DA2/Assignment 2 - Analysing a bank telemarketing campaign

Zsombor Hegedus

2020 november 29

Abstract

This paper introduces regression models to uncover a pattern of association between the success of a telemarketing campaign and information known about clients. The data used holds information of a telemarketing campaign of a Portuguese bank that was collected between May 2008 to November 2010. LPM, logit and probit models were used to analyse events with a binary outcome which indicates whether given client subscribed to a term deposit as a result of the campaign or not. A probit model using various demographic, macro and campaign related variables proved to be the best choice to uncover the patterns.

Introduction

I once read that personal finances can be considered as one of the most intimate of areas that can make people very insecure. When something is considered as such, it can often happen that one sweeps it under the rug and don't pay that much attention as to e.g. what do they do with their savings? Some might just declare themselves unable to understand the complex world of finance and withdraw their salary in cash so that they can put it in a pillow. But that might not be the best way of handling wealth. Risk-free investment options exist in the market that can yield considerable interest while banks are also in need of money from their consumers. One of the safest means for a bank to get liquidty is by increasing client deposits. A term-deposit for example is a fixed term investment in which a customer entrusts the banks to use their cash for a given period of time in exchange for some interest. In order to break the silos and incentivise people to give them cash, banks can use direct marketing in which they introduce options to clients so that they can generate income by investing their wealth in safe instruments.

In this paper I look at the pattern of association between the success of a telemarketing campaign and information available from the clients using regression models. Predicting the outcome of a campaign has been thoroughly analysed by Moro et al. (2014) while my focus is getting a better understanding in the connection between the variables used. Understanding these relationships could help marketing managers better plan their campaigns and target their clients effectively, thus saving on costs.

Data

I used *Bank Telemarketing* , a publicly available dataset with 41,188 records in which each observation is a call between the employees of a Portuguese bank and their clients. This is observational data that was collected between May 2008 and November 2010. The data has a binary variable which shows whether given client subscribed to a term loan at the bank as a result of the campaign. This is going to be the dependent variable in the analysis and takes the value of 1, in case the client subscribed to term deposit. For explanatory variables I defined 5 main categories , which are the following:

- **Demographic information about clients** - such as age, marital status, job, education
- **Financial information of clients** - whether the client had personal loan, housing loan, or any defaulted credit
- **Campaign related information** - how the client was contacted, duration of the call, number of contacts performed in current and last campaign, outcome of previous marketing campaign
- **Time variables** - in which season was the client contacted

- **Macroeconomic variables** - such as the number of employees in the bank, the 3 month Euribor rate, consumer price index and employment variation rate

In order to improve data quality, I cleaned this dataset. I first removed duplicates and missing values that were encoded with either the number 999 or with a string *unknown* - I decided not to use imputation techniques as my dataset was already large enough.I wanted to reduce the complexity of my dataset, and dropped unused variables/reduce the number of categories in the categorical variables. In this spirit, I dropped the variable showing whether the customer had defaulted credit, and the number of days before client was last contacted, as they had almost no variation (more than 95% of the values were the same).I also reduced the number of categories for *education* - the variable now only shows whether someone had university or equivalent level of education. I also transformed month variable into seasons, to have less categories. The cleaned dataset has 38,234 observations.

Only around 11% of all events ended up with a success in which the client subscribed to a term deposit. This can be seen visually in different breakdowns (filled histograms where color distinghises cases with success, where yellow is 1, which is) in the Appendix. There aren't any deviation in the data, the 10% success rate seems to be almost the same in every category, maybe except in the case of *duration*. Some summary statistics are also avaialable in Table 1:

statistics	age	duration	campaign	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	output
mean	39.86	258.24	2.57	0.17	0.08	93.57	-40.54	3.62	5167.43	0.11
median	38.00	180.00	2.00	0.00	1.10	93.44	-41.80	4.86	5191.00	0.00
min	17.00	0.00	1.00	0.00	-3.40	92.20	-50.80	0.63	4963.60	0.00
max	98.00	4918.00	43.00	7.00	1.40	94.77	-26.90	5.04	5228.10	1.00
sd	10.29	259.82	2.77	0.49	1.57	0.58	4.62	1.73	71.76	0.31
# missing	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
# used obs	38234.00	38234.00	38234.00	38234.00	38234.00	38234.00	38234.00	38234.00	38234.00	38234.00

Table 1: Summary statistics of quantitative variables

EDA

In the EDA phase my goal was twofold, and I only looked at the quantitative variables. On the one hand I wanted to see if the correlation between variables can give me a better understanding on my dataset, and if there are any highly correlated variables that I might not keep after all. On the other hand I wanted to see if there is any non-linearity between the explanatory variables and the dependent variable.

Correlations can be seen in as the third figure in the Appendix. The correlation matrix shows that there is a very strong positive correlation between macroeconomic variables, which was not very surprising given that all of them have the same underlying driver - the strength of the local/European economy. It is worth to note that the data was recorded in the midst of a financial crises. In addition to that, these metrics were updated on a monthly, or quarterly basis, and kept flat throught the course of the period so high correlation is even less surprising. It could have been argued that leaving these highly correlated variables out is sensible, however I decided not to drop them as later on they were useful contributors and improved the model fit. We could also see that they are mildy negatively correlated with the dependent variable, called *output*.

I also used a non-parametric regression model, **lowess** to see the relationship between some explanatory variables and the dependent variable. There were three interesting cases that can be seen in Figure 1. There is a non-linear relationship between *age* and the probability of success of a telemarketing campaign. The lowess resembles a quadratic function that has it's maximum at around age 45. There is a similar convex curve when using *duration* as explanatory variable. It seems that longer phone calls were associated with higher probabilities of success, which is quite intuitive, if people are not willing to consider subscribing, they wouldn't talk on the phone for too long. Lastly, *campaign* variable shows that after 15 tries, average probabilities for success were close to 0 which is also intutive - nobody likes to be harassed by their bank.

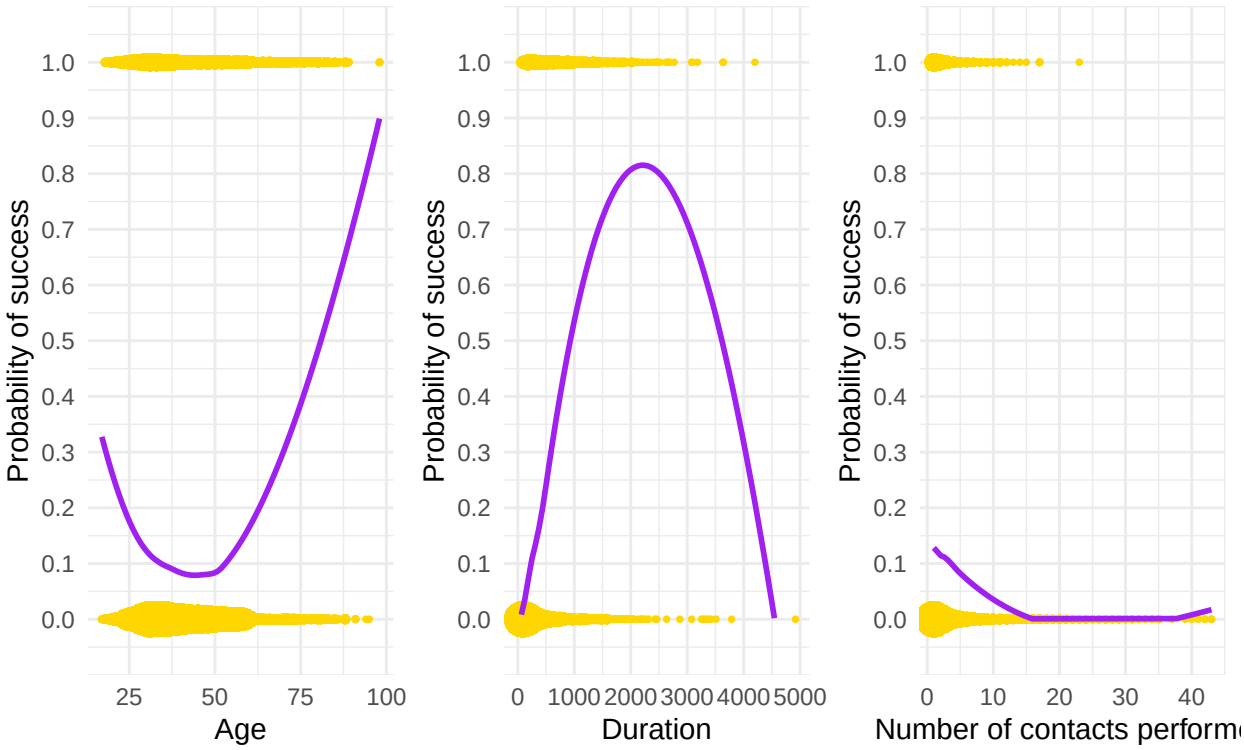


Figure 1: Three lowess functions for age, duration and campaign variabls vs the dependent variable

Model Choice

My goal is to uncover the pattern of association between the success of a marketing campaign and information available about clients. Since I want to understand the probability of such an event happening, and since my dependent variable is binary, I decided to employ methods that can model probabilities. In this chapter first I use multiple linear probability models where I gradually introduce new sets of variables, to see whether it makes sense to add them or not. After deciding on the variables to include, I will use two more sophisticated models to overcome the shortcomings of LPM, the logit and the probit.

Experimenting with LPM

My dataset is quite rich in both categorical and numerical variables, but perhaps I can drop some of them to reduce the complexity. This is a tradeoff that I am willing to take - if my ultimate goal was prediction, I might have included every variable, but in this case, I am fine to leave out some that might not improve the fit of the model, or that are not significant. I also used log transformation on *campaign* variable, and will use the second order polynomial of *age* and *duration*. Table 3 in the appendix shows three models, where the first one has variables from the demographic category only, the second has every available variable included and in the third, I removed only the following variables as they were not significant and variable importance was also seemed very low: *weekdays, fin loan, housing, cons.conf.idx, campaign_ln, previous, marital.*

There is a great improvement in the model fit with the inclusion of more and more variables, however with that complexity also increases. Model summary with all sequential steps is available in my github repo, under *lpms_comparison.html*. The last model's variables (later on referred to as *reduced model*) are the ones that I will use going forward as that is the simplest model, with a realtively good fit.

Logit and Probit

One of the shortcoming of LPM is that it can predict probabilities below 0 and above 1. In this case it was a significant issue, as almost 12k records were predicted with such unrealistic predictions. There are ways to overcome this one of them is to turn to logit and probit models. The difference between the predictions of LPM, logit and probit models in my data are visualised in the Appendix.

Choosing the final model came down to three options, the last LPM which had some dropped variables, logit and probit. In order to be able to make a choice between them I looked at model statistics that indicate how good a fit each model can provide. To do this I looked at the R^2 and a pseudo R^2 that can be used when evaulating logit and probit models. I also added a Brier score and checked if these variables had bias. These are available in Table 2 with the addition of a simple LPM included that has demographic variables only - to see the improvement. All models performed quite well, but the probit was a tiny bit more convincing (higher McFadden R^2 and lower Brier score), that is why my model choice will be the **probit model**.

Model	R2	R2.adjusted	McFadden.R2	Brier score	Bias
Simple LPM	0.04	0.04		0.10	0.00
Reduced LPM	0.35	0.35		0.06	0.00
Logit			0.42	0.06	0.00
Probit			0.43	0.06	-0.00

Table 2: Model fits

Robustness check

To see how robust is the used probit model I used a calibration curve and also checked the accuracy of its predictions compared to a baseline model. Before going further it is important to note that the goal was not prediction, and when mentioning prediction in this document what I mean is model implied dependent variables on the whole dataset, and not predicted outputs of test subsets. The calibration curve can be seen in the 5th figure of the Appendix and it shows that the model predicted outputs are very close to the 45 degree line. This is what we expected, as we saw a very small bias for the probit model - it is now visually proven that this model is well calibrated.

I also checked the accuracy of my model, where accuracy here is defined as how many times did the prediction matched exactly the actual dependent variable. This is a naive approach, but one that is only used to get more comfort over using the probit model. The goal here for the model to outperform a baseline model, that is a prediction of 0 for every observation. These seemed a sensible baseline given the low proportion of successfull campaigns in the data (ca. 11%). And the probit beat the baseline as the accuracy of probit is: 91.23%, while the accuracy of baseline is: 88.87%, so with using the model we can improve 2.36 percentage points. This is a good improvement, but not a great one, and there would be room for improvement if the model was to be used for prediction. The confusion matrix in the 6th figure of the appendix is also available to see the number of false positives and false negatives.

Results

The below regression summary summarises the coefficient estimates (column = Model 1) for the probit model and I also included the transformed marginal differences for easy interpretation (column = Model 2). I will evaluate coefficients with 5% significance:

	Model 1	Model 2
Constant	24.482**	
	(7.568)	
age	-0.039**	-0.004**
	(0.006)	(0.001)
age_sq	0.000**	0.000**
	(0.000)	(0.000)
jobblue-collar	-0.178**	-0.018**
	(0.038)	(0.004)
jobentrepreneur	-0.132	-0.013*
	(0.068)	(0.007)
jobhousemaid	-0.094	-0.010
	(0.081)	(0.009)
jobmanagement	-0.042	-0.004
	(0.048)	(0.005)
jobretired	0.021	0.002
	(0.068)	(0.008)
jobself-employed	-0.051	-0.005
	(0.064)	(0.007)
jobservices	-0.140**	-0.014**
	(0.047)	(0.004)
jobstudent	0.103	0.012
	(0.070)	(0.009)
jobtechnician	-0.023	-0.002
	(0.036)	(0.004)
jobunemployed	0.001	0.000
	(0.071)	(0.008)
high.ed	0.116**	0.012**
	(0.028)	(0.003)
seasonspring	0.124*	0.013*
	(0.049)	(0.005)
seasonsummer	0.438**	0.049**
	(0.043)	(0.005)
seasonwinter	0.006	0.001
	(0.118)	(0.016)
contacttelephone	-0.300**	-0.031**
	(0.038)	(0.004)
duration	0.004**	0.000**
	(0.000)	(0.000)
duration_sq	-0.000**	-0.000**
	(0.000)	(0.000)
poutcomenonexistent	0.266**	0.026**
	(0.036)	(0.003)
poutcomesuccess	1.068**	0.173**
	(0.052)	(0.012)
euribor3m	0.425**	0.046**
	(0.049)	(0.005)
nr.employed	-0.009**	-0.001**
	(0.001)	(0.000)
cons.price.idx	0.204**	0.022**
	(0.053)	(0.006)
emp.var.rate	-0.486**	-0.052**
	(0.045)	(0.006)
Num.Obs.	38234	38234

* $p < 0.05$, ** $p < 0.01$

Demographical variables

Most of the demographical variables are significant. For age we included the quadratic form, in which we can see that the *age_sq* is positive so the function is convex, but other than that, the coefficients don't have a clear interpretation. Significant *high_ed* has a marginal difference of 0.012, meaning that on average higher educated people have 1.2 percentage point higher chance to subscribe for a term loan than people without higher education ceteris paribus. For jobs, the reference category was people working in admin positions and only services, enterpreneur and blue-collar workers had significantly different chances (coefficients estimates were all negative) at 5% significance.

Campaign specifics, and time variables

We have many significant variables in these categories and interpretation of all of these variables would be hard, but there are a few things that are worth noting. When it comes to the seasons, people contacted in the summer were associated with a higher probability to subscribe compared to the ones contacted in autumn if we were to keep other variables unchanced. Duration is also in quadratic form, but based on variable importance it is the single most important variable in the analysis. People who were contacted with success in a previous campaign had 17.3 percentage points higher chance to subscribe than those who failed to do given every other variable is the same. This is a very important finding,it shows it is worth talking to people in another campaign if they successfully subscribe in an earlier one. Similarly interpreting the contact variable, it is worth contacting people through cellular phones.

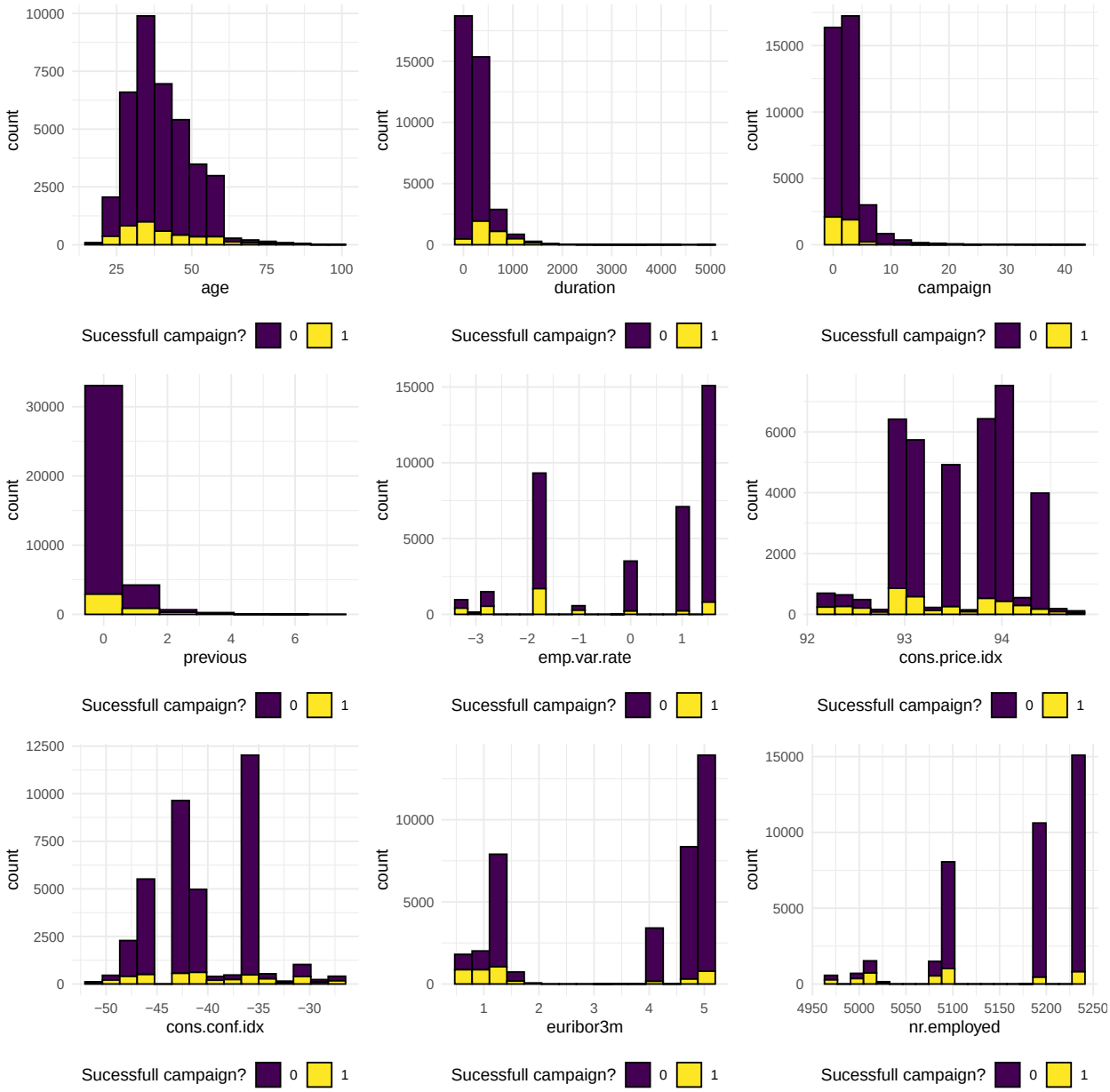
Macroeconomic variables

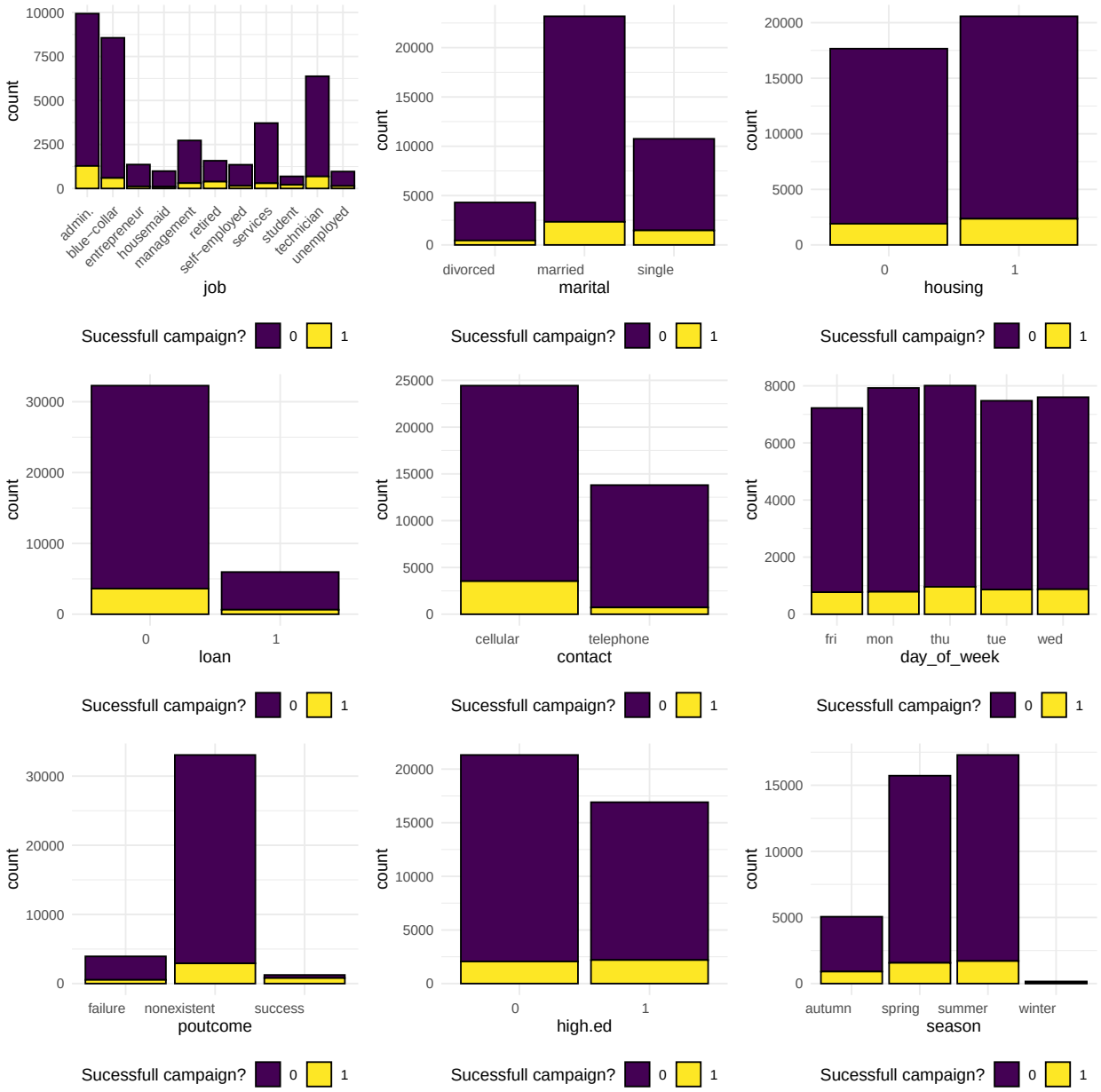
Each variable was significant even at 1%, which shows that it is worth including such variables in our models. This was also proven earlier when choosing the model, as they improved the fit of the model quite well. Without giving a proper interpretation from the marginal coefficients it is visible that higher consumer price index and euribor rate are associated with higher chances of a person subscribing for a term loan while low employment variation meant higher chance for subscription as well. These are pretty much in line with general expectations, if the interest is high and unemployment rate decreases, people are more willing to put their money in the bank. But all in all these are rather important from a perspective of prediction.

Summary

In this paper I analysed the relationship of information known about the clients and the probability of someone subscribing for a term loan in a telemarketing campaign. I used linear probability models and furthermore logit and probit models to uncover this relationship in details on data made public by a Portuguese bank. My results imply that there are lessons we can learn from the data. Based on the data, those contacted before with success might have a higher possibility to subscribe again in another one. It might also make sense to target people with a higher educational background through cellular phones rather than telephone. Macroeconomic variables also play a role, so if the anyone wishes to develop models for prediction should take these into account. I believe my findings can be useful for marketing managers to improve the efficiency of campaigns and reduce costs of their operations.

Appendix


count
age
duration
campaign
previous
emp.var.rate
cons.price.idx
cons.conf.idx
euribor3m
nr.employed
Sucessfull campaign?
0
1


count
job
admin.
blue-collar
entrepreneur
housemaid
management
retired
self-employed
services
student
technician
unemployed
marital
divorced
married
single
housing
loan
contact
cellular
telephone
day_of_week
fri
mon
thu
tue
wed
poutcome
failure
nonexistent
success
high.ed
season
autumn
spring
summer
winter
Sucessfull campaign?
0
1

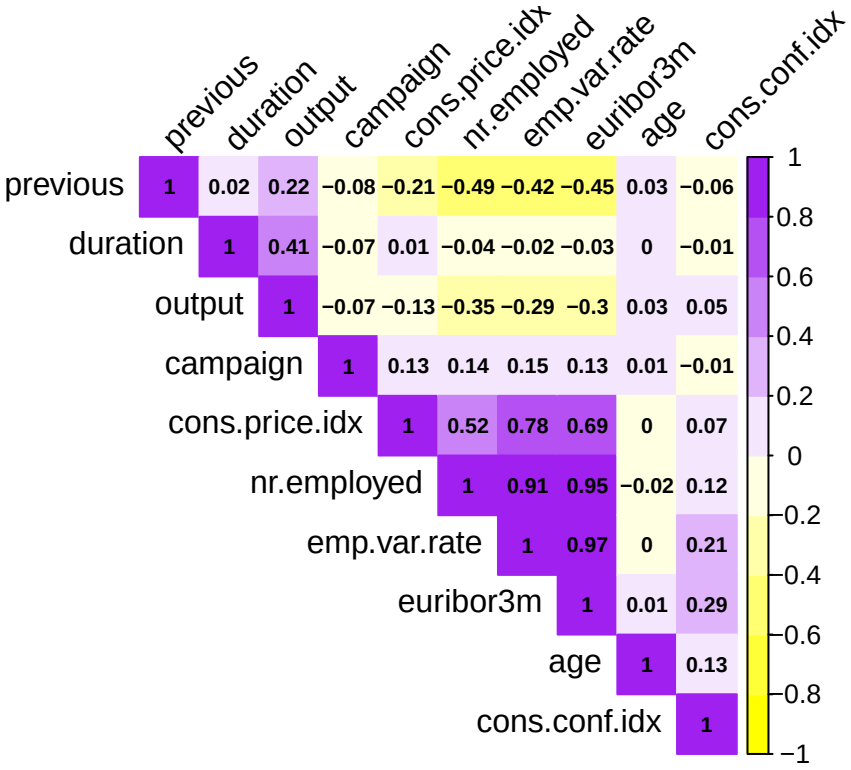
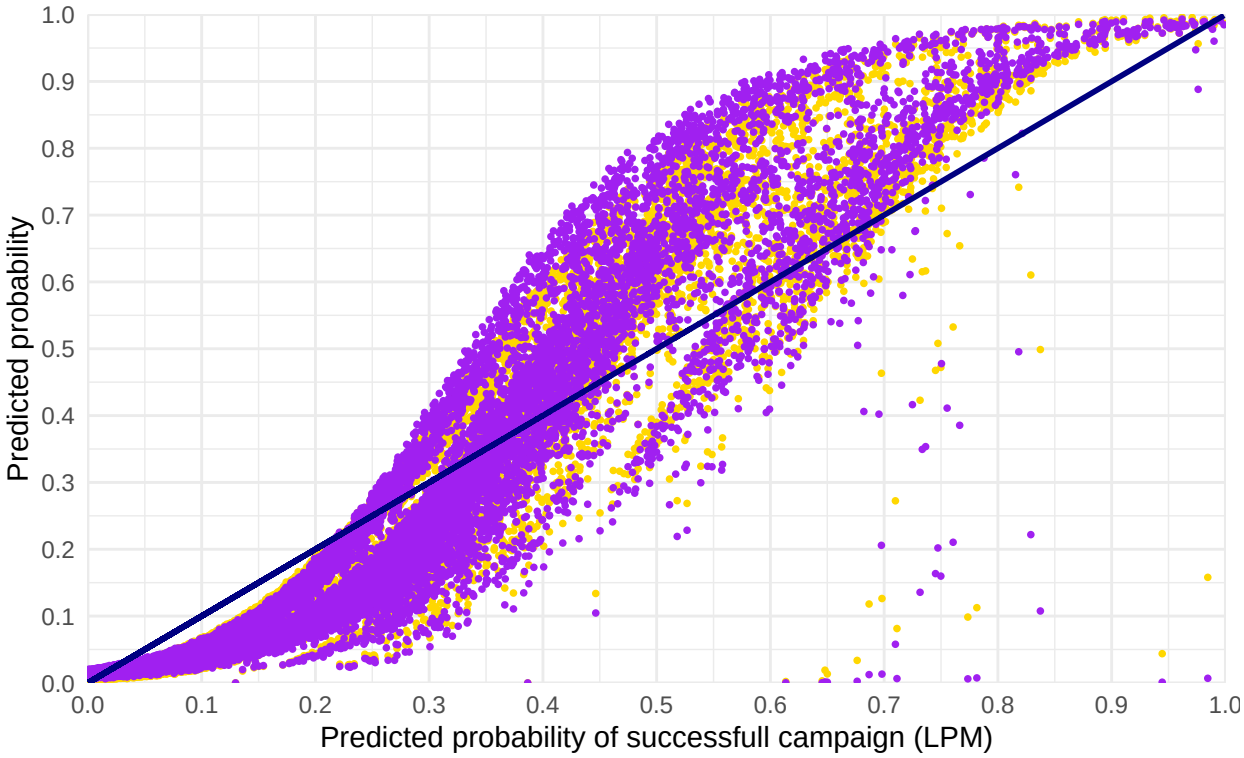
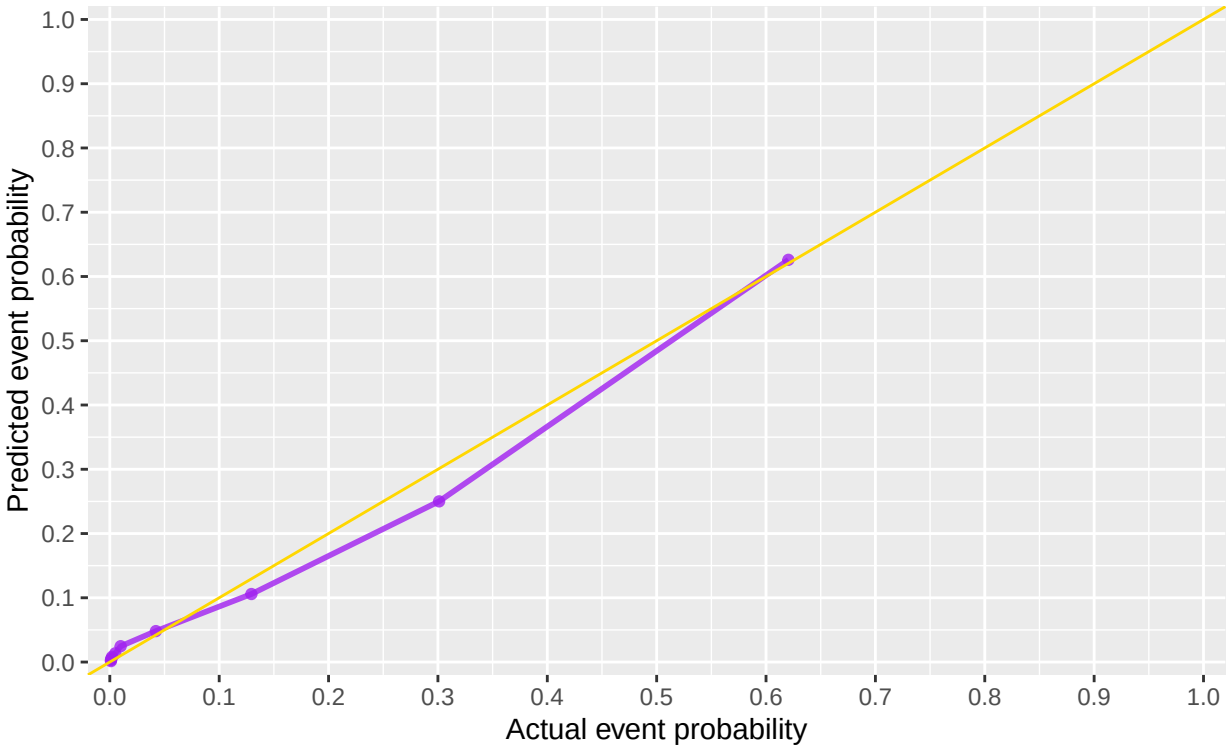
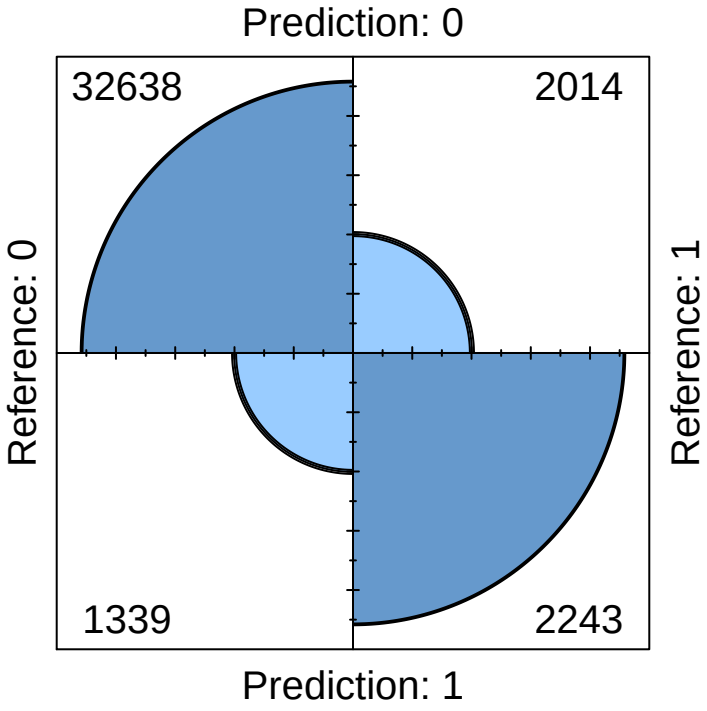

previous
duration
output
campaign
cons.price.idx
nr.employed
emp.var.rate
euribor3m
age
cons.conf.idx
1 0.02 0.22 −0.08 −0.21 −0.49 −0.42 −0.45 0.03 −0.06
1 0.41 −0.07 0.01 −0.04 −0.02 −0.03 0 −0.01
1 −0.07 −0.13 −0.35 −0.29 −0.3 0.03 0.05
1 0.13 0.14 0.15 0.13 0.01 −0.01
1 0.52 0.78 0.69 0 0.07
1 0.91 0.95 −0.02 0.12
1 0.97 0 0.21
1 0.01 0.29
1 0.13
1
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Table 3:

	Dependent variable:		
	output		
	(1)	(2)	(3)
age	−0.02*** (0.001)	−0.01*** (0.001)	−0.01*** (0.001)
age_sq	0.0003*** (0.0000)	0.0001*** (0.0000)	0.0001*** (0.0000)
jobblue-collar	−0.04*** (0.01)	−0.02*** (0.004)	−0.02*** (0.004)
jobentrepreneur	−0.04*** (0.01)	−0.02** (0.01)	−0.02** (0.01)
jobhousemaid	−0.03** (0.01)	−0.01 (0.01)	−0.01 (0.01)
jobmanagement	−0.02*** (0.01)	−0.01 (0.01)	−0.01 (0.01)
jobretired	0.02 (0.01)	−0.0004 (0.01)	−0.0005 (0.01)
jobself-employed	−0.02** (0.01)	−0.01 (0.01)	−0.01 (0.01)
jobservices	−0.04*** (0.01)	−0.02*** (0.01)	−0.02*** (0.01)
jobstudent	0.13*** (0.01)	0.03** (0.01)	0.03** (0.01)
jobtechnician	−0.02*** (0.01)	−0.003 (0.004)	−0.003 (0.004)
jobunemployed	0.02** (0.01)	−0.002 (0.01)	−0.002 (0.01)
maritalmarried	0.01* (0.01)	0.001 (0.004)	
maritalsingle	0.02*** (0.01)	0.002 (0.005)	
high.ed	0.03*** (0.004)	0.01*** (0.003)	0.01*** (0.003)
housing		−0.0001 (0.003)	
loan		−0.003 (0.004)	
day_of_weekmon		−0.01*** (0.004)	
day_of_weekthu		0.002 (0.004)	
day_of_weektue		−0.002 (0.004)	
day_of_weekwed		0.001 (0.004)	
seasonspring		−0.003 (0.01)	−0.003 (0.01)
seasonsummer		0.09*** (0.01)	0.09*** (0.01)
seasonwinter		0.04* (0.02)	0.04* (0.02)
contacttelephone		−0.04*** (0.005)	−0.04*** (0.004)
duration		0.001*** (0.0000)	0.001*** (0.0000)
previous		0.01 (0.01)	
duration_sq		−0.0000*** (0.00)	−0.0000*** (0.00)
campaign_ln		0.003 (0.002)	
poutcomenonexistent		0.05*** (0.01)	0.04*** (0.005)
poutcomesuccess		0.32*** (0.01)	0.32*** (0.01)
euribor3m		0.10*** (0.01)	0.10*** (0.01)
nr.employed		−0.002*** (0.0001)	−0.002*** (0.0001)
cons.price.idx		0.04*** (0.01)	0.04*** (0.01)
cons.conf.idx		0.0004 (0.001)	
emp.var.rate		−0.09*** (0.01)	−0.09*** (0.01)
Constant	0.53*** (0.02)	6.83*** (1.44)	7.70*** (0.83)
Observations	38,234	38,234	38,234
R^2	0.04	0.35	0.35
Adjusted R^2	0.04	0.35	0.35
Residual Std. Error	0.31 (df = 38218)	0.25 (df = 38197)	0.25 (df = 38208)
F Statistic	95.85*** (df = 15; 38218)	573.75*** (df = 36; 38197)	825.10*** (df = 25; 38208)
Note:			*p<0.1; **p<0.05; ***p<0.01


Predicted probability
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Predicted probability of successfull campaign (LPM)


Predicted event probability
Actual event probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0


Prediction: 0
32638
2014
Reference: 0
Reference: 1
1339
2243
Prediction: 1