Predicting_client_subscription

Felix Seo

2022-07-04

```
## Loading required package: ggplot2

## Loading required package: lattice

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()

## corrplot 0.92 loaded

## Warning: package 'glmnet' was built under R version 4.2.2

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 4.1-4

## Warning: package 'gglasso' was built under R version 4.2.2

## Warning: package 'rpart.plot' was built under R version 4.2.2
```

ideas for better prediction

add a column with information if it is the first or second phone call as the same customer often was contacted twice to check if term deposit.

Data preparation

Starting with looking at the data, the following attribute information follows [Moro et al., 2014].

input variables:

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','
- 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')

related with 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 employees - quarterly indicator (numeric)

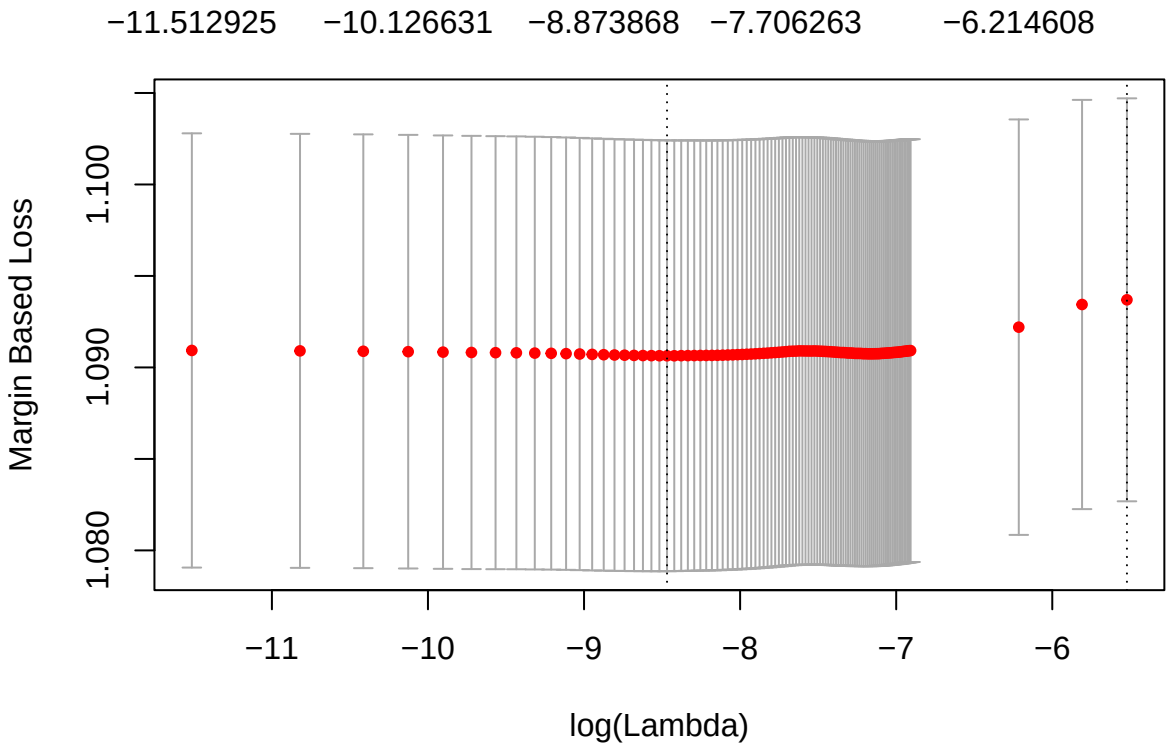
Output variable (desired target):

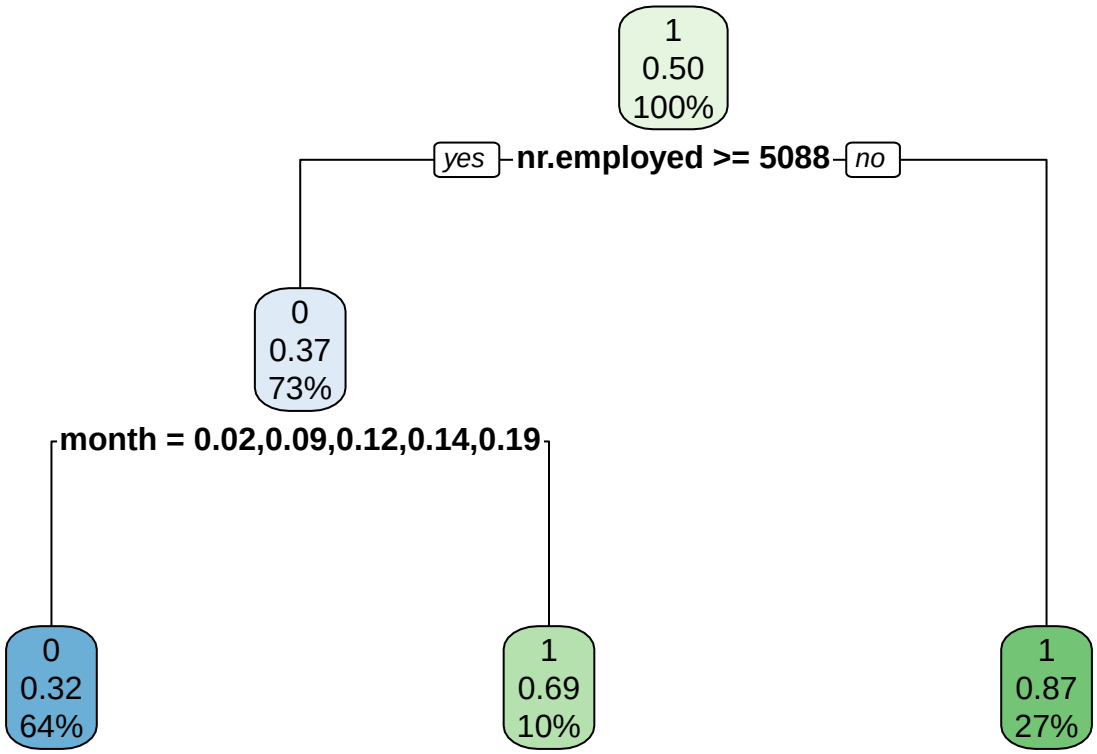
- 21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

Seen in the attribute information the 11th input variable is not desired for use in an predictive model, why it is removed. The default attribute is also removed since it only has 3 realizations of people with credit in default. This huge difference in group size will not contribute to the predicative power of the model. Attributes 2 to 7 all have the risk of containing missing values labeled "unknown". The data containing any of the missing values in attributes 2 to 7 are removed, which ended up being roughly 3000 instances (still the dataset contains approximately 38000 instances). For the categorical variables the appropriate encoding is conducted. The binary variables are transformed to 0 and 1 outcomes. The ordinal or nominal category classification for the variables are debatable. The education variable is easily classified as an ordinal type variable why we will use simple ordinal encoding. Variables like marital status, last contact day of the week and month are classified as nominal variables and one-hot encoding is conducted. The latter is arguably nominal in this context as the month seems not to have an natural ordering for this classification task.

```
## Rows: 41188 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## chr (11): job, marital, education, default, housing, loan, contact, month, d...
## dbl (10): age, duration, campaign, pdays, previous, emp.var.rate, cons.price...
##
## i Use 'spec()' to retrieve the full column specification for this data.
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.
```

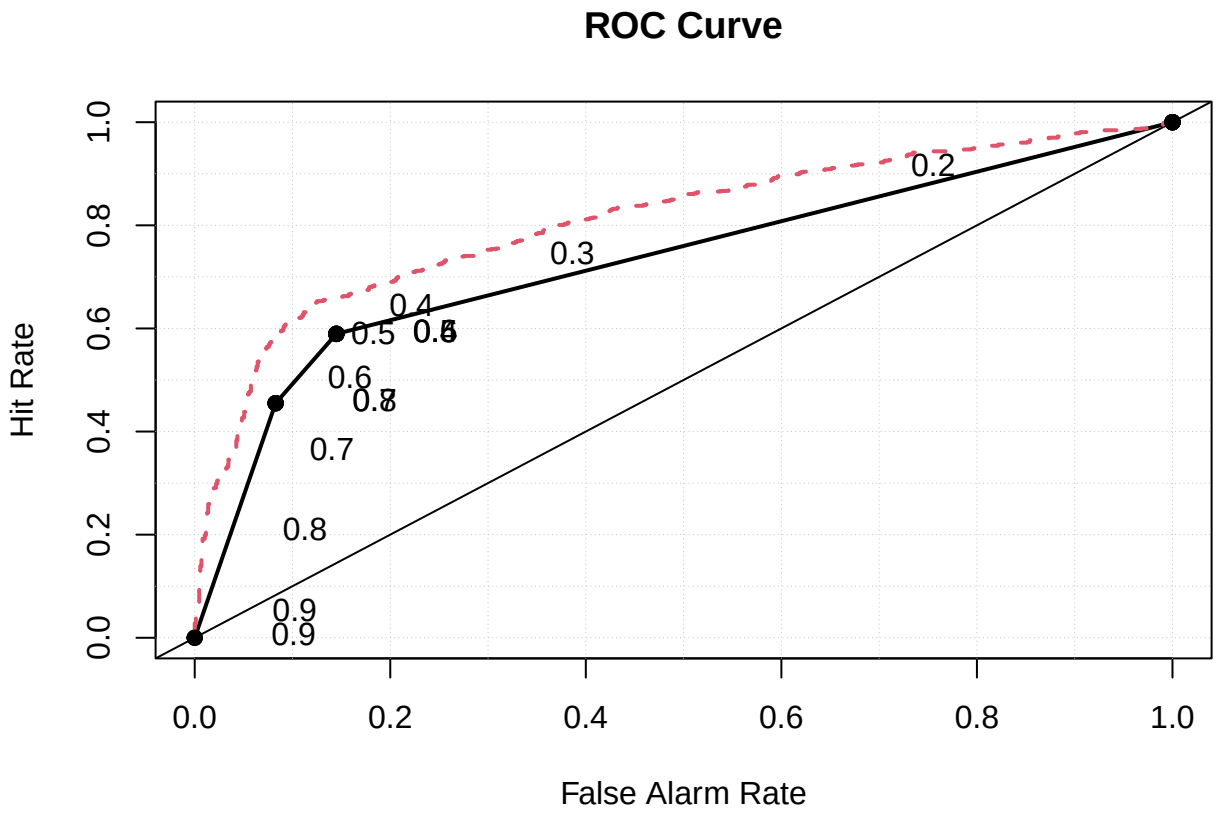
```
## # A tibble: 38,245 x 46
##      age education housing  loan contact campaign pdays previous emp.var.rate
##    <dbl>     <dbl>   <dbl> <dbl>   <dbl>    <dbl> <dbl>    <dbl>        <dbl>
##  1    48         6       1     1       1        1   999        0          1.4
##  2    27         6       1     0       1        3    10        1         -3
##  3    19         4       1     0       1        2   999        3         -3
##  4    54         1       0     0       1        2   999        0          1.4
##  5    38         3       1     0       1        3   999        0          1.4
##  6    29         3       0     0       0        1   999        0          1.1
##  7    26         4       0     0       1        3   999        0          1.4
##  8    29         6       1     1       1        1   999        0          1.4
##  9    40         6       0     0       1        4   999        0          1.4
## 10    23         4       0     0       1        1   999        0         -1.8
## # ... with 38,235 more rows, and 37 more variables: cons.price.idx <dbl>,
## #   cons.conf.idx <dbl>, euribor3m <dbl>, nr.employed <dbl>, y <dbl>,
## #   housemaid <dbl>, services <dbl>, admin. <dbl>, technician <dbl>,
## #   blue.collar <dbl>, retired <dbl>, management <dbl>, unemployed <dbl>,
## #   self.employed <dbl>, entrepreneur <dbl>, student <dbl>, married <dbl>,
## #   single <dbl>, divorced <dbl>, may <dbl>, jun <dbl>, jul <dbl>, aug <dbl>,
## #   oct <dbl>, nov <dbl>, dec <dbl>, mar <dbl>, apr <dbl>, sep <dbl>, ...
```


−11.512925
−10.126631
−8.873868
−7.706263
−6.214608
Margin Based Loss
1.080
1.090
1.100
−11
−10
−9
−8
−7
−6
log(Lambda)




```
## Registered S3 method overwritten by 'verification':
##   method    from
##   lines.roc pROC
```

```
## Warning in roc.plot.default(test$y, hej): Large amount of unique predictions
## used as thresholds. Consider specifying thresholds.
```




```
{r; eval = FALSE} bank_data_mod %>%    select(     y,      age,      education,      euribor3m,
cons.price.idx,     nr.employed,      cons.conf.idx,     campaign,     pdays,      previous
) %>%    cor(method = "spearman") %>%    corrplot(type = "upper")
```