Banking Predictions

Marko RobnikSikonja
University of Ljubljana
Marko.RobnikSikonja@fri.uni-lj.si

Romain CHATEAU
ESIGELEC | School of Engineering
r.chateau.12@groupe-esigelec.fr

Armand LEOPOLD
ESIGELEC | School of Engineering
a.leopold.12@groupe-esigelec.fr

Florent BACQUE
ESIGELEC |School of Engineering
f.bacque.12@groupe-esigelec.fr

November 16, 2016

Abstract

Nowadays, with the emerging of big data technologies, banks are occuring a big and very deep transformation. The bank technology space is being reshaped by a host of new forces, and enter into new sorts of relationships with their customers. Market targeting tends to use more and more algorithms in order to find affinities for more personalized products. This project focuses on making an approach using machine learning to find afinities on bank marketing data. We will use simple baseline and more complex classifiers and then evaluate accuracy and different optimization parameters in order to find the best fiting model.

I. Introduction

Personnal behaviors are partialy based on their specific caracteristics and environmental influences. Thanks to this fact, we can elaborate logical models which can predict multiple things.

In our case, we want to predict wheter or not a custumer will subscribe to a term deposit. To do that, we will have to find personal attributes (age, profession etc...) which have the most influence on their desicions.

To perform this study, we used Mr RobnikSikonja lectures and slides. We also made a lot of research on the internet mostly on different blogs, Stack Overflow and R documentation.

At first, we did some basic manipulations with the dataset in order to become familiar with it. We learnt the differences between the parameters and their meaning in the banking environment.

II. Dataset

Our data is provided by UCI machine learning repository, [1]. It is related with direct marketing campaigns of a Portuguese banking institution. This dataset provides a set of parameters which have or not subscribed to a term deposit. 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 ('yes') or not ('no') subscribed.

Attribute Information:
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’ , ’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’)

   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. **dayofweek**: 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 perfo rmed 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’)

III. Methods

We began our analysis by subsetting the data in three equals parts. These three parts are randomly selected. Here are the purposes of each splitted data part :

- First third of data will be used as training for baseline classifiers
- Second third of data will be used as training for stacking classifier
- First third of data will be used as final test

Figure 1: *Number of product contractors per age range*

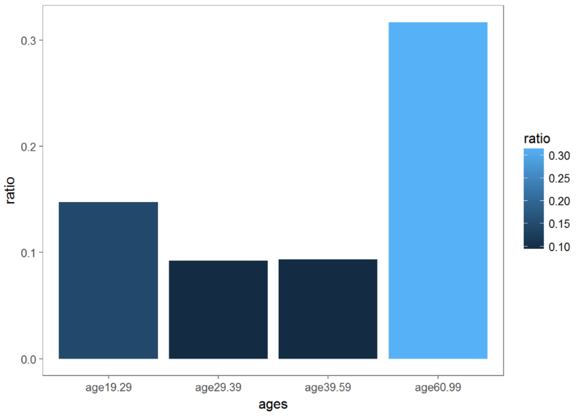
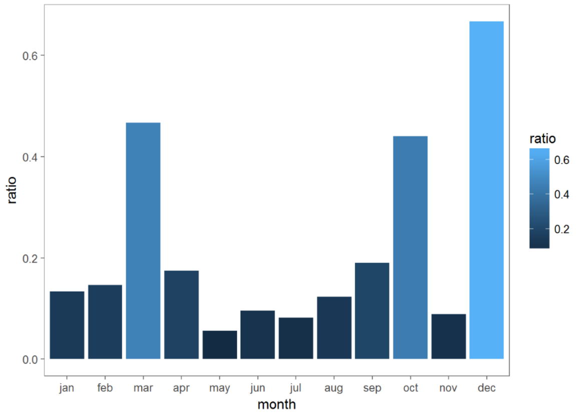


Figure 2: *Number of product contractors per months*



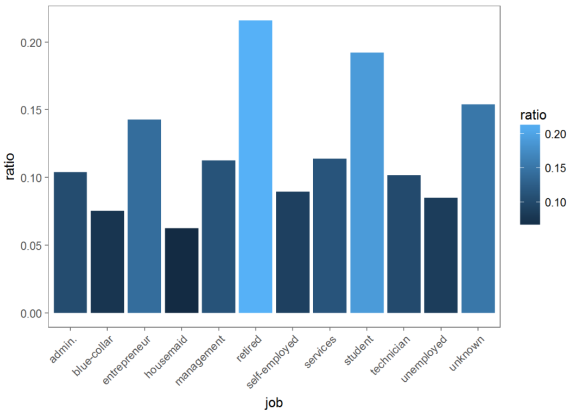
III.1. Data Overview

We made some inspection into the data. We found out that there are strong links between predictors. For instance, we can see that older people are more likely to contract a term deposit 1, which can be also seen in the job predictor as retired people 3.

After our overview, we tried to confirm these impressions by implementing predictive modeling. We want to analyse wether or not a client will subscribe to a term deposit. This is a classification situation, therefore we used the following common classifications algorithms :

- Logistic Regression
- Random Forest
- Classification Tree

Figure 3: *Number of product contractors per job*



- Naive Bayes using Bernoulli distribution
- Stacking

III.2. Logistic regression

We implement the logistic regression in its basic form, using all predictors in the formula.

III.3. Random Forest

The parameters we tuned on the model are :
- the parameter "mtry" which is calculated by $\sqrt{p}$ with p being the number of predictors (in our case $\sqrt{p} = 4$)
- the parameter "ntree" is set to 1000 which is the number of randomly generated tree

III.4. Classification tree

We did a basic analysis with the classification tree.

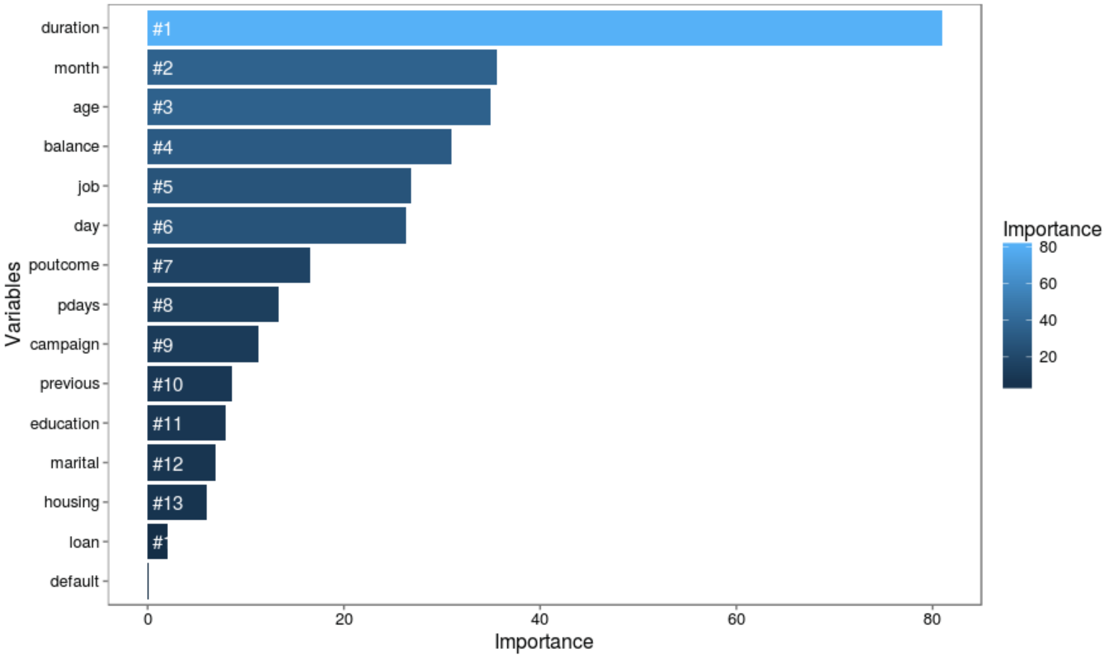
III.5. Naive Bayes using Bernoulli distribution

We did a basic analysis with the naive bayes too.

III.6. Stacking

After analysing their accuracy, specifity and sensitivity we tried to improve these baselines classificators using stacking method. In order

Figure 4: *Importance of attributes in Random Forest model*



to do that, we stacked the results of probabities predicted by baseline classifiers on the second training set.

IV. Evaluation

To evaluate our different models, we calculated the accuracy and the Area Under the ROC Curve.

V. Results

According to our results, we can see that the stacking model is less accurate than the baselines classifiers. The logistic regression model is better than the others. Table 1

VI. Marketing campaign simulation

In order to test the influence of the major attributes to find wether or not a customer will

Table 1: *Global Evaluation Table*

	Accuracy	AUC	Std. Dev
Logistic Regression	91	0.945	0.0058839
Classification Tree	89	NA	0.0061201
Naive Bayes	86	0.821	0.0090036
Random Forest	90	0.664	0.0060135
Stacking Best	89	0.928	NA

subscribe to a term deposit. We have created a new dataset of "favorable customers" with generated data. We set the age of customer to a random value between 60 and 80 years 1, the job to retired 3 and the month to december 2 (best values for the targeted predictors). The goal is to see if a targeted advertising campaign could be usefull for the bank. We did not modified the duration value because the bank can not know in advance the duration of the phone call.

We applied our random forest model to the "favorable customers" dataset and get the following results : Table 2

Table 2: *Simulated Marketing Campaign*

–	Original data	Modified data
Nbr of subs	159	619
Ratio of subs	0.1065684	0.4148794

We see that the selected predictor's based dataset gives a better result which means that the predictors have a great influence on the result. Targeting this special type of customer can lead the bank to multiply its chance of success by 4.

Finally, this analysis shows that machine learning algorithm can highlight attributes which can enhance the performance of a compagny.

VII. Deliverables

The results of our project will be displayed on a R markdown file which we used to make our analysis. We also have this report who gives deeper insight in our methodology and approach.

VIII. Conclusions

After our analysis, we have found that the following attributes are the most relevant : month, balance, age, duration, job. [2]
The best model that fit our data is logistic regression.

For future work suggestions , we could have improved the parameters of baseline classifiers to enhance their precision and performance.

Considering the unexpectedly low accuracy of our stacking model, we could replace the baseline model by a voting system, which could produce better results.

A deeper analysis of each parameter of the dataset could highlight other axes of studies. For instance the dataset provides a variable which defines the number of calls for one client during the current campaign.

References

[1] UCI - Bank Marketing Data Set UCI - 2012

[2] Link to the html version of the Rmd Link - 2016

[3] 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- 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

[4] Hany. A. Elsalamony, Alaa. M. Elsayad Bank Direct Marketing Based on Neural Network - August 2013

[5] Charles X. Ling and Chenghui Li Data Mining for Direct Marketing: Problems and Solutions - 1998

[5] Thiago S. Guzella and Walmir M. Caminhas a-review-of-machine-learning-approaches-to-spam-filtering - 2009