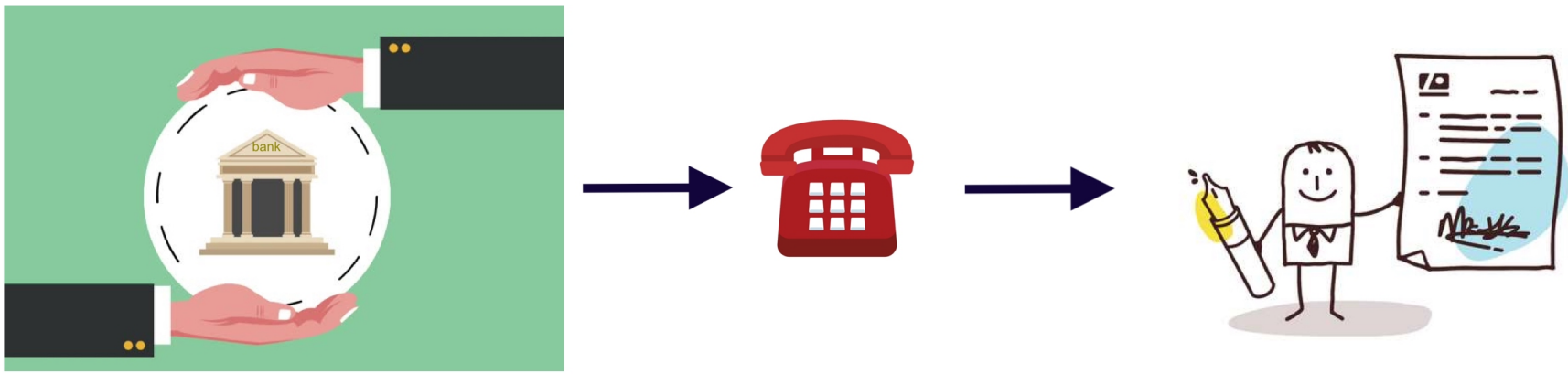
Bank Marketing Campaign - Intelligent Customer Profiling

Franscesca de Robertis, Federico Matteo
University of Padua
Master degree in Data Science
Statistical Learning

Problem statement



Problem statement

The data-set has been downloaded from the UCI Machine Learning repository website.
Its aim is to predict whether a person will sign a deposit. We need to

1. Find the best strategy to improve for the next marketing campaign
2. Enhance the financial institution effectiveness for future marketing campaigns

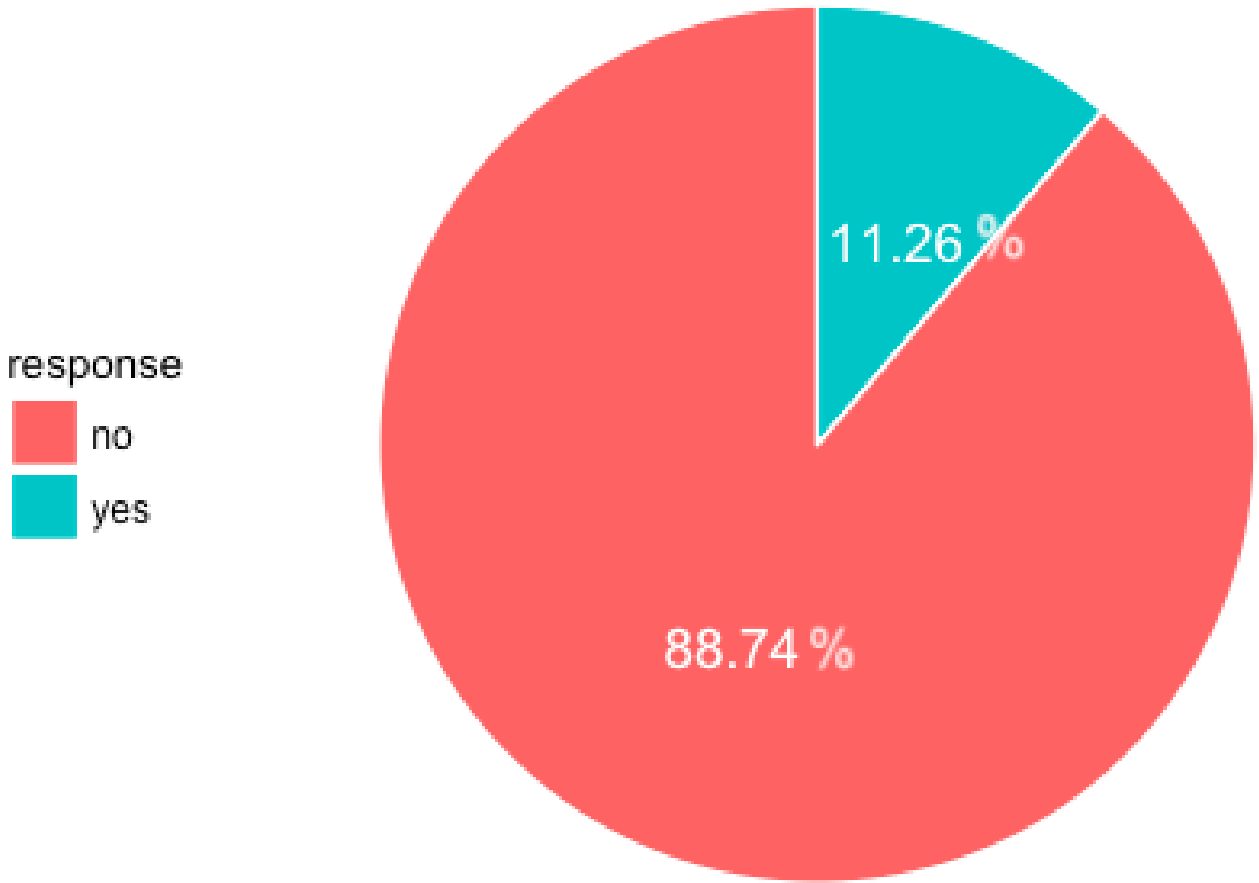
In order to answer these questions, we have to analyze the last marketing campaign the bank performed and identify the patterns that will help us in finding conclusions to develop future strategies.

Attributes analysed

Categorical Variables	Quantitative Variables	Economic indices
Day of the week	Age	Emp.var.rate
Job	Campaign	Cons.price.idx
Marital	Duration	Cons.conf.idx
Education	pdays	Euribor3m
Default	Previous	Nr.employed
Housing		
Loan		
Month		
Contact		
Poutcome		

Study of the response

The data-set is highly unbalanced in the response, we keep this aspect in the analysis and in the stratified sampling procedure.



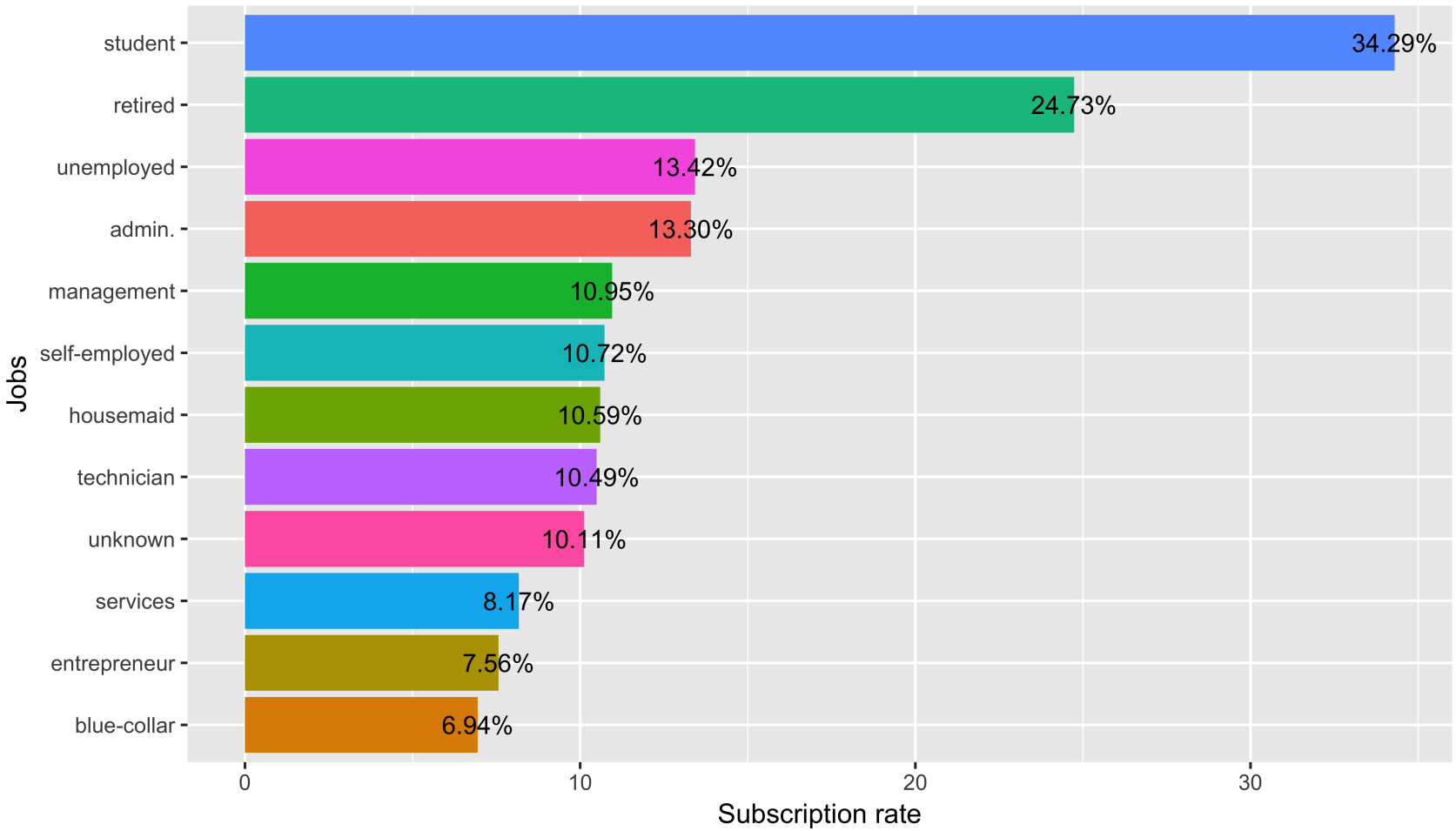
Before starting the analysis

1. We keep the same ratio of the response, i.e. 11.27% of yes.
2. In addition, from the beginning we split up the data set in this manner:
    - 50% of the data for the Training Set
    - 25% of the data for the Validation Set
    - 25% of the data for the Testing set

And perform the graphical analysis and modelling only using the training set.

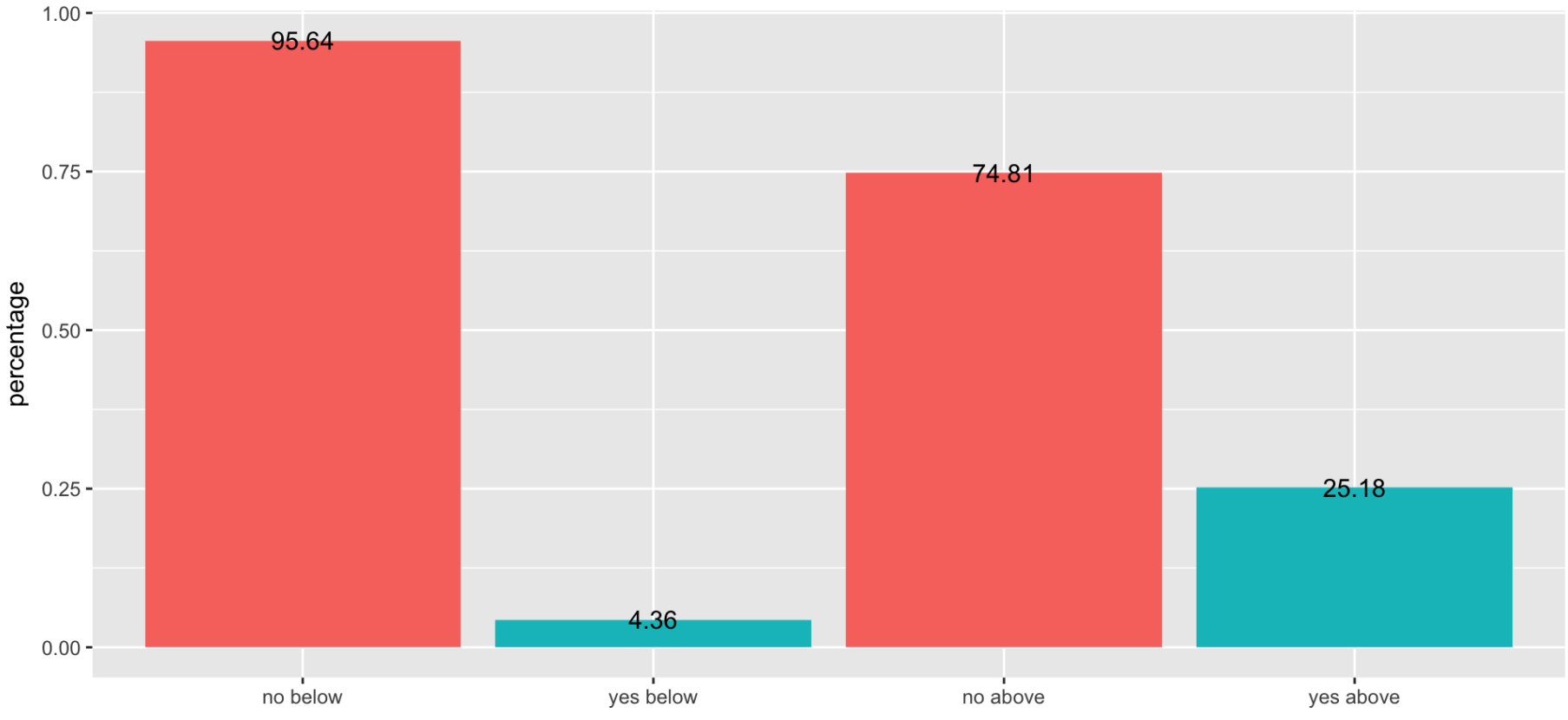
Study of categorical variables

Bar-plot of jobs conditioned to the subscription rate.



Study of quantitative variables

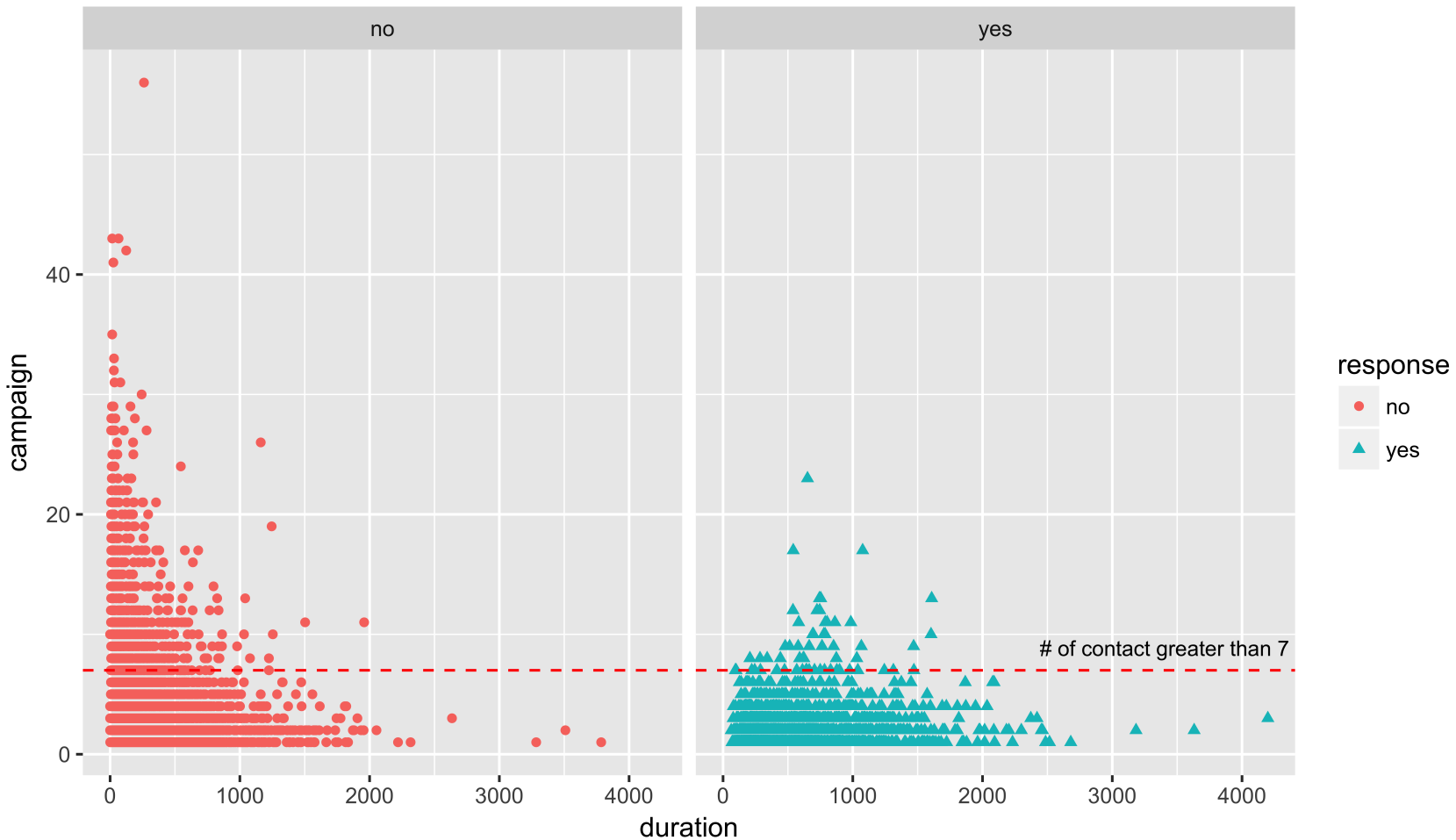
Bar-plot based on average duration



Increase in percantage of people signing a long term deposit when they are above the average duration

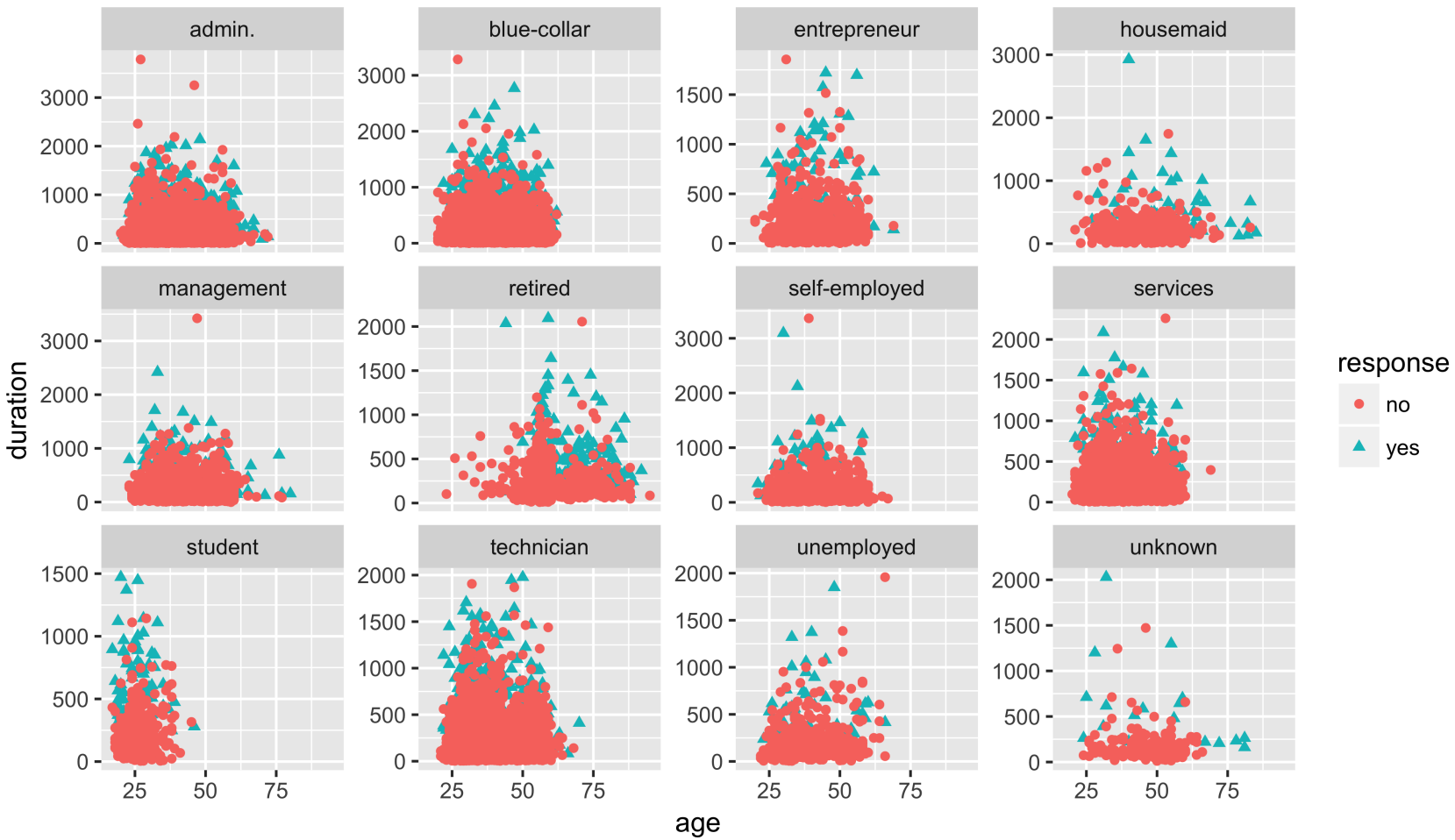
Study of quantitative variables

Campaign versus Duration



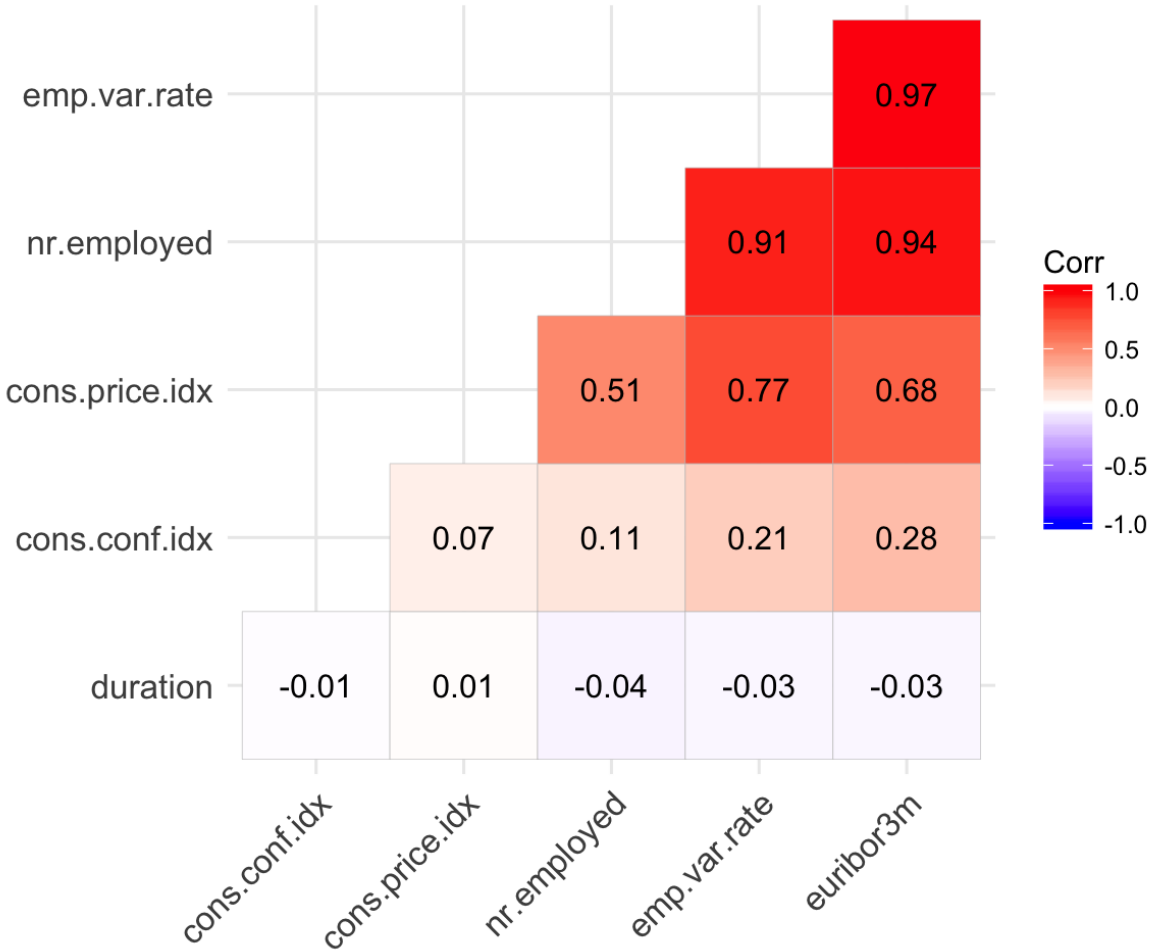
Study of quantitative variables

Job conditioned to duration, age and response



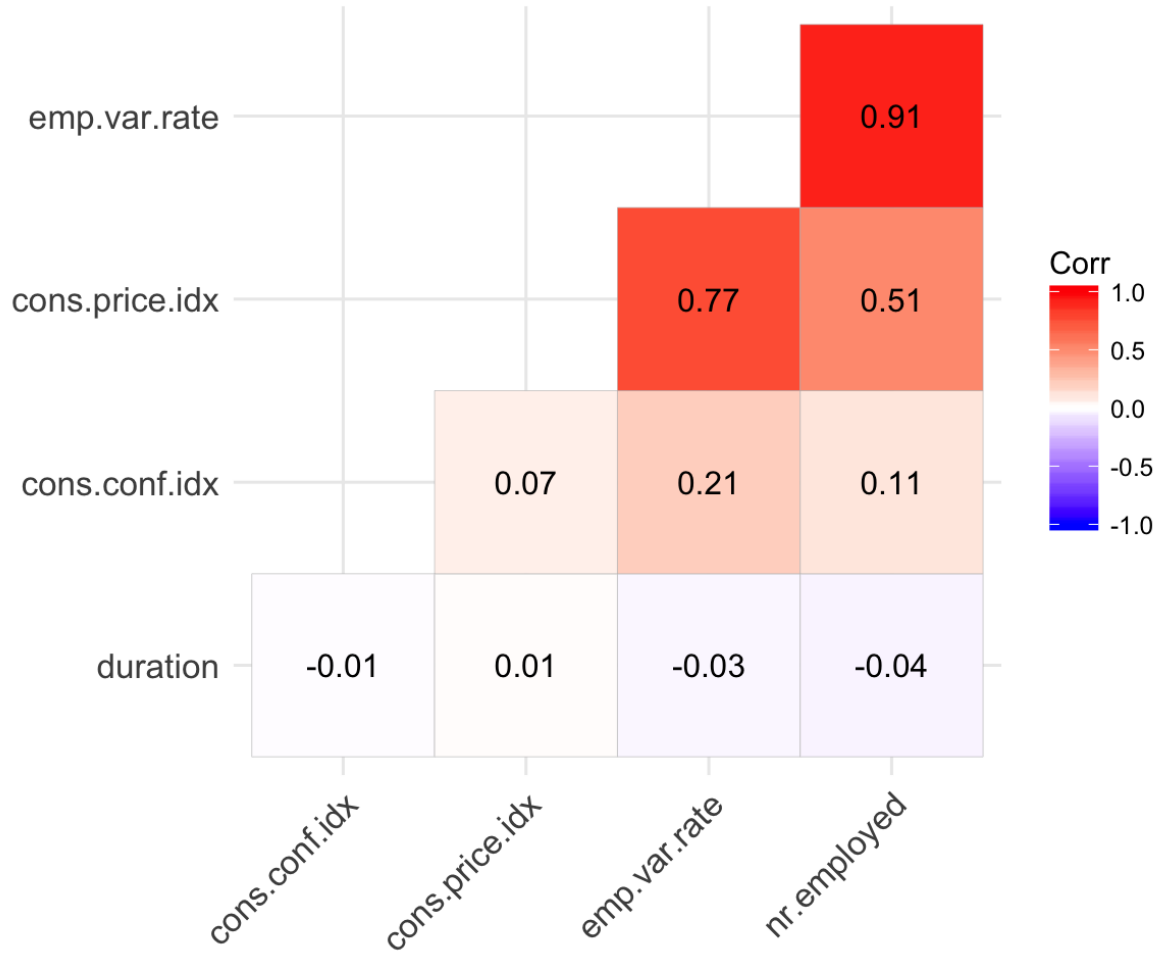
Study of quantitative variables

Correlation plot



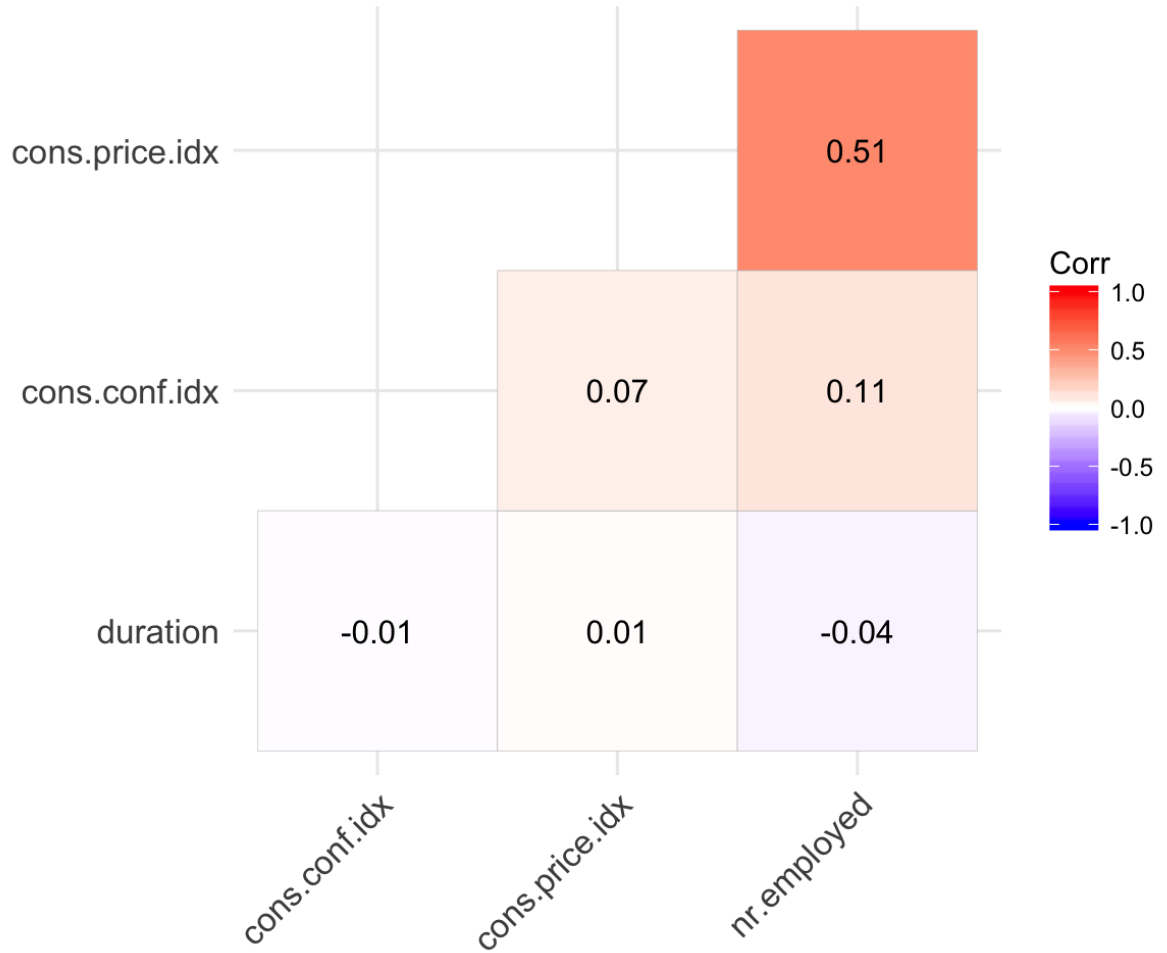
Study of quantitative variables

Correlation plot



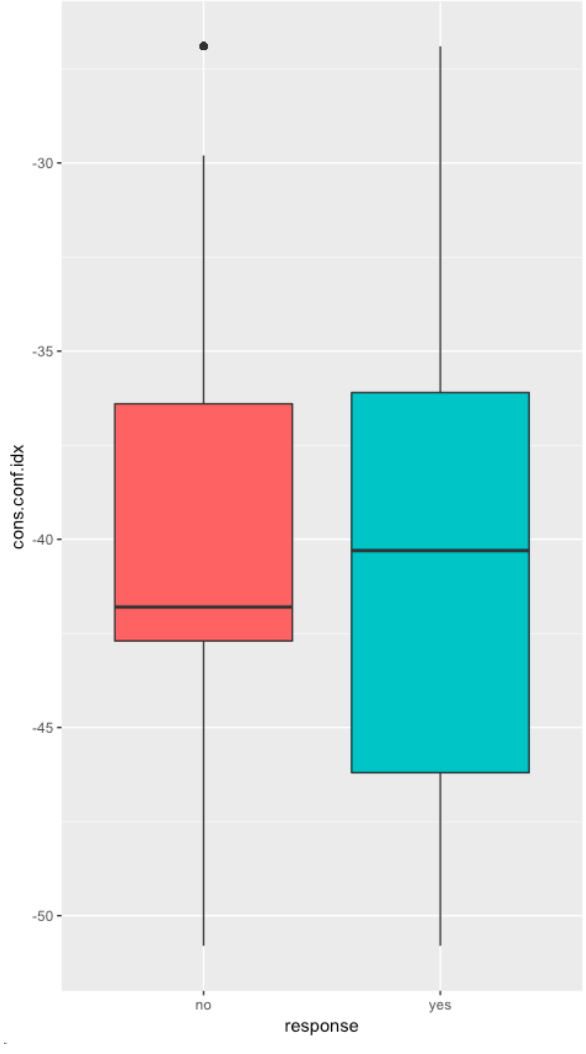
Study of quantitative variables

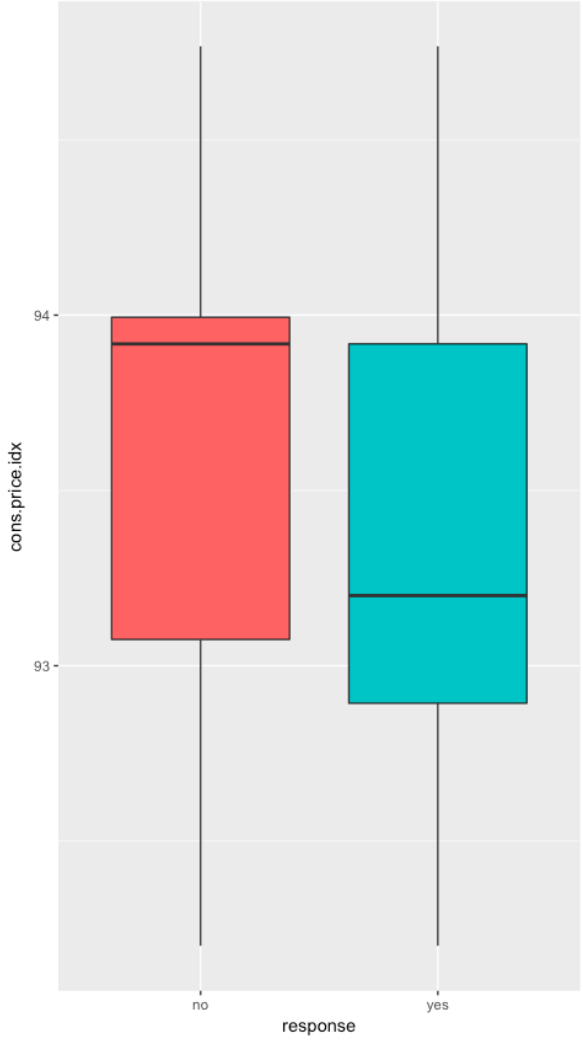
Correlation plot

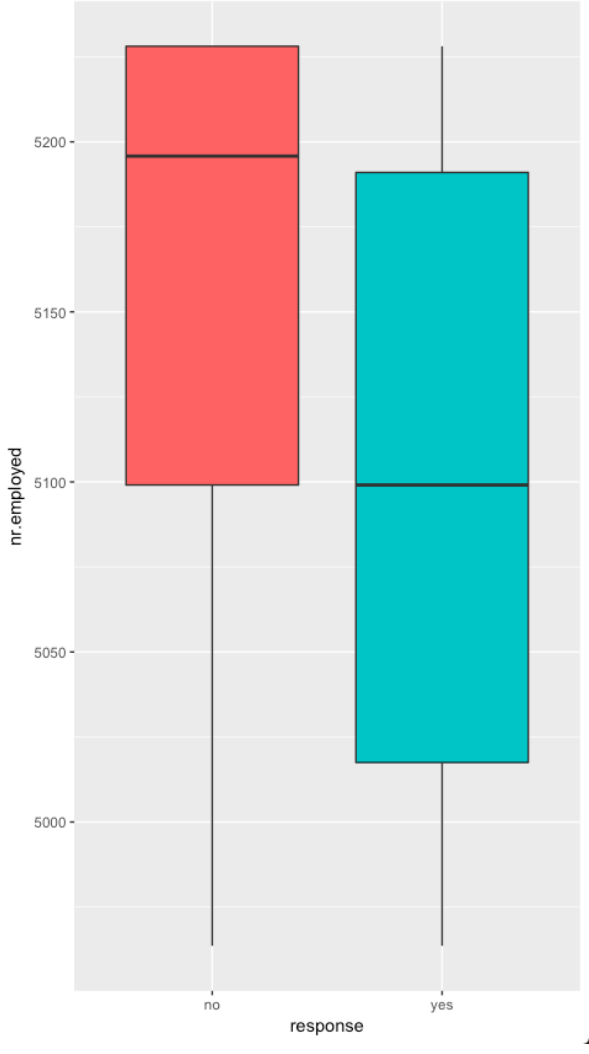


Study of quantitative variables

Box-plots of indices considered in the logistic regression model

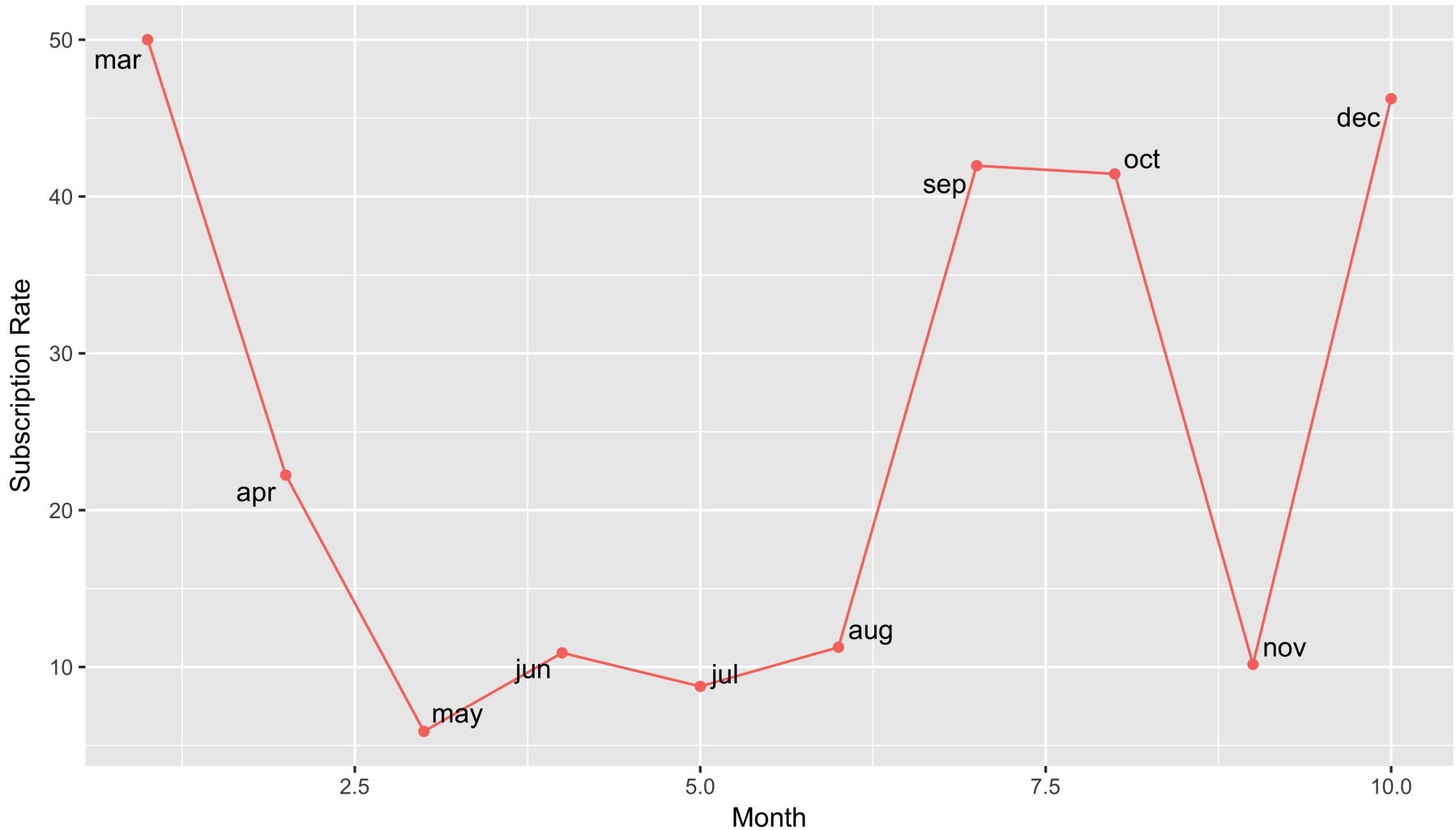






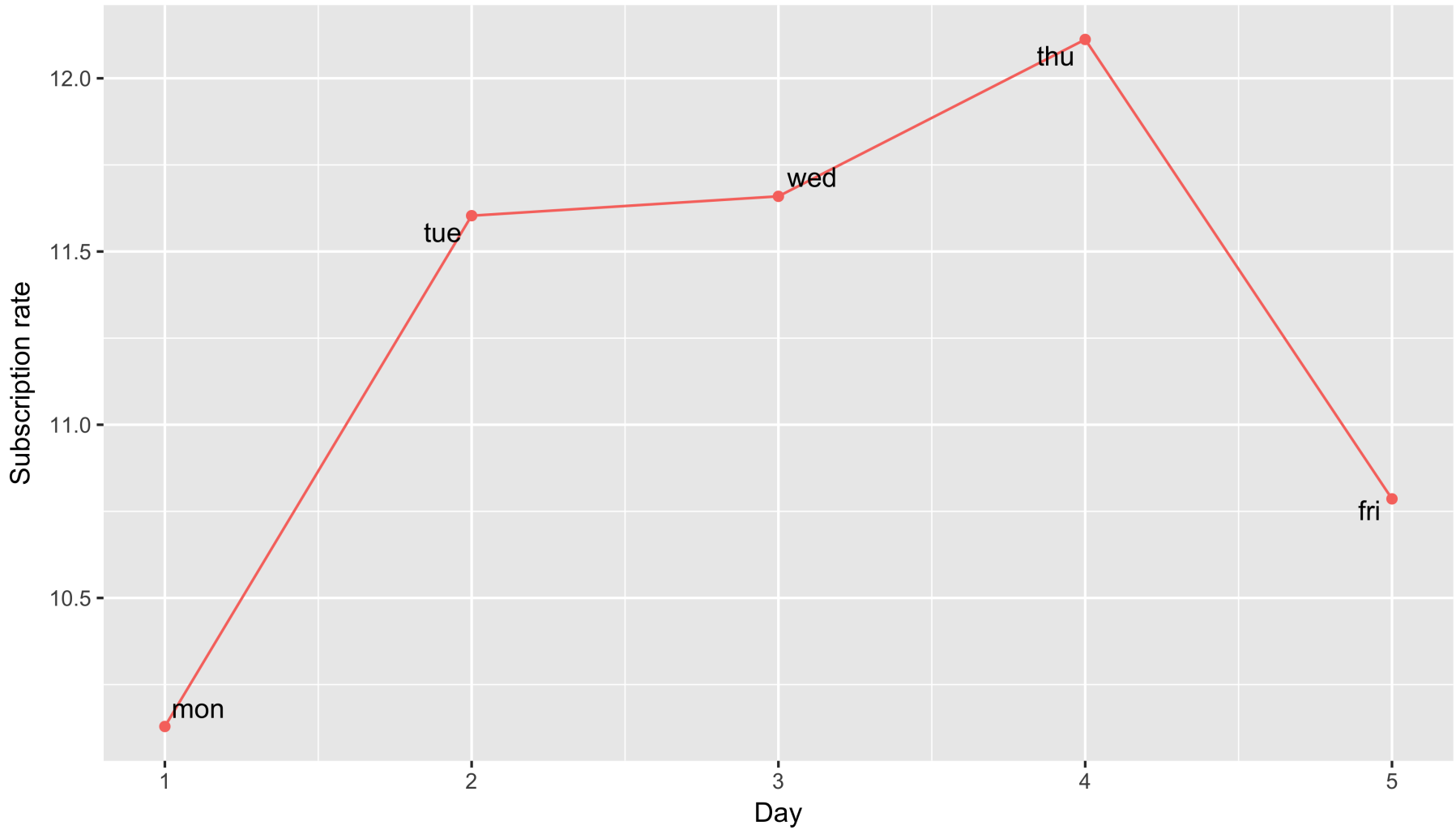
Study of quantitative variables

Subscription rate conditioned on the months



Study of quantitative variables

Subscription rate conditioned on the days



What is classification?

Logistic Classifier

We keep the following variables:

Costrained based meth.	**Exhaustive and greedy meth.**
month	month
contact	contact
poutcome	poutcome
nr.employed	nr.employed
cons.conf.idx.	cons.conf.idx
Default	
Campaign	
Day of Week	
Previous	

Confusion matrix

Model with variables: month, contact, poutcome, nr.employed, cons.conf.idx., Default, Campaign, Day of Week, Previous

		Actual values	
		no	yes
Predicted values	no	8056	476
	yes	1080	684
	Threshold	Specificity	Sensitivity
	0.1629	0.8817	0.5897

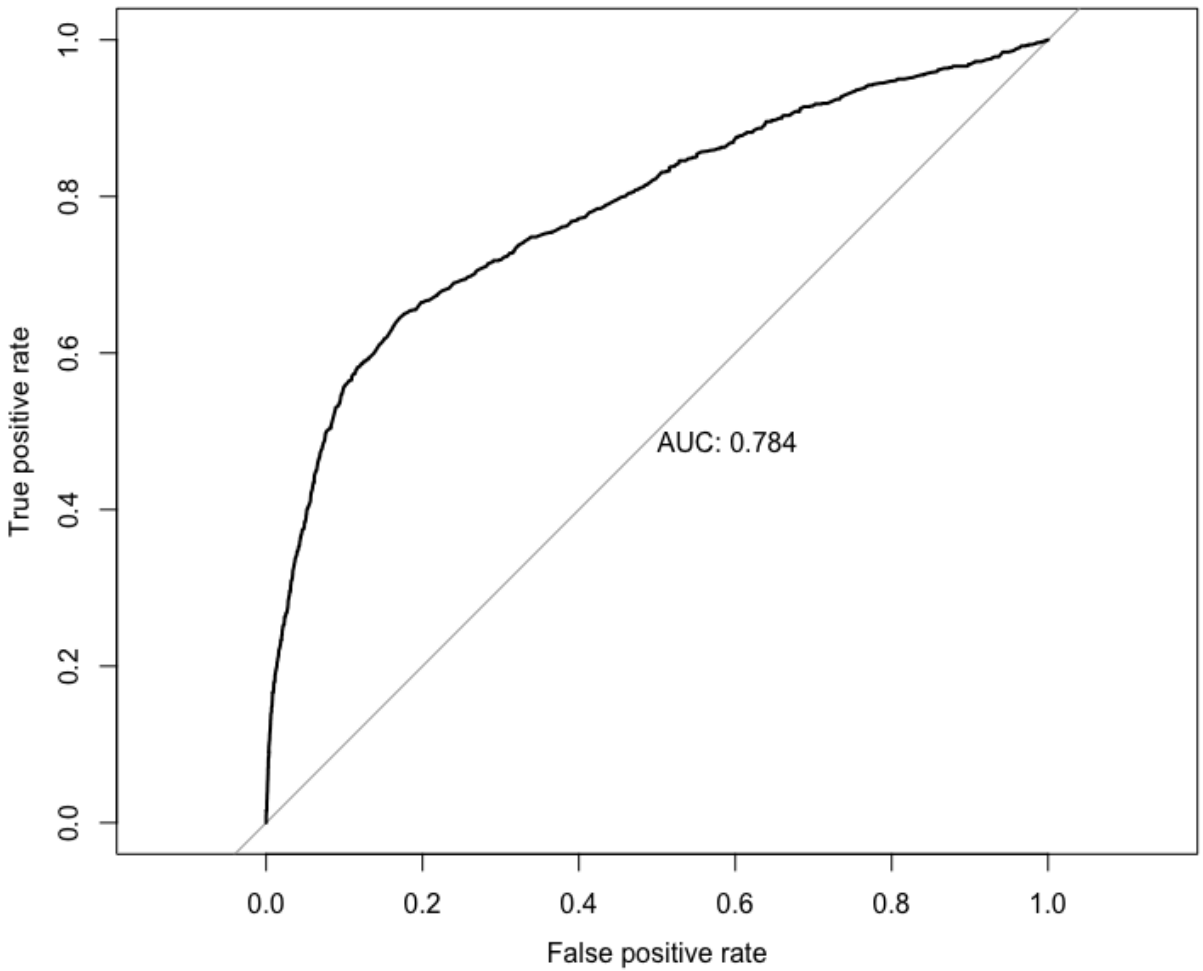
Confusion matrix

Model with variables: month, contact, poutcome, nr.employed, cons.conf.idx

		Actual values	
		no	yes
Predicted values	no	8046	485
	yes	1091	675
		Threshold	Specificity	Sensitivity
		0.1540	0.8806	0.5819

Logistic Classifier

Area Under the Curve (AUC) of the best Logistic Classifier obtained



K-Nearest Neighbors

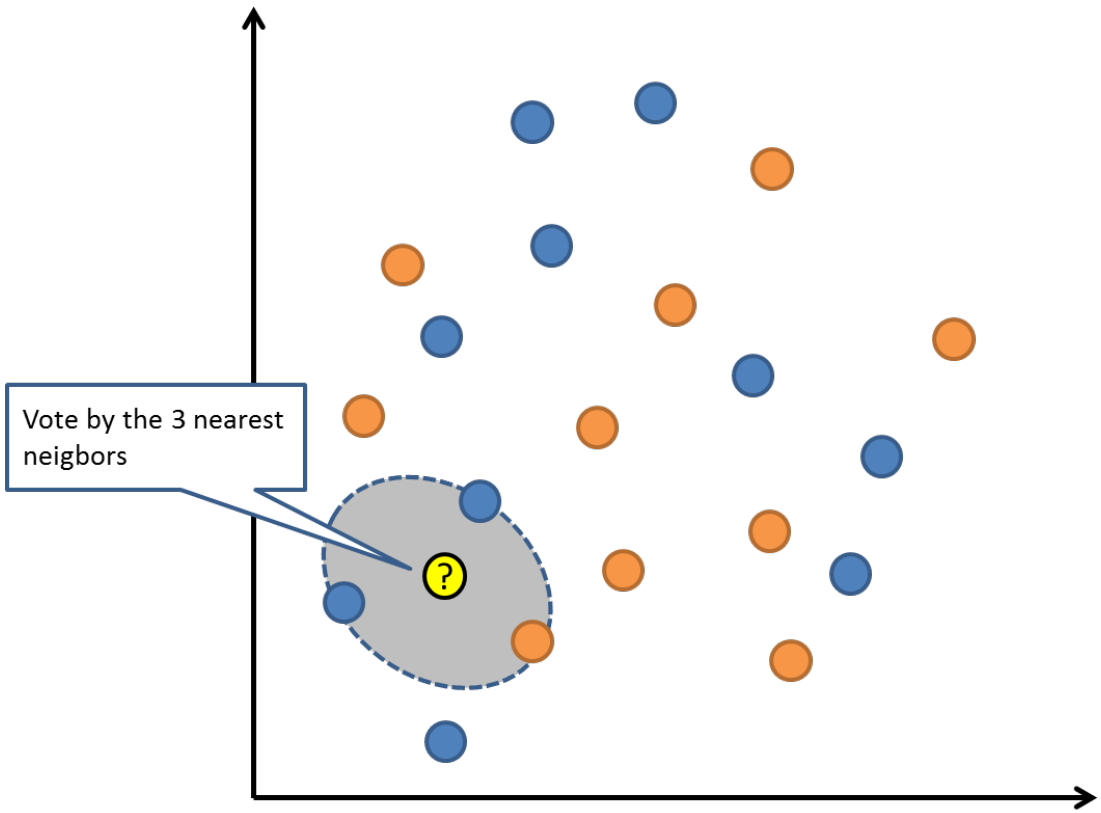
A non parametric classifier with

- A simple idea of functioning
- No coefficients to estimate
- A black box machine: there is no way to understand which variables are important to consider
- Better results in classification (maybe?)

We need to train the model and figure it out

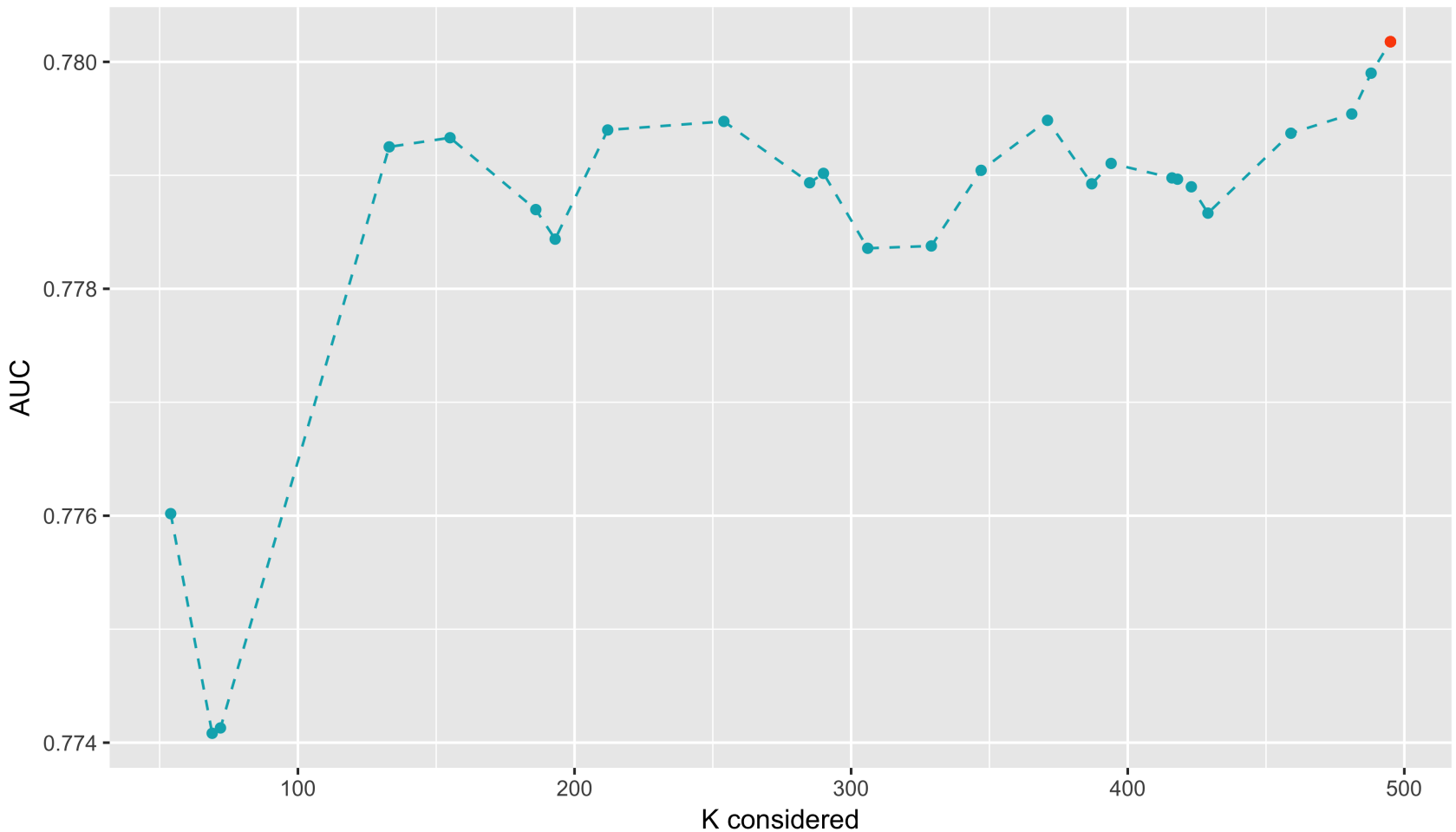
K-Nearest Neighbors

A two dimensional example



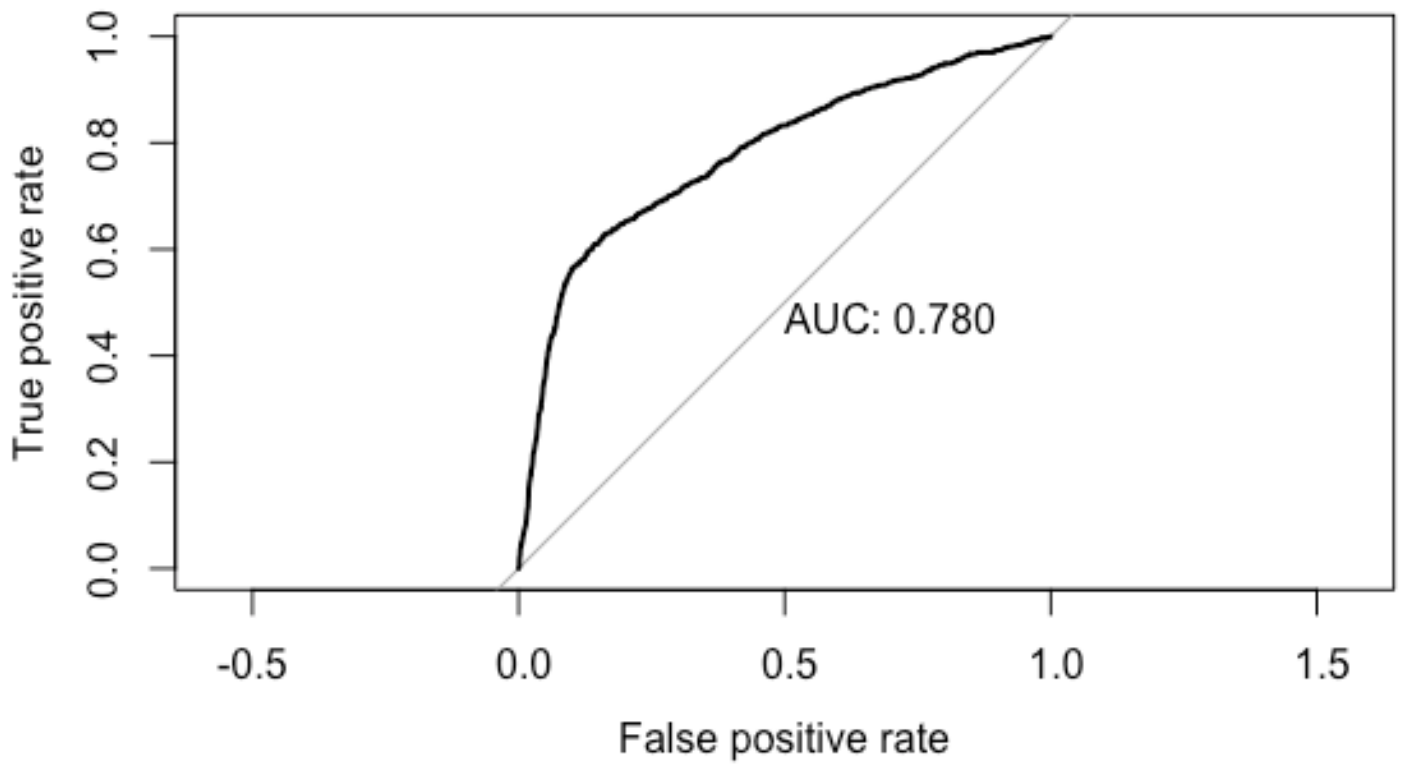
K-Nearest Neighbors

AUC obtained at different k-th values



K-Nearest Neighbors

Area Under the Curve (AUC) of the best knn model obtained



Inference from the model

From the analyses carried out we can infer that the most relevant inputs in predicting the Success rate of a bank direct marketing campaign are the following:

- **months** of August, June, March, May, November, September
- **contact** the person via telephone
- consider the **outcome** of the previous marketing campaign
- **number of employees** - quarterly indicator
- **consumer confidence index** - monthly indicator

Inference from the model

Exploiting the odds ratio, a powerful instrument that can be used when working with a logistic classifier, we can see that:

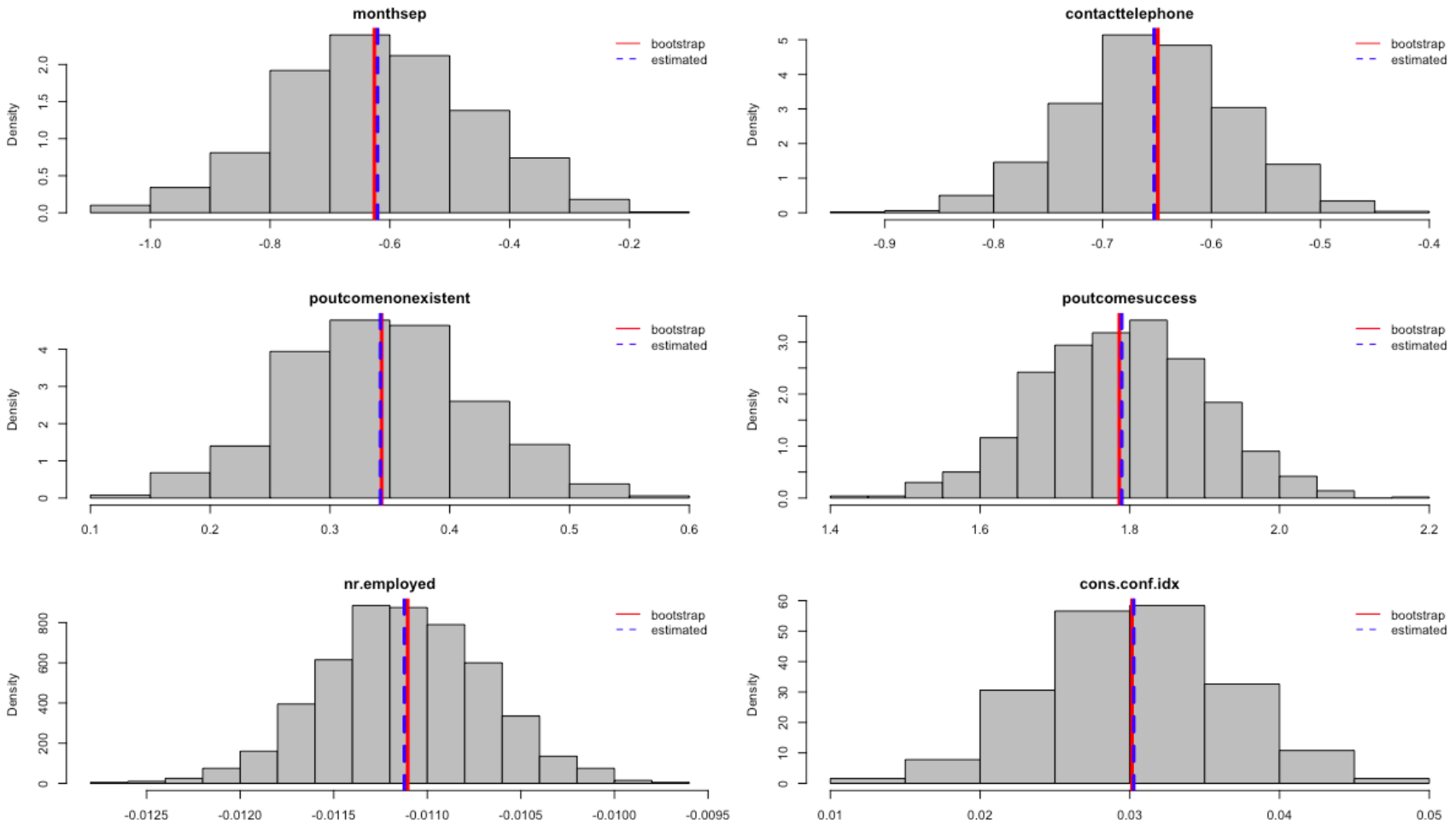
1. In march the probability of success increases of 2.53%
2. If the person has already signed a deposit previously, it is likely to sign a new one with a probability of 5.96%
3. When the number of employees and the consumer confidence index increase by one unit, the probability of success decreases of 0.98% and increases 1.03% respectively

Inference from the graphical analysis

- These results partially confirmed the graphical analysis, in fact the variables **campaign** and **jobs** have not been revealed by the model
- However, we would suggest to consider also **students** and **retired people** as strategic targets
- And **not** to contact a person more than **seven times** in a single campaign

Bootstrap

Histograms of coefficients after performing the bootstrap



Bootstrap

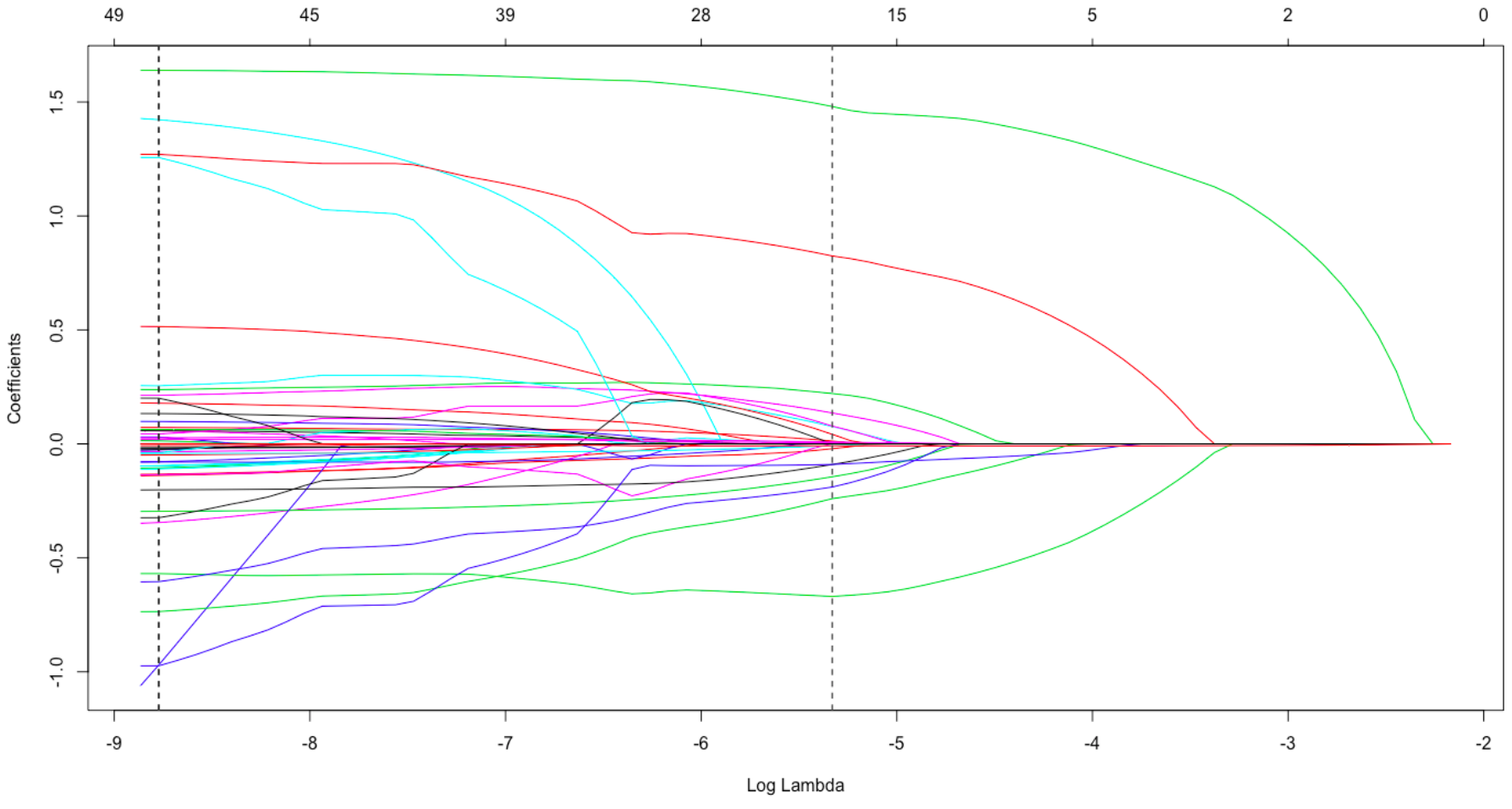
- We can see that bootstrap and estimated values coincide for each coefficients, except for some irrelevant bias
- We consider this result as a proof that even if initially we split up the data-set we still have enough observations to allow the asymptotic theory of the maximum likelihood estimators of the coefficients to be valid

Gam and Lasso models

- We have also tried to apply other statistical models
- The first one has been a Gam model, but with no significant results (see the script).
- On the contrary, the lasso is quite interesting In this case we see that all the coefficients that have not been pushed to zero are the ones considered in the logistic classifier, but also the others considered important in the EDA, including jobs and campaign

Gam and Lasso models

Pushing coefficients towards zero



Conclusions

References:

1. Data-set: [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014
2. and definitely too many R packages to quote...

Conclusions

Thank You for Attention !