Question 1: Modeling Binary Choices

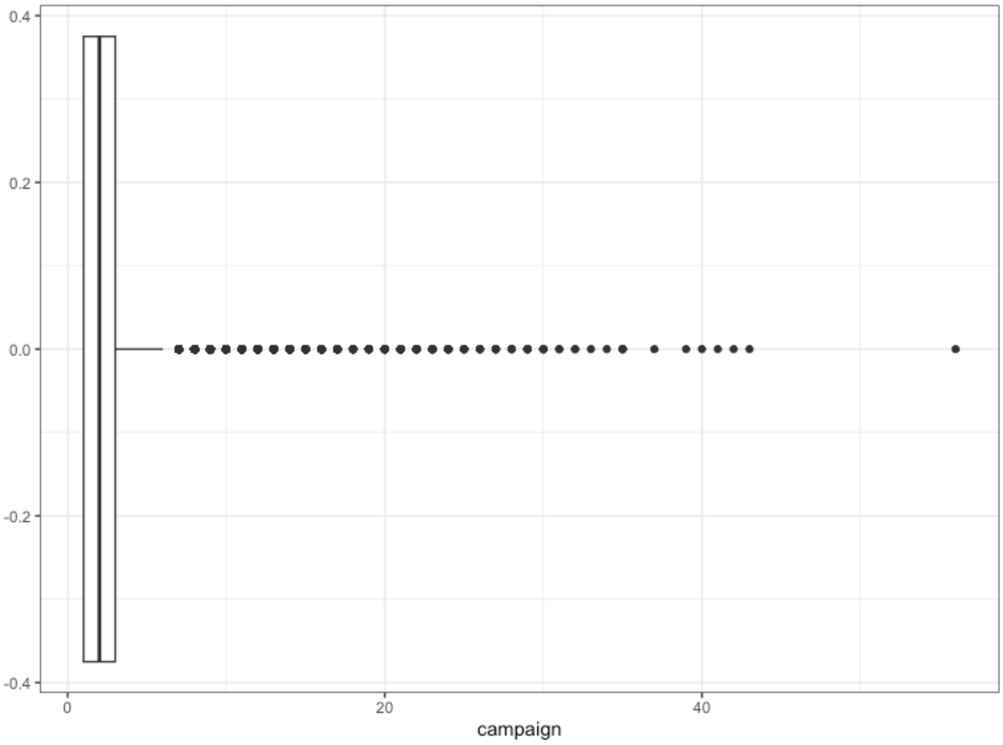
1. Explore the data

This step can be observed in detail on our code in the appendix.

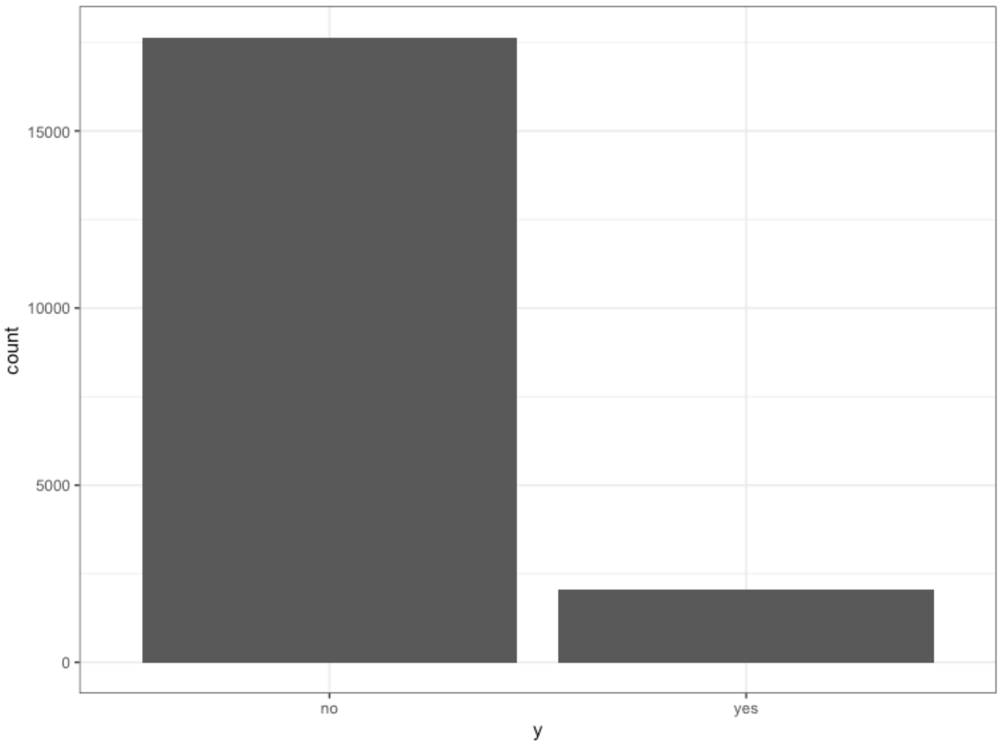
We looked at the data set, used the summary function. We converted the datatype of the columns holding categorical values into factors. Because we found a lot of "unknown" values, we decided to assign the NA value to them. To make more sense, we also assign the NA value to the value "999" in the column "pdays". With these decisions, we found ourselves with 20% of NA's in the entire dataset. We decided to show whether a client was contacted or not, rather than showing the number of days since the client was last contacted.

We made several plots/graphs:

- to see how the variables are correlated with the output variable.
- to see how the variables are correlated together: we notice here that several of the variables are highly correlated and should not be included in the same linear models due to issues with multicollinearity.
- to see outliers, extreme values: we notice here several outliers (the dots), we update the data frame by removing them from the "campaign" and "previous" columns. There is no outlier for the "employment variation rate", the "consumer price index", the "euribor 3-month rate", the "number of employees" and the "consumer confidence index" attributes. Here is for instance the plot that highlights the outliers of the "campaign" variable.



- to see how the outcome variable "y" was distributed:



2. Investigate the customer’s decision to subscribe a term deposit

2.a. Dependent and independent variables

Our DV is whether the client subscribed a term deposit. This variable is called “y”.

We choose five IVs:

- The contact communication type: “contact”,
- The marital status: “marital”,
- The type of job: “job”,
- Whether or not the client has a personal loan: “loan”,
- The number of contacts performed during the campaign for a specific client: “campaign”.

These are the modified variables with the fewest NA’s (“unknown” value). These are also not highly correlated variables. Furthermore, these variables allow us to have information about both the marketing campaign and about the client more personally.

2.b. Hypotheses

We can choose the independent variable “campaign” for the hypothesis. We can expect both negative and positive effects on “y” from it.

For instance, contacting a customer too many times during the marketing campaign could have a repulsive effect and rather discourage them from carrying out an action, especially here to subscribe a term deposit in the bank.

Conversely, contacting a customer very few times, especially without a follow-up, could lead to the customer forgetting the marketing campaign and of course not subscribing a term deposit. The customer might also tend to take out a term deposit at another bank, thus conflating the marketing campaign with another bank.

On the other hand, a well-balanced number of contacts and reminders can have a positive effect on “y”.

2.c. Choose (an) appropriate method(s) and explain your choice

Because the exercise focuses on predicting the “y” output, which is binary: whether the client subscribed a term deposit. To do so, we are going to create a binomial logistic regression model using the variables we chose previously (contact, marital, job, loan and campaign) as predictors. This model is called “model a”.

We split our modified data frame into estimation (train) and validation (test) samples with the functions “training” and “testing”. We then create a model with the “glm” function.

2.d. Robustness check/Model validation

To validate our model, we are going to create a second model, called “model b”. We would like to focus on more external attributes to see how it affects the outcome. We decided to look at other attributes, both linked to the marketing campaign itself and also the personal characteristics of the client: "poutcome", "previous", "education", "age" and "day_of_week".

Next, we will estimate our model a and model b. We previously split our data frame into train and test samples. We run our models and compute validation measures (accuracy rate, AIC), using the test sample.

The model a has an accuracy rate of 84% and the model b an accuracy rate of 90%. The best model seems to be the model b, because it has the lower AIC and the better accuracy.

2.e. Main model estimation

As we saw it in the previous question, the best model is our model b. Using the entire “subscribe” data set and the glm function, we re-estimate it. It is called “final model” on our code.

The Akaike Inferior Criterion of our final model is 12,524.080.

2.f. Report findings, interpret the results, and formulate conclusions/discussions

We can observe that several variables are significant on the output “y”. A “non-existent” or a “success” outcome of the previous marketing campaign are really significant ($p<0.01$). A non-existent outcome of the last marketing campaign impacts negatively “y”, while a successful outcome of the last marketing campaign has a positive influence on the output. The attribute “education” has also a great effect on the outcome: especially with a university degree ($p<0.01$), then there is a negative influence on “y”, if the customer has a 6 ($p<0.05$) or 9 years education ($p<0.1$). If its education is unknown or if the customer is illiterate, it has a positive impact on whether or no the client will subscribe a term deposit ($p<0.05$). The age will also greatly influence the output (p<0.01). Finally, if the last contact to the client was on a Thursday, it impacts positively the “y” (p<0.1)

2.g. Write a short part about managerial implications: Imagine that you are reporting these findings to the management team of this company. How would they turn these insights into business actions?

Concretely, our findings could be very interesting for a more efficient management of the marketing team. We now know that a client previously contacted on a Thursday is more likely to subscribe to a term deposit. It is therefore time to focus on this type of customer in the marketing campaign targets. Moreover, Thursdays can become the main day to conduct marketing campaign and contacting clients.

Similarly, it now makes sense to look at a customer who has previously been successfully contacted. It is also possible to look at older customers, with a university degree or who are illiterate.

With our studies, it’s also important to notice that contacting too many times a client can bothered him, so we need to take this information in consideration.

As a result of these analyses and to turn these insights into business actions, we can draw up a typical portrait of the customer to be targeted in priority, compared to other clients, in our marketing campaigns, to favor a positive outcome (they subscribe a term deposit).

2.h. If you had a chance to come back in time and collect more data, what kind of additional data would you collect to improve the model? Explain why.

To improve the model, we could collect more data.

Especially here, we now know that Thursday is the best day to contact a customer to sign up for a long-term deposit. Why not look more closely at the ideal time of the day to contact a customer? Collecting data during call times would therefore be a relevant idea to refine our analysis.

We could also look deeper into some personal data like for example the gender of the client.

It might also be interesting to focus on the person in charge of the marketing campaign. Who is the person who contacts the customer? In particular, one could collect data on the number of years of professional experience this person has or on the quality of the feedback he or she has received within the company.