Finding Opportunity in Crisis

Jens Meydam

2019-06-13

Preliminary Remark

This project is an exercise. All results are derived from the data set and the accompanying documentation.

Executive Summary

This is an analysis of the bank marketing data set made available via the UCI Machine Learning Repository maintained by the University of California, Irvine.

The data were collected during direct marketing campaigns run by a Portuguese bank. The classification goal is to predict whether a client will accept the offer - made via phone call - to invest money in a term deposit. Note that the data were collected from May 2008 to November 2010, covering the period from just before until two years into the financial crisis, and that term deposits are considered a safe investment.

The date of the observations is not given, but the 41,188 observations are ordered by date. There is a sharp drop in the 3 month Euribor interest rate aoround observation 25,000. If we measure the efficiency of campaigns in terms of the relative frequency of positive responses, campaigns became dramatically more efficient - and in that sense more successful - starting around observation 27,500.

While the timing suggests that the state of the economy played a role, the improvement of the success rate might also be due to a learning effect: the campaigns towards the end may have learned lessons from previous campaigns and this might be an important reason why later campaigns were more efficient/successful.

Could the improvement of campaign performance have come earlier, before the onset of the financial crisis?

At the end of this study, a commonly highly effective machine learning algorithm is trained and tested both on the data before the onset of the financial crisis and on the data after, excluding as far as possible information on the state of the economy. The model trained on the earlier data has no predictive power, which suggests that the data contain no useful signal. In contrast, the model trained on the later data has considerable predictive power and could have been used for optimizing campaigns. In fact, later campaigns appear to have been informed by such a model, systematically targeting subgroups and calibrating the number of calls.

Our claim is not so much that the earlier data are in fact completely useless, but that the later data are strikingly more useful. Dramatic changes in the economy may have led to the amplification of signals in the data and/or to the emergence of entirely new signals. We suggest this as a potentially fruitful avenue for further research.

To conclude, it appears that the marked improvement of campaign performance was the result of a new, optimized approach that exploited changing attitudes in a country in crisis, and that a similar improvement would not have been possible before. In particular, it is conceivable that certain identifiable subgroups among the clients may have become more susceptible to sales pitches emphasizing security and wealth protection, but this is mere speculation.

Data Set

Bank marketing data set made available via the UCI Machine Learning Repository maintained by the University of California, Irvine.

Specific Data Set Used

> bank-additional-full.csv with all examples (41,188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014].

Download Link

> http://archive.ics.uci.edu/ml/machine-learning-databases/00222/bank-additional.zip

Full Citation for 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

As stated on the website, the data were collected during

> [. . .] direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe [to] a term deposit (variable y).

Background Information on Term Deposits

> https://www.investopedia.com/terms/t/termdeposit.asp
>
> A term deposit is a fixed-term investment that includes the deposit of money into an account at a financial institution. Term deposit investments usually carry short-term maturities ranging from one month to a few years and will have varying levels of required minimum deposits.
>
> The investor must understand when buying a term deposit that they can withdraw their funds only after the term ends. In some cases, the account holder may allow the investor early termination - or withdrawal - if they give several days notification. Also, there will be a penalty assessed for early termination.
>
> [. . .] Term deposits are an extremely safe investment and are therefore very appealing to conservative, low-risk investors.

Attributes from Bank Client Data

1. age (numeric)
2. job: type of job (categorical: “admin.”, “blue-collar”, “entrepreneur”, “housemaid”, “management”, “retired”, “self-employed”, “services”, “student”, “technician”, “unemployed”, “unknown”)
3. marital: marital status (categorical: “divorced”, “married”, “single”, “unknown”; note: “divorced” means divorced or widowed)
4. education (categorical: “basic.4y”, “basic.6y”, “basic.9y”, “high.school”, “illiterate”, “professional.course”, “university.degree”, “unknown”)
5. default: has credit in default? (categorical: “no”, “yes”, “unknown”)
6. housing: has housing loan? (categorical: “no”, “yes”, “unknown”)
7. loan: has personal loan? (categorical: “no”, “yes”, “unknown”)

Attributes Related to the Last Contact of the Current Campaign

8. contact: contact communication type (categorical: “cellular”, “telephone”)
9. month: last contact month of year (categorical: “jan”, “feb”, “mar”, . . . , “nov”, “dec”)
10. day_of_week: last contact day of the week (categorical: “mon”, “tue”, “wed”, “thu”, “fri”)
11. duration: last contact duration, in seconds (numeric). **Important note**: this attribute highly affects the output target (e.g., if duration=0 then y=“no”). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Other Attributes

12. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14. previous: number of contacts performed before this campaign and for this client (numeric)
15. poutcome: outcome of the previous marketing campaign (categorical: “failure”, “nonexistent”, “success”)

Social and Economic Context Attributes

16. emp.var.rate: employment variation rate - quarterly indicator (numeric)
17. cons.price.idx: consumer price index - monthly indicator (numeric)
18. cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19. euribor3m: euribor 3 month rate - daily indicator (numeric)
20. nr.employed: number of people employed - quarterly indicator (numeric)

Output Variable (Target)

21. y: has the client subscribed a term deposit? (binary: “yes”, “no”)

Added for GBM: y_bernoulli (0 = “no”, 1 = “yes”)

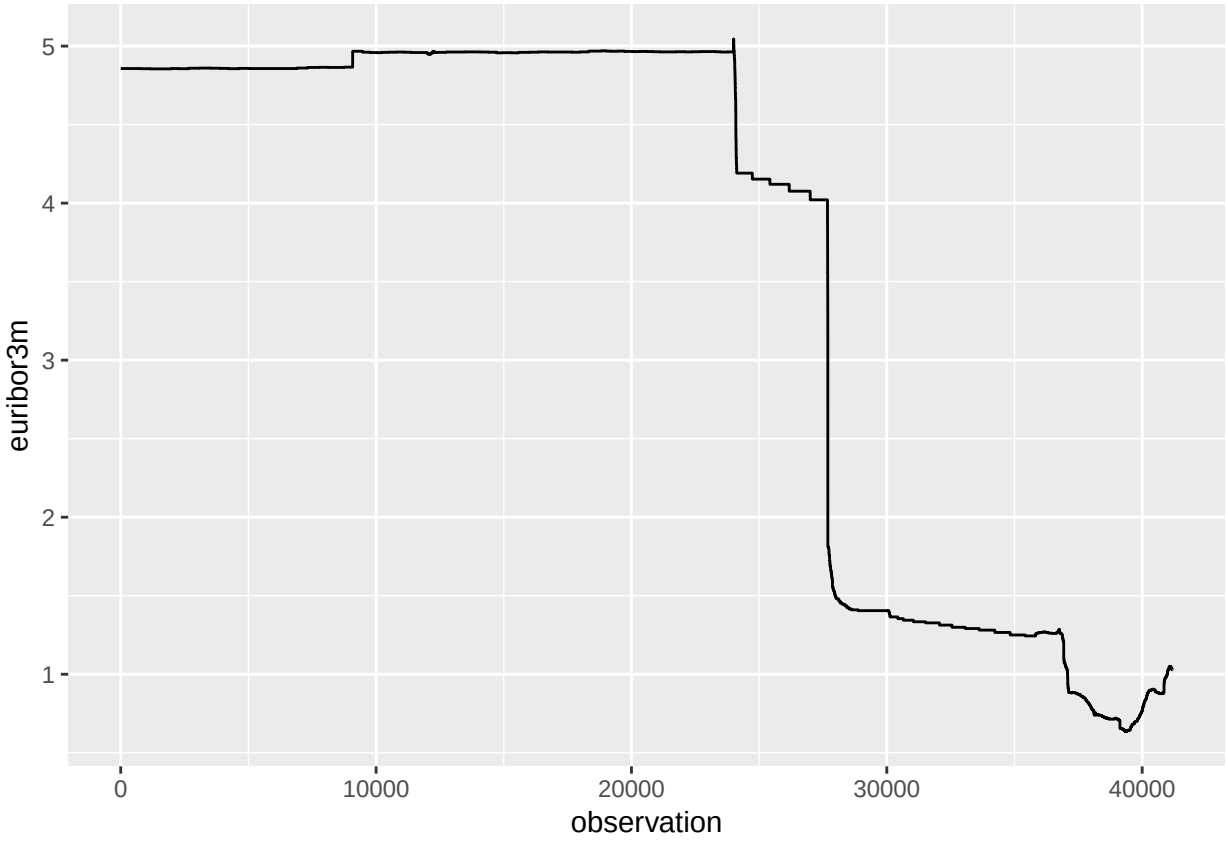
Exploratory Data Analysis

Social and Economic Context Attributes

Most of these variables are correlated.

The date of the 41,188 observations is not given, but are ordered by date, from May 2008 to November 2010. Therefore, plotting the data in sequence will give an approximation to a proper time series.

There is a sharp drop in the 3 month Euribor interest rate aoround observation 25,000:



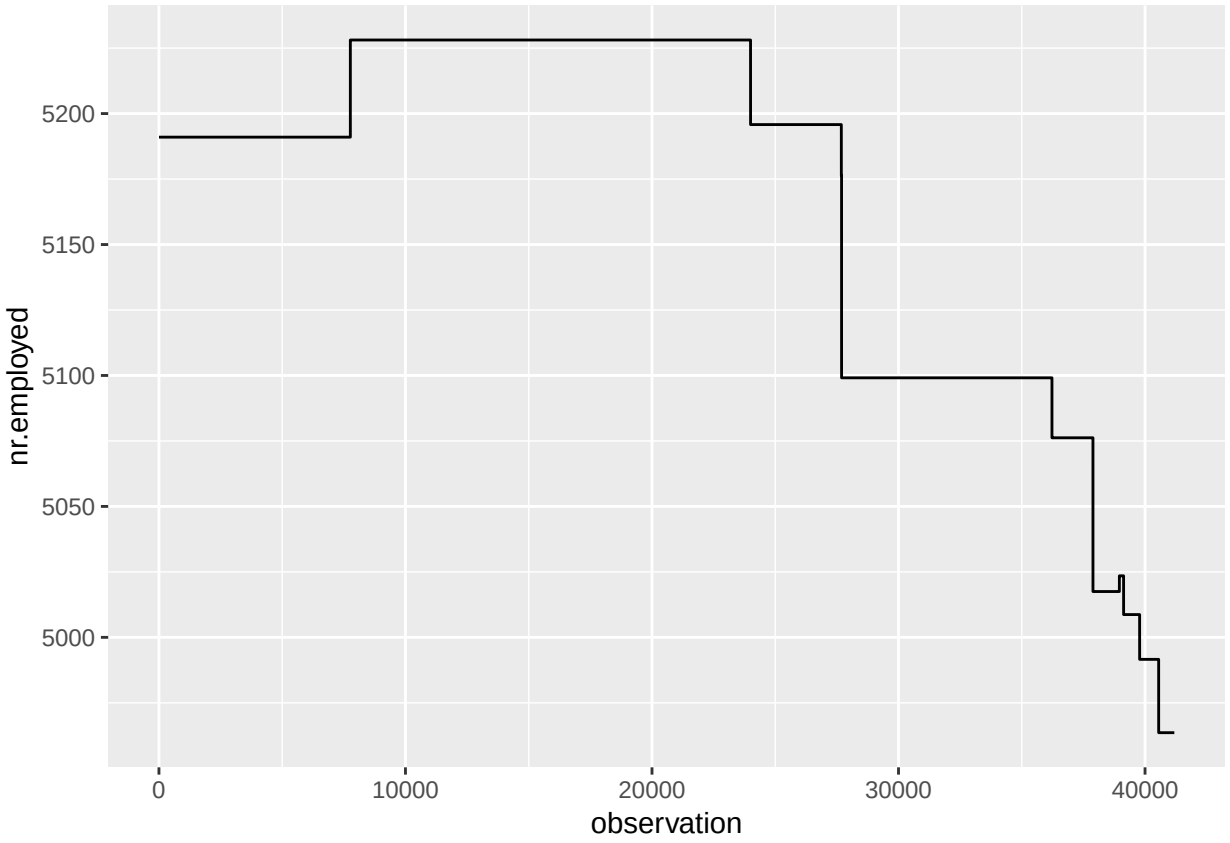
Note:

- The observations cover the period from just before to two years after the onset of the global financial crisis. Lehman Brothers collapsed on September 15, 2008. (https://en.wikipedia.org/wiki/Financial_crisis_of_2007%E2%80%932008)
- The observations are not evenly distributed in time.
- Higher variability after observation 36,000 is partly due to fewer observations per time unit (fewer/lower intensity campaigns).

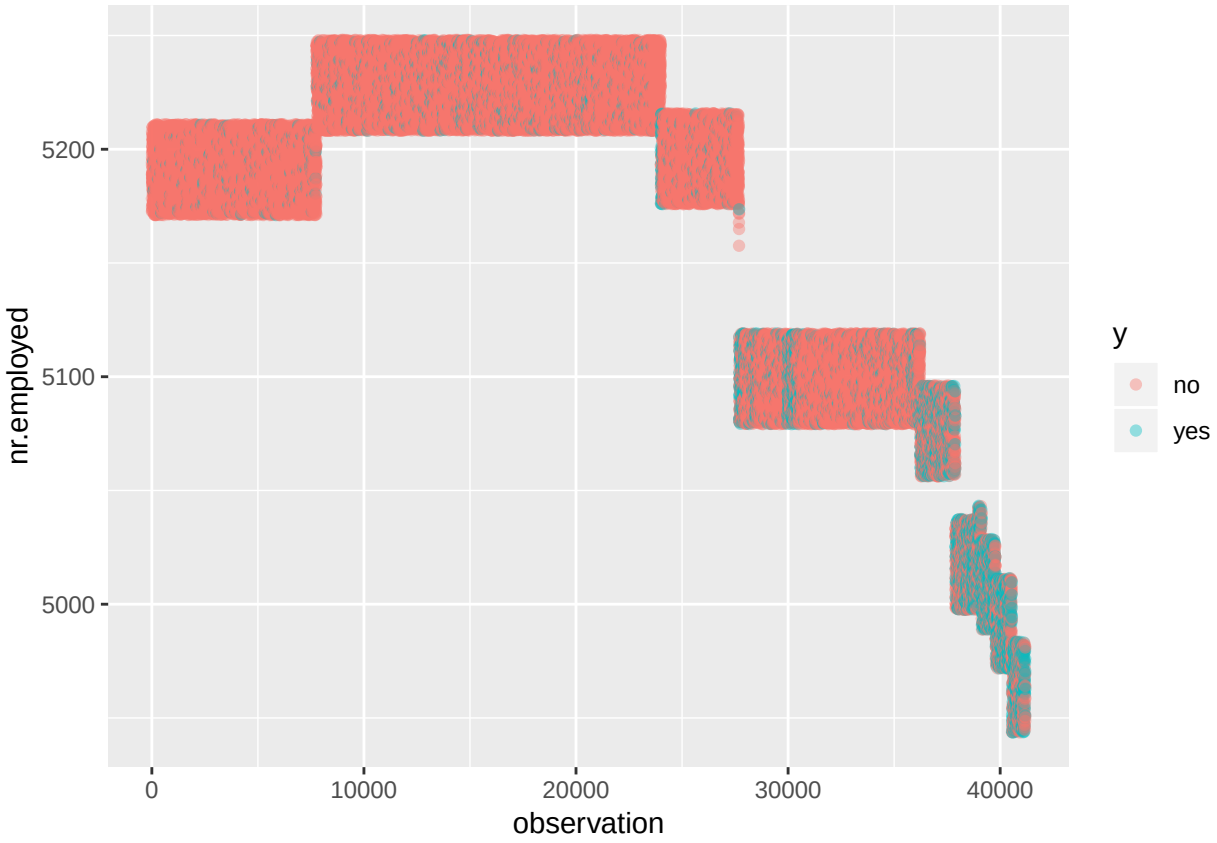
Next we will focus on the the number of people employed.

Number of People Employed

Note the solid downward trend after the onset of the crisis:



Now we will also show the responses ("yes"/"no"). We will show responses as jittered points to make gradual changes in frequency easier to see:



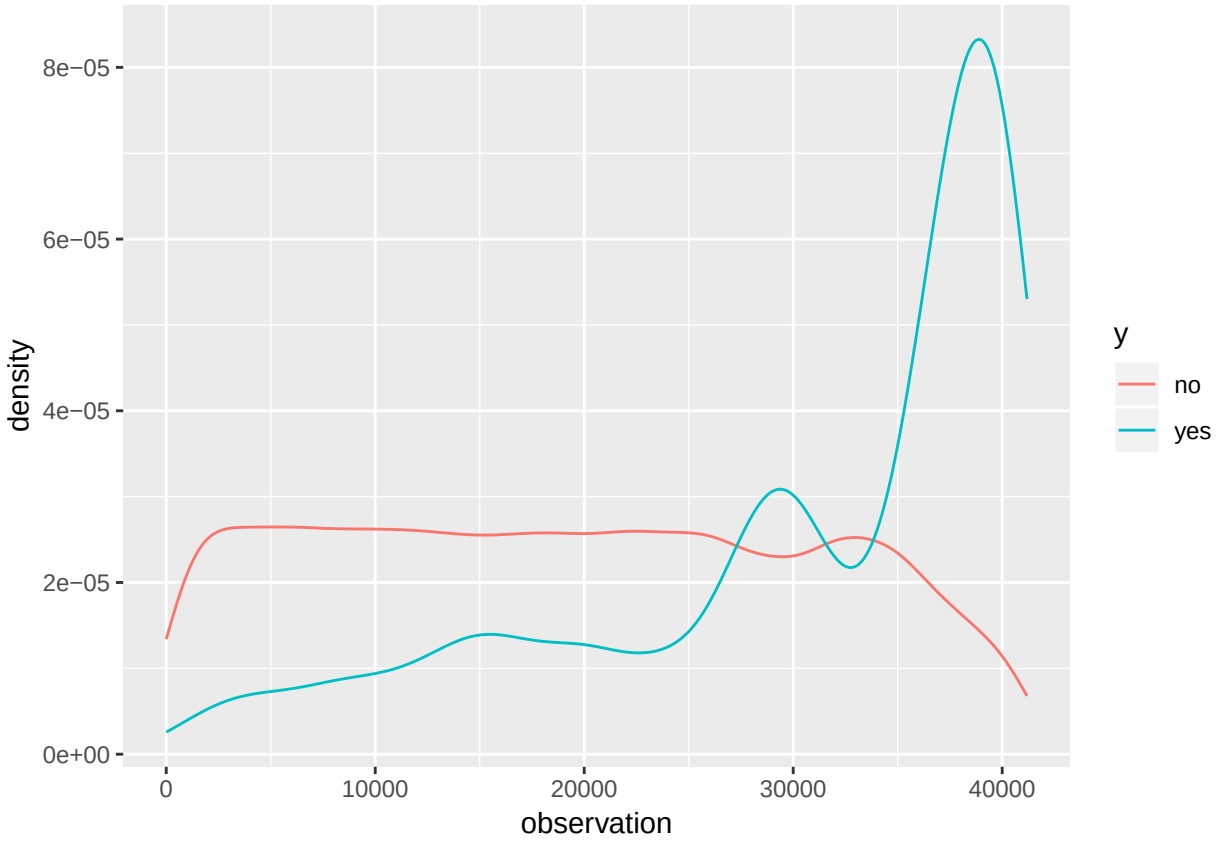
Note how "yes" responses become more dominant towards the end. Keep in mind that observations are not evenly distributed in time. The number of people employed is a quarterly indicator; apparent accelerated change towards the end is an effect of having fewer observations per time unit.

While there are "yes" responses throughout, the "no" responses appear to drown out the "yes" responses for most of the plot.

Responses over Time

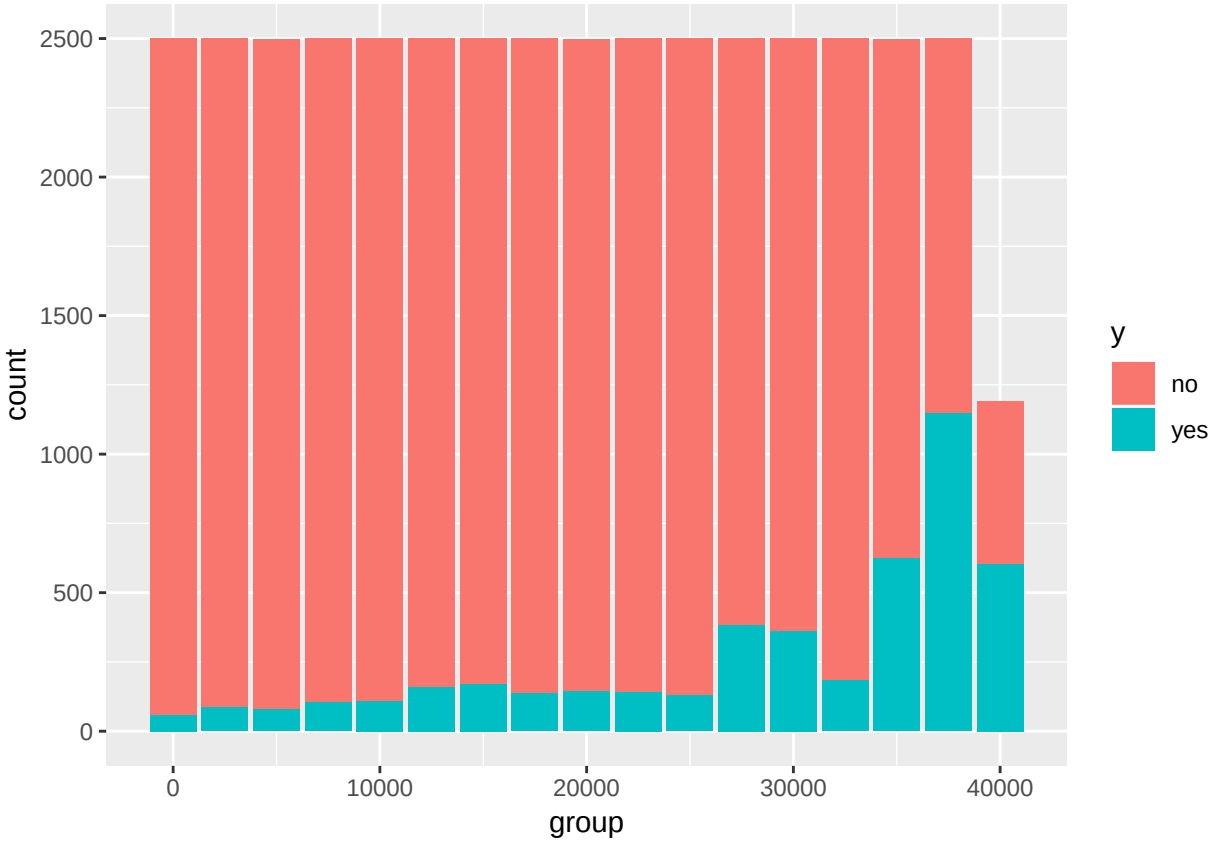
Next, to see trends for both responses more clearly, we will plot density curves of "yes" and "no" responses. These curves are smoothed and normalized so that each curve integrates to 1.

Note the slow, then halting increase in the frequency of "yes" responses before the financial crisis, followed by steep bump just before observation 30,000, and then a solid rise (the drop at the end is an artifact of the plot).



Finally, to show the changing share of "yes" and "no" responses, we will bin observations in groups of 2,500.

Note the much lower prevalence of "yes" responses in the first two thirds of the plot. Again, the lower bar at the end and consequently the apparent drop both of "yes" and "no" responses is an artifact of the plot. In fact, the share of "yes" responses rises to over 50% in the last bar.



Thoughts

In the first visual displays of the data, until about observation 36,000, the "no" responses drowned out the "yes" responses - even using high levels of jitter and transparency left some uncertainty concerning the changing absolute frequency of "yes" responses. After observation 36,000 the relative frequency of "no" responses drop dramatically, making the "yes" responses clearly visible.

Further analysis confirmed that "yes" responses, while present from the beginning, became much more frequent starting around observation 27'500, and ended up being more frequent than "no" responses.

If we measure the efficiency of the campaigns in terms of the ratio of "yes" to "no" responses, campaigns became dramatically more efficient - and in that sense more successful - starting around observation 27'500.

The most efficient campaigns towards the end can also be identified via nr.employed, since the number of people employed is decreasing for the second half of the observations and is significantly lower towards the end. The employment variation rate (emp.var.rate) can also be used to detect observations from 27,500 onwards.

There might be a causal connection between the efficiency/success of campaigns and a combination of economic indicators (in particular, the number of people employed and the employment variation rate), but the improvement of the success rate might also be due to a learning effect: the campaigns towards the end may have learned lessons from previous campaigns and this might be an important reason why later campaigns were more efficient/successful.

It is worth noting that after an initial dramatic improvement around observation 27,500 performance first went back to the previous level, then returned to the improved level, and then improved even further. It is conceivable that this discontinuity and further accelerated improvement is a result of different campaigns using different approaches, with systematic optimization of the approach after a brief delay after a first breakthrough around observation 27,500. This new, optimized approach might have exploited changing attitudes in a country in crisis. In particular, it is conceivable that certain identifiable subgroups among the customers may have become more susceptible to sales pitches emphasizing security and wealth protection.

Other Attributes

As we will see later, three attributes are particularly relevant:

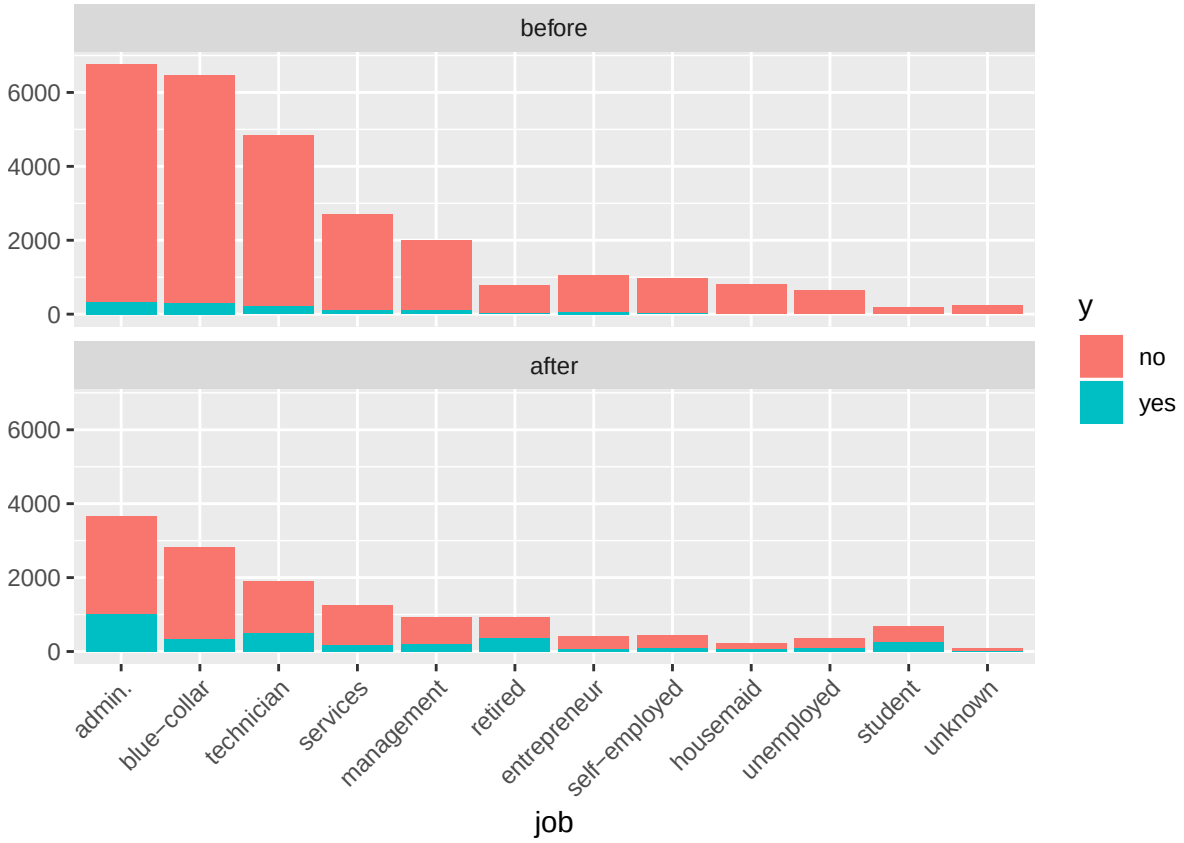
1. Job
2. Age
3. Education

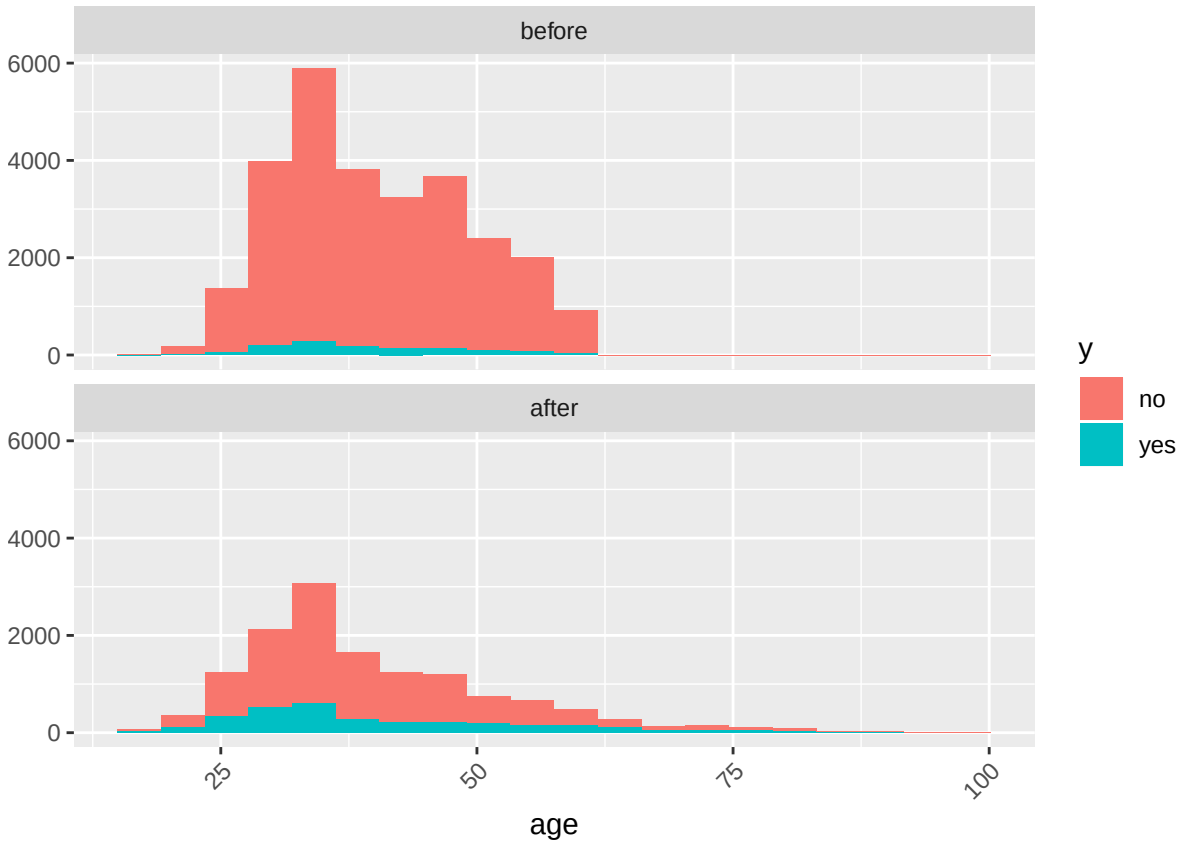
We will show graphs for these three attributes, contrasting the frequencies of the various values "before" and "after", while also showing the response y.

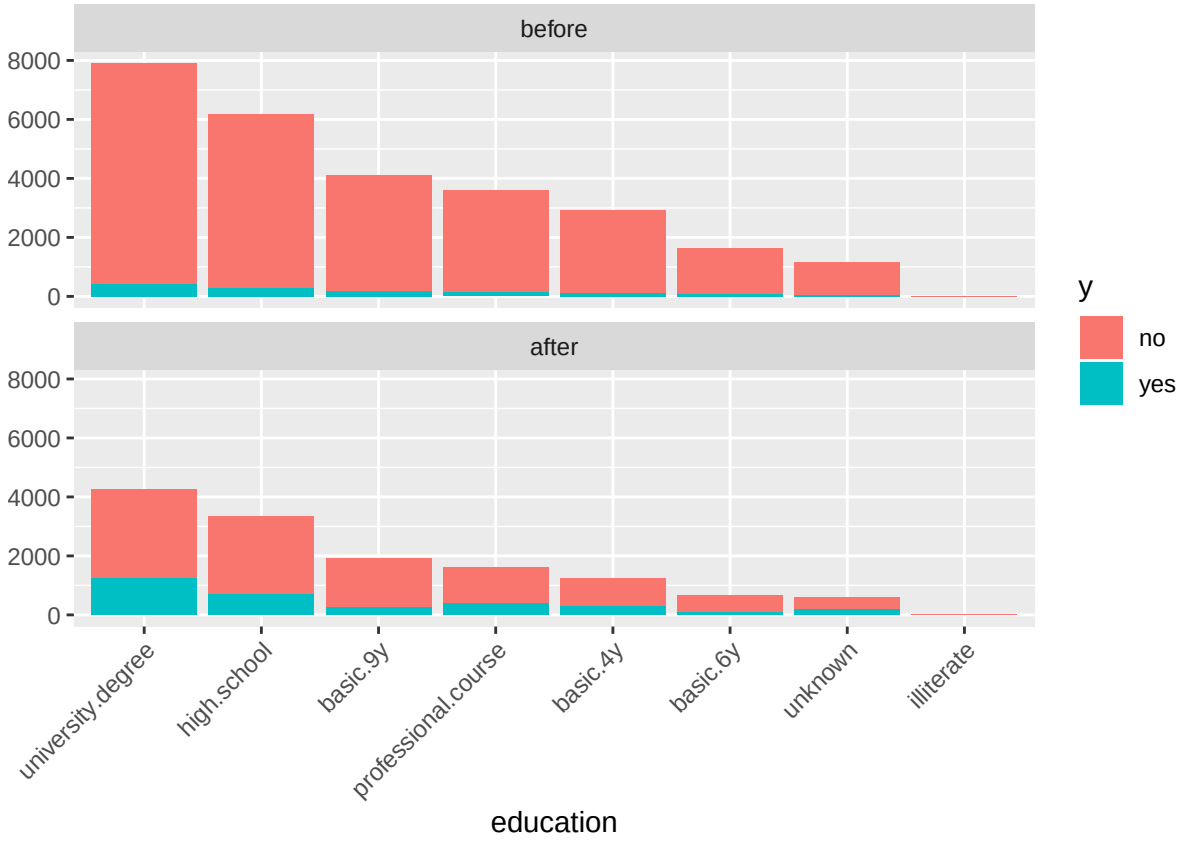
Note the changes after the onset of the financial crisis. In general, the share of positive responses has increased. Categories and ranges with relatively many positive responses show a marked increase, some more so than others. Some categories and ranges that were previously hardly represented or not represented at all become noticeable, with very high shares of positive responses, in particular the categories "student" and "retired" and the associated age ranges.

Thoughts

A general effect of the financial crisis is highly likely. Furthermore, the emergence of new groups with high success rates points to changes in the targeting of the campaigns.




before
after
6000
4000
2000
0
25
50
75
100
age
y
no
yes


before
after
8000
6000
4000
2000
0
university.degree
high.school
basic.9y
professional.course
basic.4y
basic.6y
unknown
illiterate
education
y
no
yes

Model

Given the obvious changes both in the economy and in the targeting of the campaigns after the onset of the financial crisis, a model based on the whole data set is bound to give misleading results. The data set is therefore split into two data sets, "before" and "after" the onset of the financial crisis, using observation 27,500 as the cut-off. Each of these new data sets is then split into a training and a test set.

Since there are many categorical variables and at least some effects of continuous variables (age!) may be nonlinear, a tree-based model is a natural choice. We will fit a simple boosted model using the gbm package.

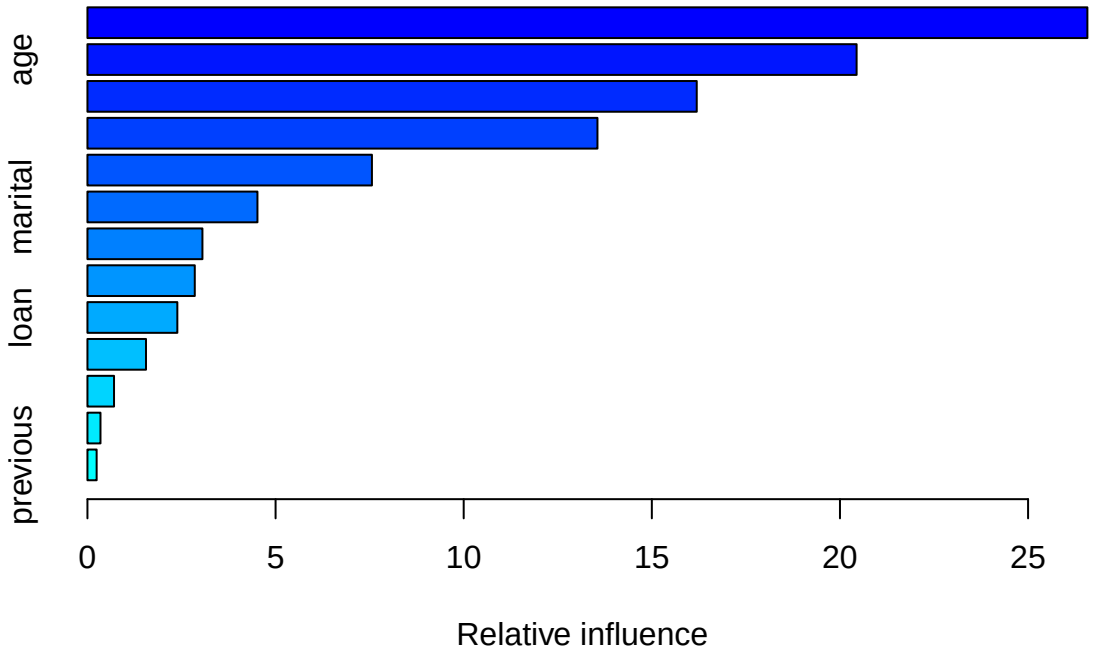
The emphasis here is on understanding what is roughly possible with the data before and after the onset of the financial crisis, while leaving out all economic variables and the month to reduce timing effects.

Tuning the models by optimizing parameters would be a natural next step.

Before Onset of Financial Crisis

```
boost_before_3 <- gbm(y_bernoulli ~ .,
                      data = subset(train_before,
                                    select = -c(duration, month, euribor3m, emp.var.rate,
                                                cons.price.idx, cons.conf.idx, nr.employed,
                                                y)),
                      distribution = "bernoulli",
                      n.trees = 5000,
                      interaction.depth = 4)

summary(boost_before_3)
```

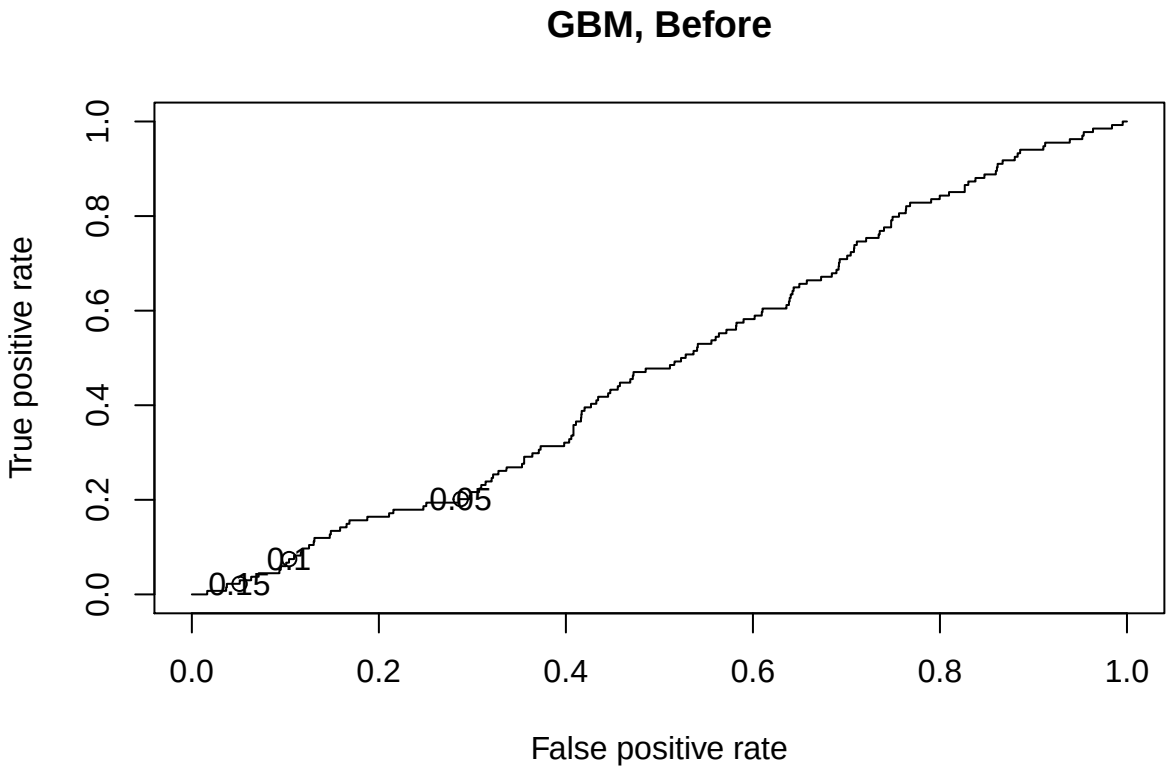



```
##                     var    rel.inf
## job                 job 26.5760640
## age                 age 20.4428023
## education     education 16.1938455
## day_of_week day_of_week 13.5550010
## campaign       campaign  7.5608956
## marital         marital  4.5181140
## contact         contact  3.0561994
## housing         housing  2.8526186
## loan               loan  2.3893808
## default         default  1.5565461
## poutcome       poutcome  0.7068396
## pdays             pdays  0.3469229
## previous       previous  0.2447702
```

Test

```
yhat_boost_before <- predict(boost_before_3,
                             type = "response",
                             newdata = test_before,
                             n.trees = 5000)

plot(performance(prediction(predictions = yhat_boost_before,
                            labels = ifelse(test_before$y_bernoulli == 1, "yes", "no")),
                 "tpr",
                 "fpr"),
     print.cutoffs.at = c(0.05, 0.10, 0.15),
     main = "GBM, Before")
```

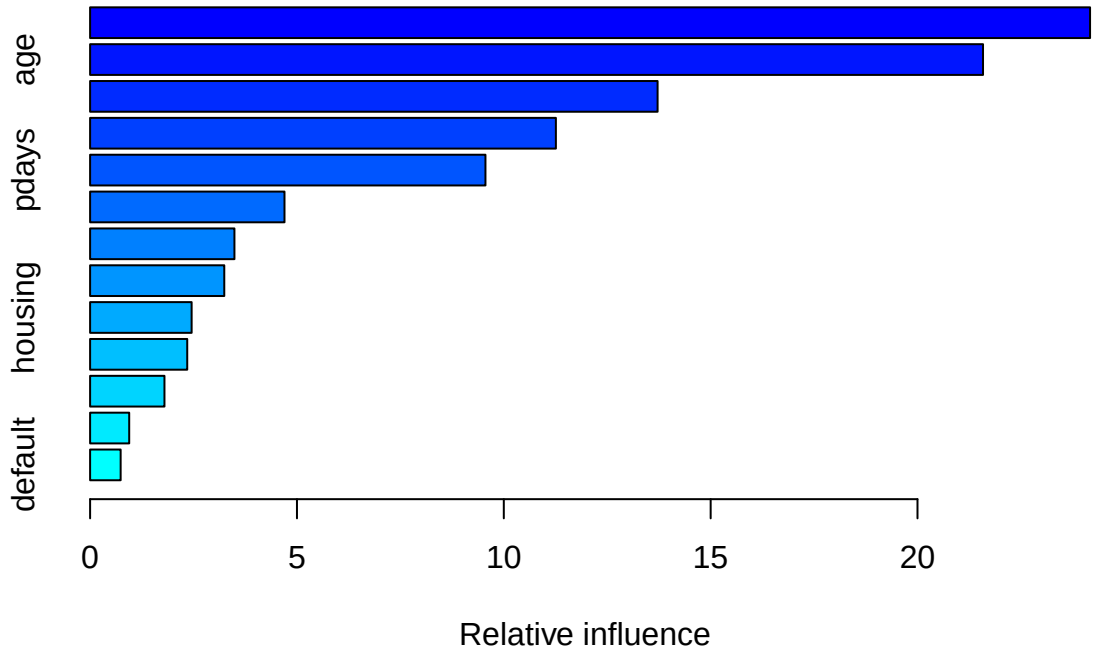


This model is basically useless. Inspecting the ROC curve shows that the model performs at the level of random guessing.

After Onset of Financial Crisis

```
boost_after_3 <- gbm(y_bernoulli ~ .,
                     data = subset(train_after,
                                   select = -c(duration, month, euribor3m, emp.var.rate,
                                               cons.price.idx, cons.conf.idx, nr.employed,
                                               y)),
                     distribution = "bernoulli",
                     n.trees = 5000,
                     interaction.depth = 4)

summary(boost_after_3)
```

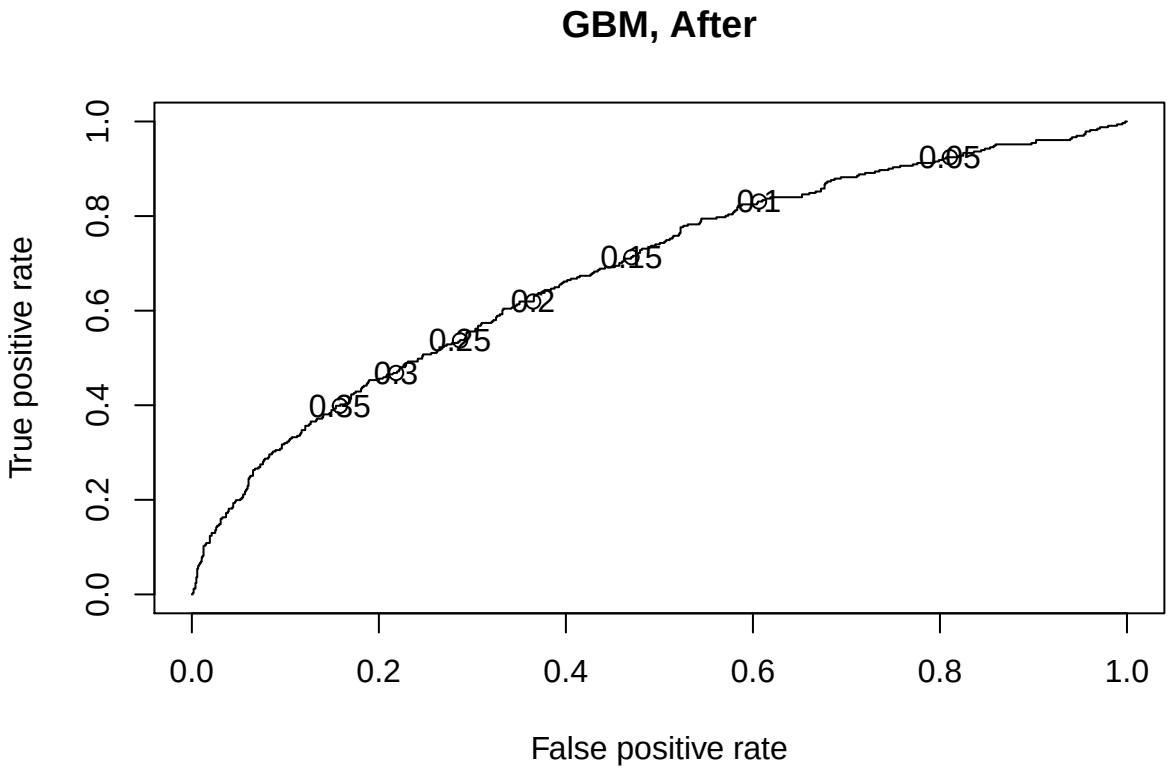



```
##                     var    rel.inf
## job                 job 24.1710027
## age                 age 21.5834137
## education     education 13.7176167
## day_of_week day_of_week 11.2587708
## pdays             pdays  9.5564042
## campaign       campaign  4.6986626
## poutcome       poutcome  3.4887042
## marital         marital  3.2433103
## housing         housing  2.4541944
## previous       previous  2.3475798
## loan               loan  1.7973868
## contact         contact  0.9448086
## default         default  0.7381453
```

Test

```
yhat_boost_after <- predict(boost_after_3,
                            type = "response",
                            newdata = test_after,
                            n.trees = 5000)

plot(performance(prediction(predictions = yhat_boost_after,
                            labels = ifelse(test_after$y_bernoulli == 1, "yes", "no")),
                 "tpr",
                 "fpr"),
     print.cutoffs.at = c(0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35),
     main = "GBM, After")
```



This model clearly has predictive power. A cut-off can be chosen to calibrate campaigns, for example to improve the proportion of positive responses.

Results

A commonly highly effective machine learning algorithm was trained and tested both on the data before the onset of the financial crisis and on the data after, excluding as far as possible information on the state of the economy.

The model trained on the earlier data has no predictive power, which suggests that the data contain no useful signal.

In contrast, the model trained on the later data has considerable predictive power and could have been used for optimizing campaigns. In fact, as pointed out earlier, later campaigns appear to have been systematically targeting subgroups. It can be assumed that later campaigns were based on models similar to the one presented here.

Our claim is not so much that the earlier data are in fact completely useless, but that the later data are strikingly more useful. Dramatic changes in the economy may have led to the amplification of signals in the data and/or to the emergence of entirely new signals. We suggest this as a potentially fruitful avenue for further research.

Conclusion

It appears that the marked improvement of campaign performance was the result of a new, optimized approach that exploited changing attitudes in a country in crisis, and that a similar improvement would not have been possible before.

In particular, it is conceivable that certain identifiable subgroups among the clients may have become more susceptible to sales pitches emphasizing security and wealth protection, but this is mere speculation.