Summary of Job, Marital and Education Variables

default	housing	loan	Subscribed
no :31309	no :18059	no :32868	Yes: 3987
unknown: 8571	unknown: 955	unknown: 955	No: 35896
yes: 3	yes :20869	yes : 6060	

Table 4 - Summary of Default, Housing and Loan Variables

Associations

Demographics

In Figure 1 significant variation of the success percentage is observed. Starting from the peak at around 45% at 17 years, a steady decline is observed which flattens at around 25 years old and remains stable until 60. After 60, the success percentage increases noticeably and due to the low number of observations, especially after 80, wild variation is observed. The color of the line represents the amount of observation we have for each age in log scale to make comparisons between extreme variations more visible.

From the graph, we can clearly conclude that the relationship is not linear and that there are distinct age groups that behave differently. To capture this non-linearity, a new variable "age group" was made as depicted in Figure 2, which holds ages from 17-25, 26-62 and 63+ years old. It is noted that the numbers inside bars represent the amount of observation available for said part. Figure 2 depicts success percentage for these age groups. Noticeable differences are observed, with the group of 63+ showing the highest subscription rate and the 26-62 group showing the lowest.

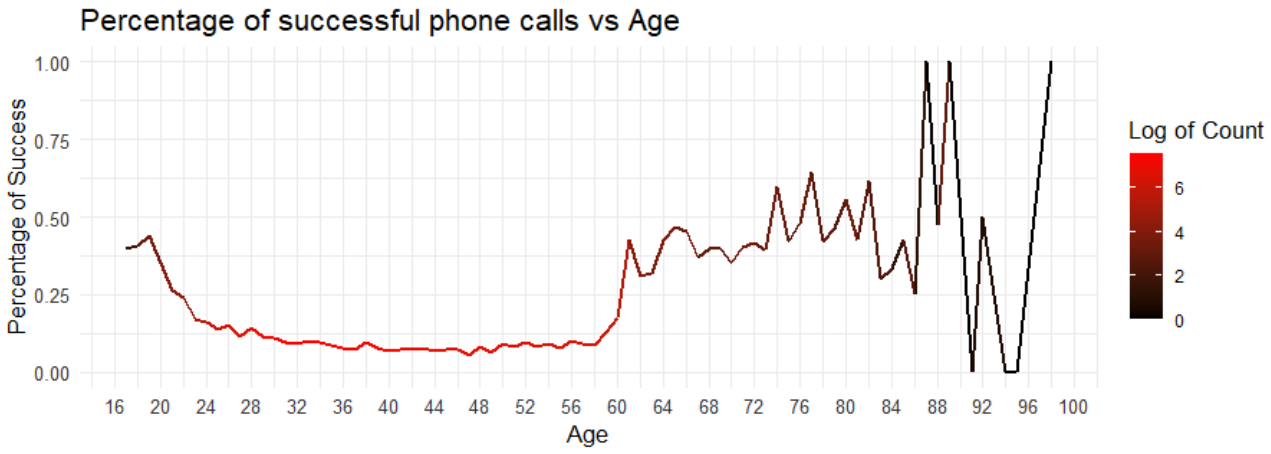


Figure 1 Variation of percentage of successful calls and client age

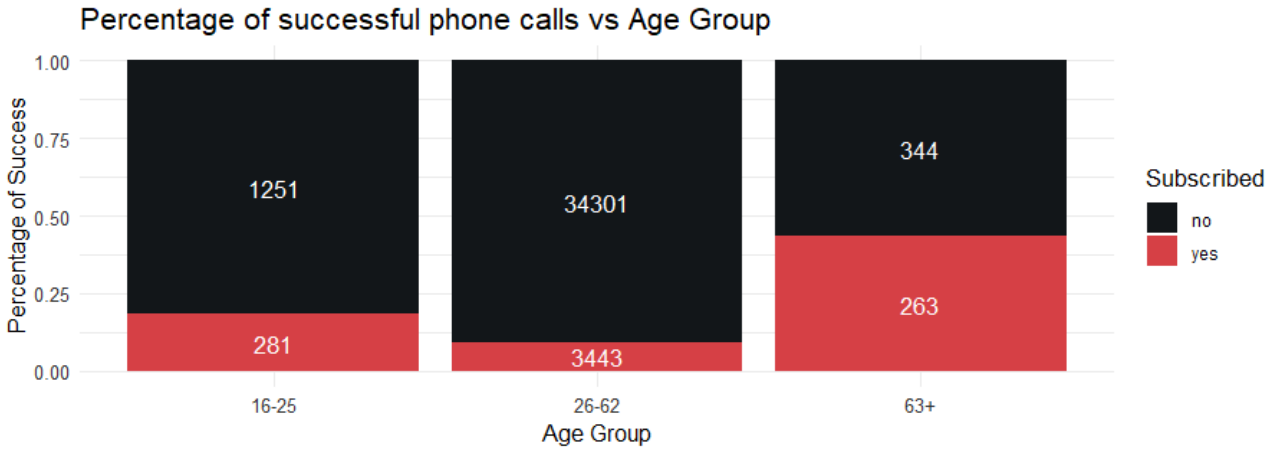


Figure 2 Percentage of successful calls per age group

On the other hand, marital status does not display significant explanatory power. Whether the depositor married, single or divorced, the percentage of successful calls is roughly the same and around the mean success percentage of 10% (Figure 3).

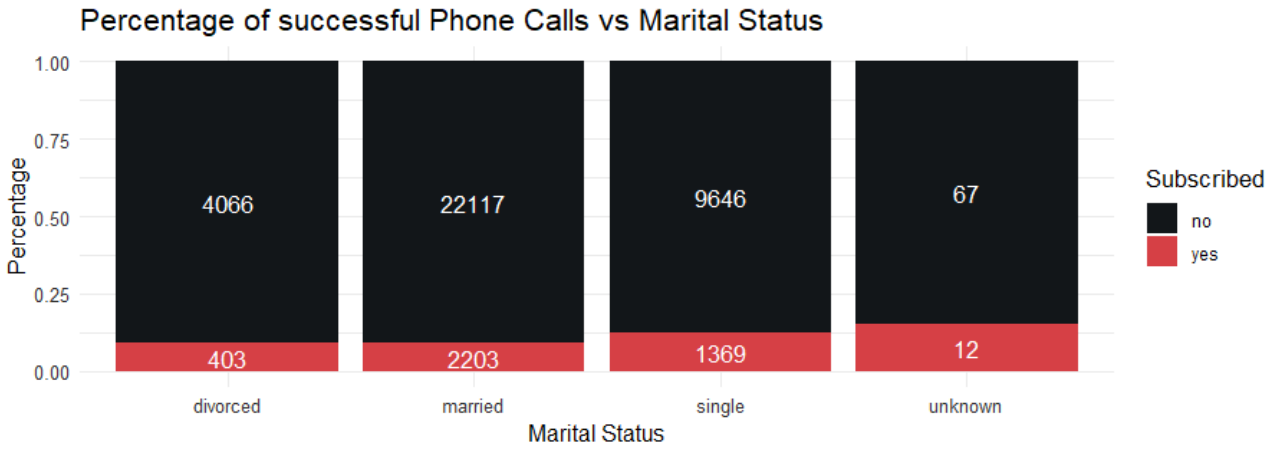
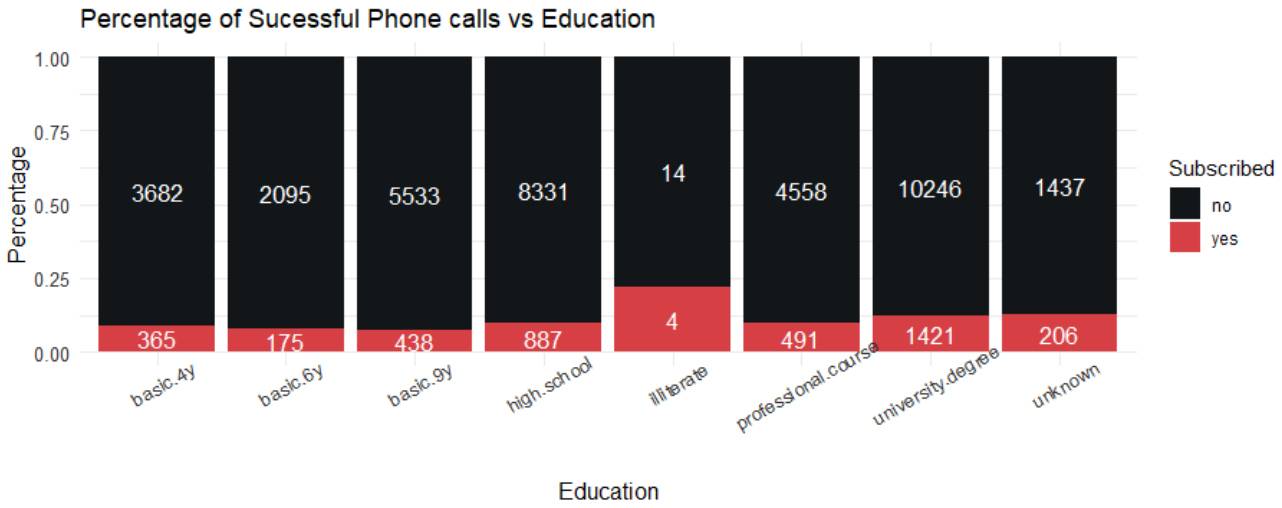


Figure 3 percentage of successful calls per marital status

Similarly, there isn't much difference between different education levels. Only illiterates have a noticeable higher percentage of subscription, but the total sample for that category is very low (n = 18), thus no reliable conclusions can be made.



Lastly, two of out 12 job categories exhibit explanatory power of our response variable. Students and retirees have a subscription rate above the mean of 10%, while the rest of job categories are roughly the same are the global sample mean (Figure 4).

Credit status does seem to play a minor role in determining subscriptions. Clients for which it is certain that that they haven't defaulted on their loans go forward with long term deposits more often than those which it is not. On the other hand, whether on not the depositor has taken either a mortgage or a commercial loan doesn't affect subscription rates (Figure 5 ,Figure 6).

All in all, as far as demographics are concerned, the only factor with real explanatory power of subscriptions age; neither education nor marital status provide much differentiation in our response variable. Second in order comes job, especially for students and retirees. However, these variables are highly correlated with our age groups, thus the combined end result will be more or less same.

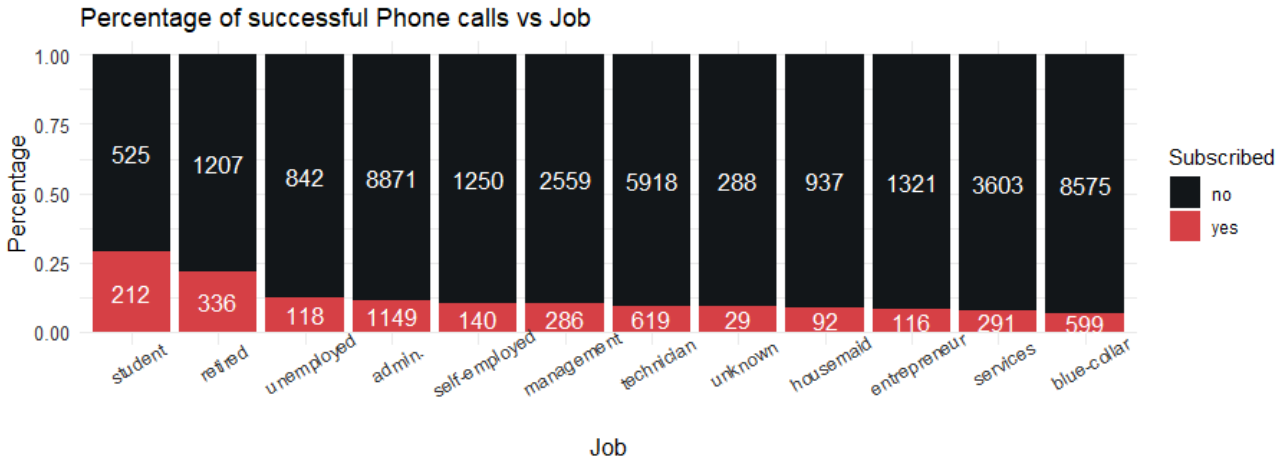


Figure 4 Percentage of successful calls per job

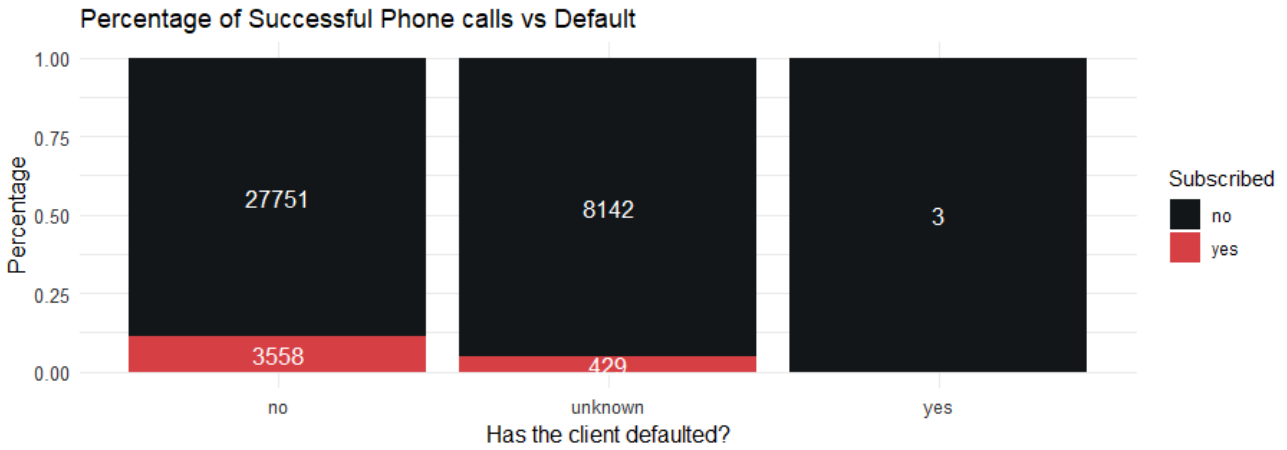


Figure 5 Success rates vs Default Status

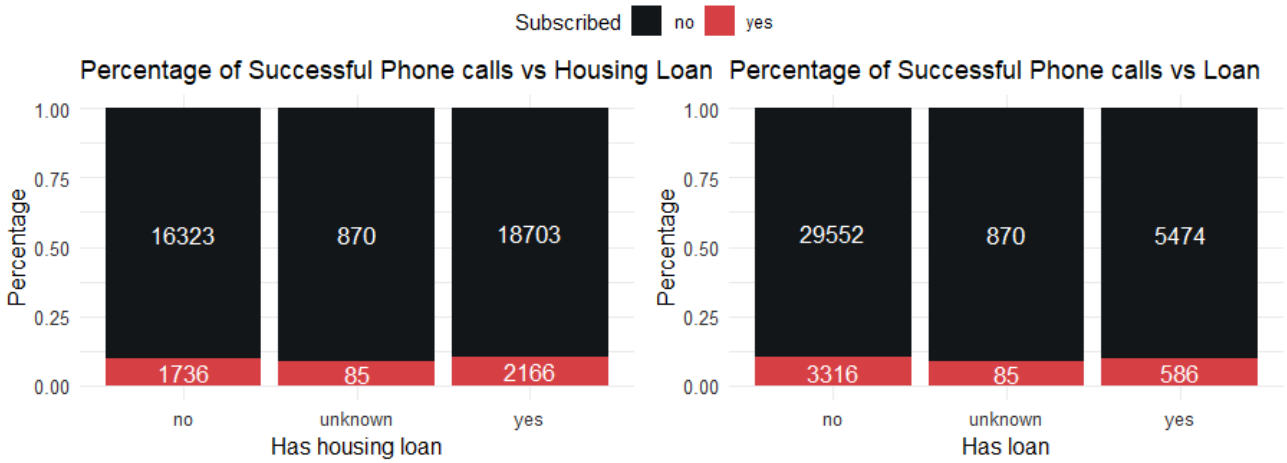


Figure 6 Success Rate vs Mortgages, Loans

Contact History

When examining contact history, there is no prior expectation about relationships; reasonable arguments for both positive and negative relationships (or none at all) could be provided. Thus, low effort will be given to provide explanations about observed behaviors. Having said that, the month of the call seems to play a significant role in determining success. Calls made in December, September, October and March show

almost a 50% success rate, a significant increase to our mean 10% success rate (Figure 7). Curiously, the lowest number of calls happen at those months as well.

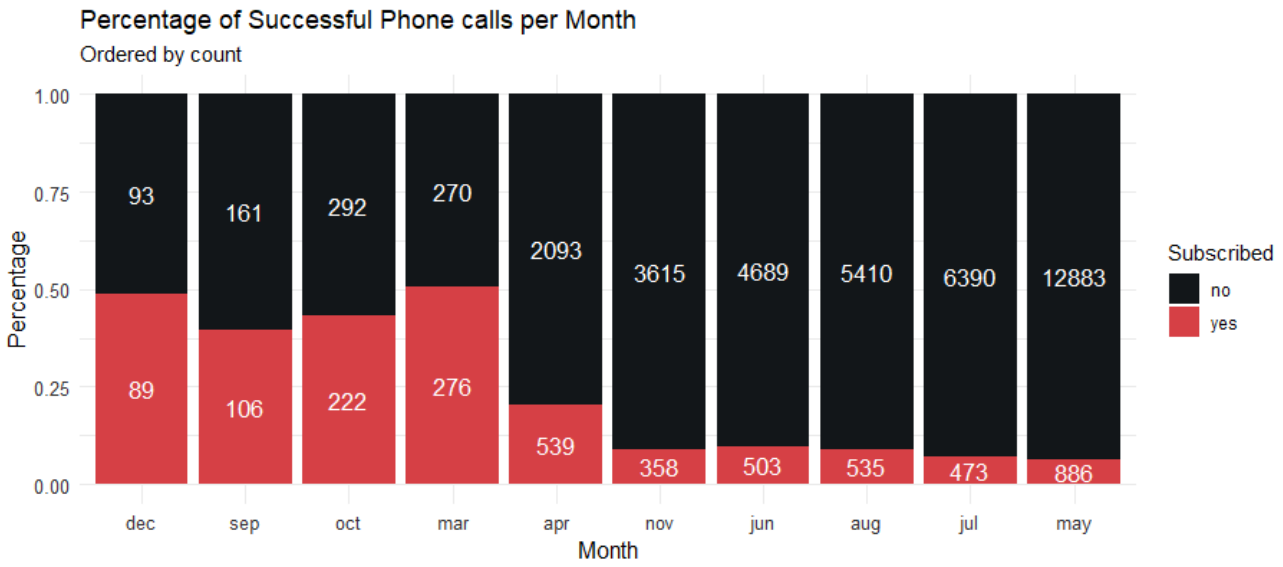


Figure 7 Success Rate per Month

The number of contacts for this campaign is correlated negatively with success rate. As more contacts were made from the bank to the specific client, the less likely it was that the client would actually go forward with a long-term deposit as seen in Figure 8.

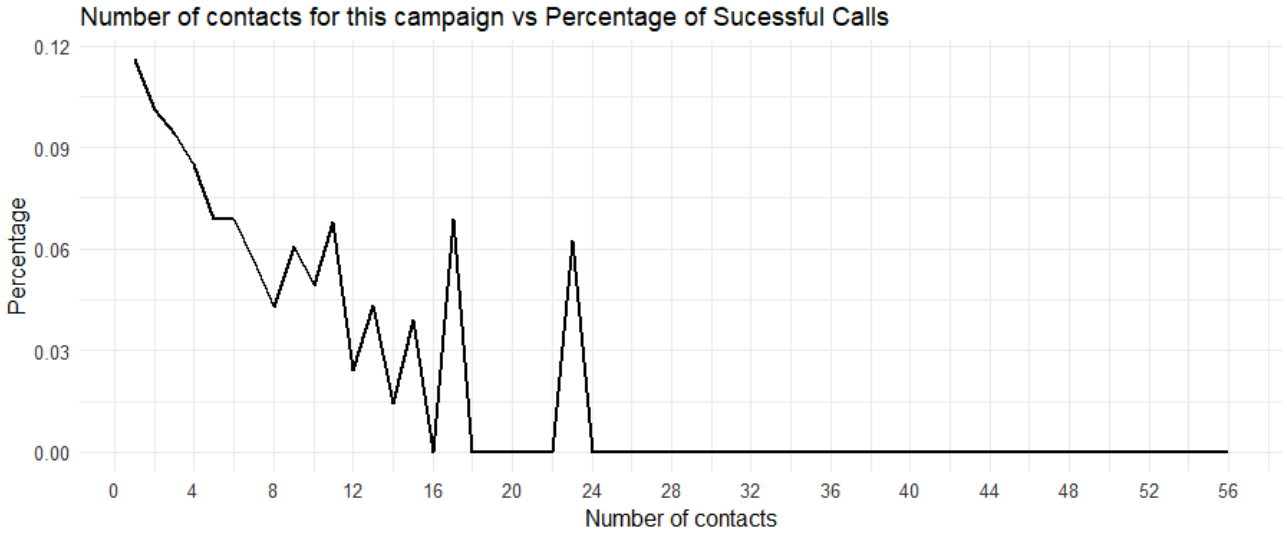


Figure 8 - Success Rate per Number of Contacts for this Campaign

However, the number of previous contacts made before the current campaign are very positively collerated with success rates. The probabilty of the client subscribing to the long term deposit increases almost linearly with the number of contacts made with this client before the campaign. This of course slightly cotrandicts the above observation. Lastly, the contact device seems to play a defining factor. When the client was contacted using a cellular device, the probability of subscription more than doubles when compared to standard telephone (from 8% to 18%) (Figure 10).

Last but not least, a major factor that defines success rate is the previous outcome of a similar campaign. If the previous outcome of a similar campaign was successful, the mean percentage of success was above 50%, whereas when the previous outcome was a failure or non-existent, this percent drops to around 10%.

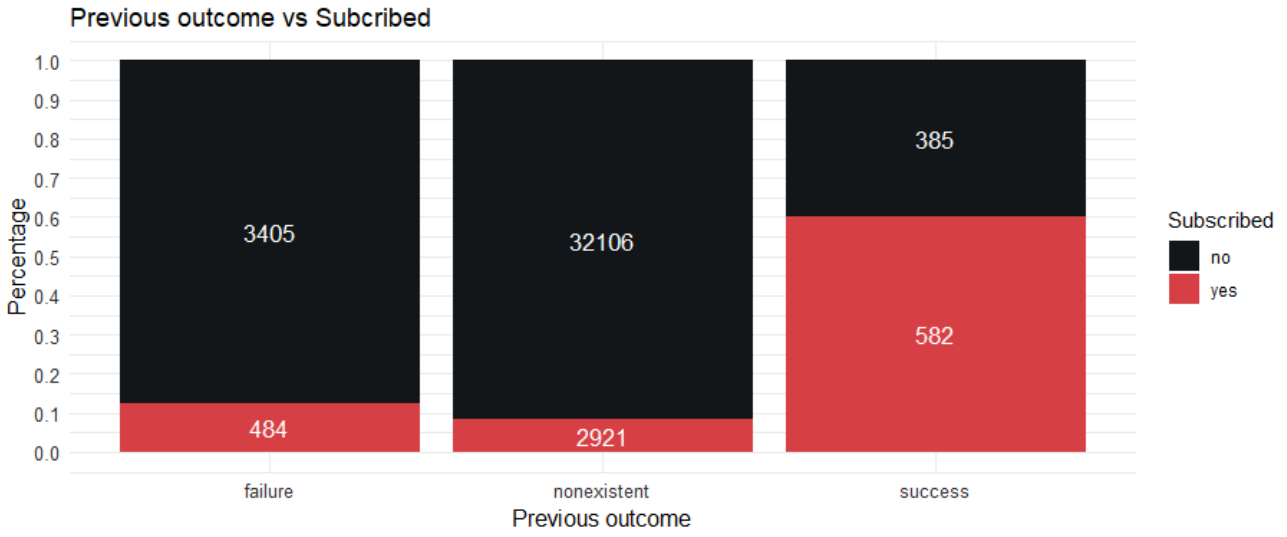


Figure 9 - Success Rate per Previous Campaign Outcome

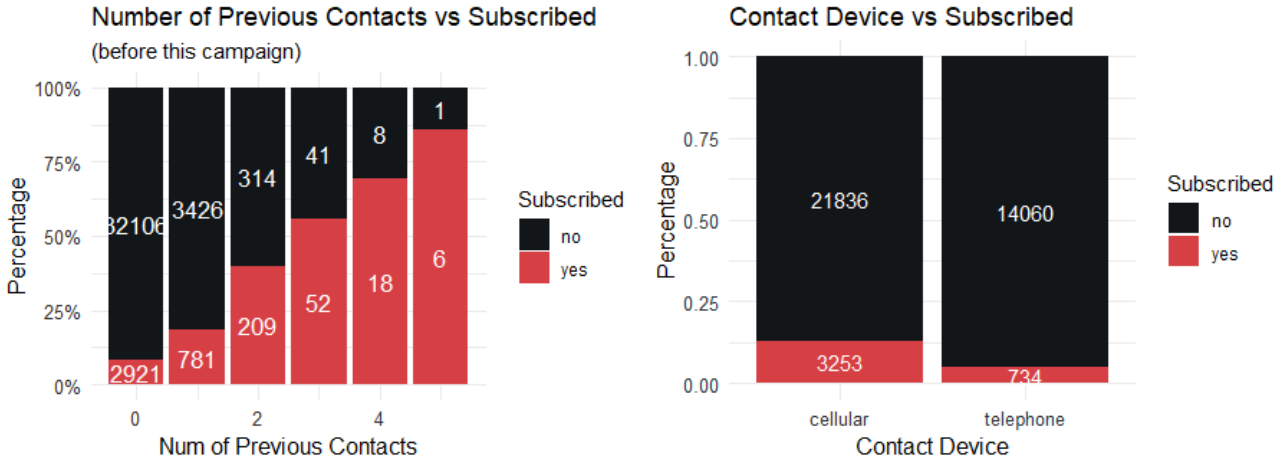


Figure 10 - Previous Contant and Contact Device vs Success Rate

Macroeconomic Factors

The most interesting macroeconomic variable is the 3-month Euribor rate, as this rate is directly correlated with most deposits rates and with interest rates in general. From Figure 11 we observe the boxplots of 3month Euribor per Subscription. The median value of euribor3m when the answer is "No" is almost 5, whereas when the answer is "Yes" the median values drops to below 2. This implies a negative relationship between the 3month Euribor and the probability that the client will subscribe to our long-term deposit. Of this contradicts prior expectations, as depositors have a greater incentive to subscribe to long-term deposits when interests are high. We have to keep in mind that during 2008-2010, a global debt crisis occurred, which may affect how economic decisions are made. In that respect, this negative correlation may be a spurious one, which may not be verified when we sample data from a different, more stable, economic period.

Figure 11 Boxplot of Euribor3m and Subscription Rate

MODEL BUILDING

To build our model we implemented stepwise procedure using BIC as the model selection criterion, which penalizes large models to a greater degree than AIC. Since the purpose is to interpret and find the important the important factors rather than predict, BIC is the recommended selection criterion. Unsurprisingly, the stepwise method kept all of the variables that our analysis indicated as important, such as Default Status, Contact Type, Month, Duration, Previous Outcome and 3-month Euribor. Surprisingly, all of our macroeconomic indexes are very statistically important, even though they present significant collinearity. Removing them from the model makes our model worse, judging from statistically significant decreases of model Deviance. Our final model is the following (Table 6):

$$\log\left(\frac{\pi_{subscribed}}{1-\pi_{subscribed}}\right) = -370.5 - 0.329 defaultUnknown - 7.204 defaultYes$$
$$-0.4712 ContactTelephone + 2.981 Aug + 0.899 Dec + 0.861 Jul$$
$$-0.95 Jun + 2.315 Mar - 0.083 May + 0.304 Nov + 1.254 Oct$$
$$+0.939 Sep + 0.0047 Duration + 0.415 PNonExistent + 1.720 PSuccess$$
$$-2.817 emp.\text{var}.rate + 3.847 cons.price.idx - 0.061 cons.conf.idx$$
$$+0.998 euribor3m - 0.29 ageGroup26\_62 + 0.094 ageGroup63 +$$

In our case, the log proportion of success and failure follows a linear relationship. In practice, this means our model coefficients represent percentage increases/decreases of the success/failure proportion. Thus, coefficients do not have standard interpretation, given that changes in the independent values translate to different change in the response, depending on the base probability.

For example, if it is unknown if the client has defaulted or not, the probability of him subscribing to the long term deposit is multiplied by $e^{-0.329} = 0.72$, in other words, a decrease of 28%. Likewise, an increase by one in one of numeric variables such as consumer confidence index, will result in an increase of $e^{3.85} - 1 = 46.83 - 1 = 45.83$. Table 5 contains the full table of relative increases in probability. To get a better understanding of modeled relationships, we plot independent variables with the fitted probability of success, as seen in Figure 12.

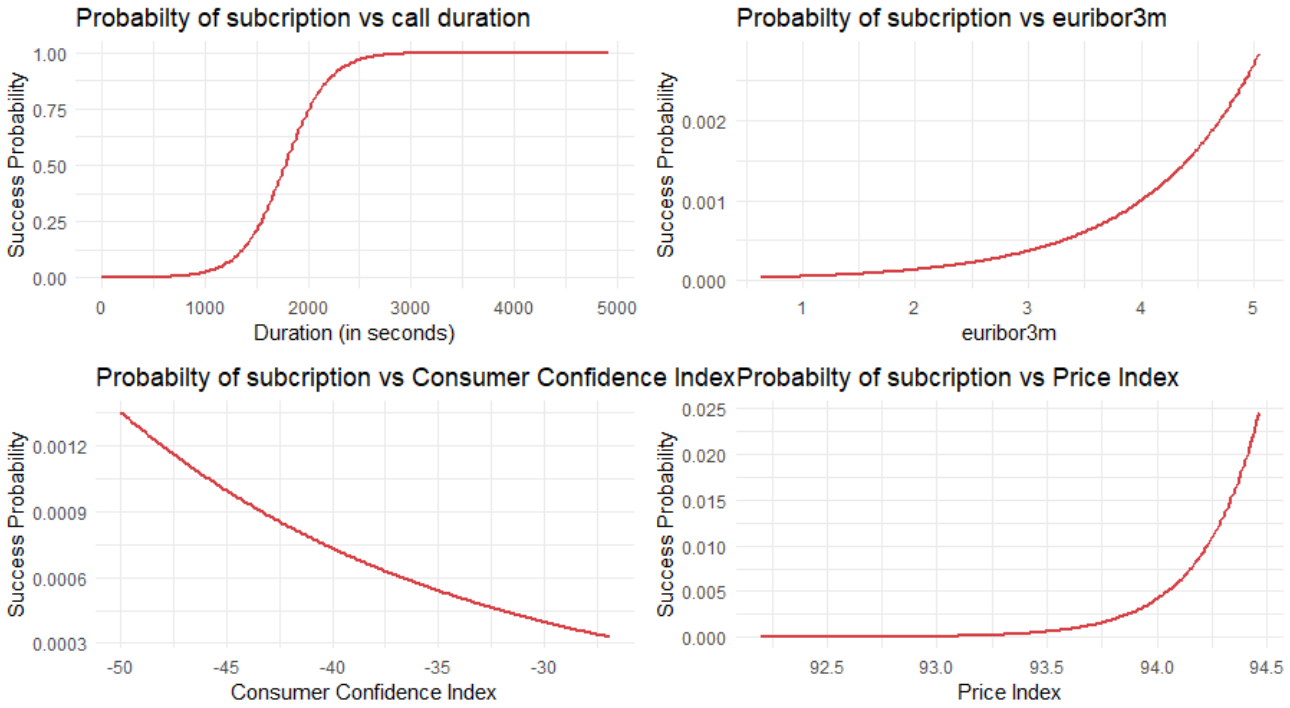


Figure 12 Fitted Probability of Success vs Numeric Variables. All else variables are considered constant. Specifically, we take the mean if the variable is numeric and the mode if the variable is a factor

It is evident that as calls last longer, it becomes more likely that the client will subscribe to the long-term deposit, with all the other variables constant (in our case, the other variables are either the mean sample value if the variable is numeric or the mode if the variable is a factor). Likewise, when Euribor increases, so does the probability of subscription.

Variable	Relative Increase in Probability
defaultunknown	0.72
defaultyes	0.00
contacttelephone	0.62
monthaug	19.71
monthdec	2.46
monthjul	2.37
monthjun	0.39
monthmar	10.12
monthmay	0.92
monthnov	1.36
monthoct	3.50
monthsep	2.56
duration	1.00
poutcomenonexistent	1.52
poutcomesuccess	5.58
emp.var.rate	0.06
cons.price.idx	46.83
cons.conf.idx	0.94
euribor3m	2.71
age_group26-62	0.75
age_group63+	1.10

Table 5 - Relative Multiplication of base Probability per one unit of incrase in our indepedent variables

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-370.55	18.47	-20.06	0.00
defaultunknown	-0.33	0.07	-4.96	0.00
defaultyes	-7.20	113.48	-0.06	0.95
contacttelephone	-0.47	0.08	-6.03	0.00
monthaug	2.98	0.20	14.86	0.00
monthdec	0.90	0.21	4.32	0.00
monthjul	0.86	0.12	7.28	0.00
monthjun	-0.95	0.12	-8.14	0.00
monthmar	2.31	0.12	19.68	0.00
monthmay	-0.08	0.08	-1.00	0.32
monthnov	0.30	0.14	2.16	0.03
monthoct	1.25	0.18	6.85	0.00
monthsep	0.94	0.21	4.45	0.00
duration	0.00	0.00	62.27	0.00
poutcomenonexistent	0.42	0.07	6.19	0.00
poutcomesuccess	1.72	0.10	17.65	0.00
emp.var.rate	-2.82	0.14	-19.80	0.00
cons.price.idx	3.85	0.19	19.97	0.00
cons.conf.idx	-0.06	0.01	-7.72	0.00
euribor3m	1.00	0.10	10.42	0.00
age_group26-62	-0.29	0.09	-3.29	0.00
age_group63+	0.09	0.13	0.73	0.47

Table 6 - Model Coefficients, Standard Errors and P-Values

DISCUSSIONS AND RESULTS

How good is the model? Model Residual Deviance is equal to 15680 and Null Deviance, i.e. when we assume for every case the sample mean probability of subscription, is 25925. This difference is significant enough to conclude that our model is better than the naïve approach to assume the sample mean for every case. Formally, we take 2 times difference the Log Likelihood of our model and the log Likelihood of the Null Model. Under the null hypothesis, the difference follows a chi square distribution with p-1 degrees of freedom. Specifically,

$$2\big(-7839.887 - (-12962.43)\big) = 10245$$

The P-Value of this value under the null is near Zero, so we reject the null hypothesis that the two models are equal.

However, by looking at Figure 13 we can make two important conclusions. First, for most of our observations, as inferred by the density of the dots, is located at the lower left corner as it should. This means that our model assigns a low probability of subscription for the majority of observations. Secondly, from the graph it is evident that there are observations where our models assign a low probability (under 20%) but the actual outcome was client subscription (top left corner). Likewise, there are observations where our model is almost sure of client subscription, but the actual outcome was no subscription (bottom right corner). On the same note, the ideal graph would have accumulated the points on the bottom left and top right corner respectively. The fact that neither the top nor the bottom line do not show any kind of group on the "correct" corners suggests that our data and model, while much better than the simplest naïve approach, still misses some critical component of the underlying client behavior pattern that we are trying to explain.

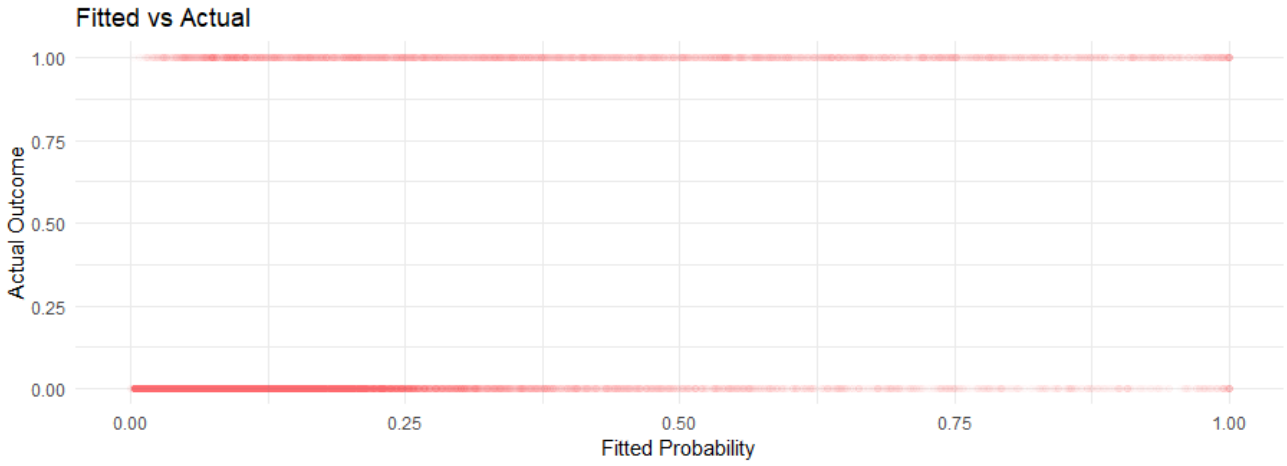


Figure 13 Predicted Probability vs Actual Outcome