# PREDICTING BANK TELEMARKETING SUCCESS

Final Exam due Dec 8, 2020 07:59 +08 Past due

# PREDICTING BANK TELEMARKETING SUCCESS

The success of marketing campaigns can be highly specific to the product, the target audience, and the campaign methods. In this problem, we examine data from direct marketing campaigns of a Portuguese banking institution between May 2008 and November 2010. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be or not subscribed.

In this analysis, the goal would be predicting the dependent variable **y**, which takes value 1 if the the client subscribed to a term deposit, and 0 otherwise. The data we will be using bank.csv is a subset of the original data, containing 5000 examples and 20 input variables. The variable information is as follows:

- **age**
- **job** - type of job
- **marital** - marital status
- **education** - Shows the level of education of each customer
- **default** - Whether a customer has credit in default
- **housing** - Does the customer have a housing loan?
- **loan** - Does the customer have a personal loan?
- **contact** - The contact communication type
- **month** - Last contact month of year
- **day_of_week** - Last contact day of Week
- **duration** - Last contact duration in seconds (*Note: this variable is not known before making the call*)
- **campaign** - Number of contact performed for the client during the campaign
- **pdays** - number of days that passed by after the client was last contacted from a previous campaign (value of 999 means the client was not previously contacted)
- **previous** - number of contacts performed before this campaign and for this client
- **poutcome** - outcome of the previous marketing campaign
- **emp.var.rate** - employment variation rate - quarterly indicator
- **cons.price.idx** - consumer price index - monthly indicator
- **cons.conf.idx** - consumer confidence index - monthly indicator
- **euribor3m** - euribor 3 month rate - daily indicator
- **nr.employed** - number of employees - quarterly indicator

---

## Problem 2 - Call Durations by Job

0.0/2.0 points (graded)

Build a boxplot that shows the call duration distributions over different jobs. Which three jobs have the longest average call durations? (if it's hard to see from the boxplot, use tapply function.)

- [ ] admin.
- [ ] blue-collar
- [ ] entrepreneur
- [ ] housemaid ✔
- [ ] management
- [ ] retired

Calculator

- [ ] self-employed

- [ ] unemployed

Explanation
By examining tapply(bank$duration, bank$job, mean), we can see the three jobs with highest mean call durations.

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

## Problem 3 - Multicolinearity

0.0/2.0 points (graded)
As good practice, it is always helpful to first check for multicolinearity before running models, especially since this dataset contains macroeconomic indicators. Examine the correlation between the following variables: emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, and nr.employed. Which of the following statements are correct (limited to just these selected variables)?

- [ ] cons.conf.idx does NOT seem to have severe multicolinearity with the other variables.
- [ ] emp.var.rate and nr.employed have the highest correlation between two different variables.
- [ ] cons.price.idx and cons.conf.idx have the lowest correlation between two different variables.

Explanation
Use cor function to get the correlation matrix and inspect.

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

## Problem 4 - Splitting into a Training and Testing Set

0.0/5.0 points (graded)
Obtain a random training/testing set split with:

```
set.seed(201)

library(caTools)

spl = sample.split(bank$y, 0.7)
```

Calculator

spl = sample.split(bank$y, 0.7)

Split months into a training data frame called "training" using the observations for which spl is TRUE and a testing data frame called "testing" using the observations for which spl is FALSE.

Explanation
Use the subset function to put the TRUE observations in the training set, and the FALSE observations in the test set.

Why do we use the sample.split() function to split into a training and testing set?

- ○ It is the most convenient way to randomly split the data
- ○ It balances the independent variables between the training and testing sets
- ○ It balances the dependent variable between the training and testing sets ✔

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

## Problem 5 - Training a Logistic Regression Model

0.0/2.0 points (graded)

Train a logistic regression model using independent variables age, job, marital, education, default, housing, loan, contact, month, day_of_week, campaign, pdays, previous, poutcome, emp.var.rate, cons.price.idx, and cons.conf.idx, using the training set to obtain the model. Notice that we have removed duration (since it's not available before the call, so shouldn't be used in a strictly predictive model), euribor3m and nr.employed (due to multicolinearity issue).

Which of the following characteristics are statistically significantly POSITIVELY (at 0.05 level) associated with an increased chance of subscribing to the product?

- [ ] age ✔
- [ ] default is unknown
- [ ] contact via telephone
- [ ] month is August ✔
- [ ] month is March ✔
- [ ] day_of_week is Monday
- [ ] poutcome is nonexistent ✔
- [ ] emp.var.rate

Calculator

- [ ] cons.price.idx ✔
- [ ] cons.conf.idx

Explanation
The model can be trained with the glm function (remember the argument family="binomial") and summarized with the summary function.

Submit You have used 0 of 3 attempts

Answers are displayed within the problem

## Problem 6 - Interpreting Model Coefficients

1 point possible (graded)
What is the meaning of the coefficient labeled "monthmar" in the logistic regression summary output?

- ( ) When the month is March, the odds of subscribing to the product are 261.8% higher than an otherwise identical contact. ✔
- ( ) When the month is March, the odds of subscribing to the product are 261.8% higher than the avarage contact.
- ( ) When the month is March, the odds of subscribing to the product are 28.6% higher than an otherwise identical contact.
- ( ) When the month is March, the odds of subscribing to the product are 28.6% higher than the avarage contact.

Explanation
The coefficients of the model are the log odds associated with that variable; so we see that the odds of subscribing are exp(1.286)=3.618284 those of an otherwise identical contact. This means the contact is predicted to have 3.618284-1=2.618284 higher odds of subscribing.

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

## Problem 9 - Interpreting AUC

1 point possible (graded)
What is the meaning of the AUC?

- ( ) The proportion of the time the model can differentiate between a randomly selected client who subscribed to a term deposit and a randomly selected client who did not subscribe ✔

Calculator

The proportion of the time the model correctly identifies whether or not a client subscribed to a term deposit.

Explanation

The AUC is the proportion of time the model can differentiate between a randomly selected true positive and true negative.

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank

Previous Next

Calculator