PH125.9x HarvardX - Capstone Examination in Data Science

Report on Direct Bank Marketing Project

Philippe Lambot

April 24, 2019

****

"CHOOSE YOUR OWN" PROJECT: DIRECT BANK MARKETING

****

****

****

“CHOOSE YOUR OWN” PROJECT: DIRECT BANK MARKETING

****

Important Foreword - Requirements

In this project, there are two programs: Script.R and Report.Rmd.

Script.R has been run in

- RStudio Version 1.1.456 - © 2009-2018 RStudio, Inc.

Report.Rmd has been knitted to HTML in

- RStudio Version 1.1.456 - © 2009-2018 RStudio, Inc.

The version of R that I use on my PC is

- R version 3.5.1 (2018-07-02) – “Feather Spray”
- Copyright (C) 2018 The R Foundation for Statistical Computing
- Platform: x86_64-w64-mingw32/x64 (64-bit)

The operating system on my PC is Windows 10.

I cannot guarantee Script.R running on other versions. Neither can I guarantee Report.Rmd being knitted to HTML on other versions.

I. Introduction

This **predictive and supervised data science project** is about **direct bank marketing** campaigns that a Portuguese bank organized by phone calls among customers to collect term deposits.

This project is presented in the framework of the Capstone Examination in Data Science organized by HarvardX (PH125.9x). It is a “choose your own” project and it is the second step from the Capstone Examination.

A. Dataset

Data has been downloaded from the *UCI Machine Learning Repository* site, following one suggestion of HavardX’s. It is real data.

The dataset is a file with 4521 observations and 17 variables. The 4521 observations relate to 4521 customers whom the bank has contacted by phone calls to propose subscribing a term deposit. Among the 17 variables, there is a variable, y, indicating whether the customer has or has not subscribed a term deposit. From the other 16 variables, 15 will be used as predictors to predict whether the subscriber will or will not subscribe a term deposit.

B. Objective of the Project and Terminology

The objective is as follows: with a model that predicts who will subscribe a term deposit and who will not, reaching, on a validation set, one of the three alternatives from the table hereunder.

Objective	Subscribers_Reached	Global_Coverage_Reduction
Alternative 1	>= 76 %	>= 49 %
Alternative 2	>= 75 %	>= 50 %
Alternative 3	>= 74 %	>= 51 %

The percentage of subscribers reached is calculated as follows: the number of customers that are predicted as subscribers by the model and that really are subscribers, divided by the number of subscribers. It is multiplied by 100 to have percentage points instead of decimals. In more technical words, the percentage of subscribers reached is sensitivity (or recall) multiplied by 100. It equals TP / (TP + FN) * 100.

The global coverage reduction is the percentage of customers that the model considers as non-subscribers. It is called “global coverage reduction” because it indicates by how much the list of customers to be contacted can be reduced while still reaching a specified percentage of subscribers. In more technical terms, the global coverage reduction is equal to (FN + TN) divided by the number of customers (and multiplied by 100).

Alternative 1 (76-49), alternative 2 (75-50) and alternative 3 (74-51) are perfect substitutes for each other and are considered as perfectly equivalent in terms of reaching the objective. If one alternative is met, then the objective of the project is fully met. For instance 76-49 on the validation set perfectly attains the objective since it meets alternative 1. Other acceptable example: 74-51 also reaches the objective since it meets alternative 3. Third acceptable example: 77-50 also attains the objective; it meets alternative 1 and alternative 2 (it even beats alternatives 1 and 2).

The objective must be reached on a 30% validation set that can only be used at the very last step with the final model. Of course, the final model cannot use, in predicting, any information coming from the validation set dependent variable, i.e. the variable indicating whether the customer has really subscribed a term deposit or not. The validation set dependent variable can only be used at the very last step to check up whether prediction is right or not and so to measure performance.

When checking up validity of results in this binary classification challenge, results are rounded to the nearest percentage point. As an example, 49.499% becomes 49%.

Outside the validation set, the remaining 70% of data can be further split, rearranged and analyzed in any way.

By the way, in this project, the words “objective” and “target” are synonyms.

C. Key Steps

There are several key steps in the project, as indicated in the table of contents on the first page from Report.pdf.

After downloading the dataset, it will be prepared and split. A 30% validation will be extracted, as required in the objective of the project. The remaining 70% of data will be further split into an 80% training set and a 20% test set, which will both be used to train models.

Another section will describe the attributes, i.e. the dependent variable and the predictors. It will be followed by exploratory analysis and visualization. Insights will be gained through this process. One predictor will be discarded because of a fundamental bias.

Four machine learning models will then be trained on the training set and on the test set. After trying a dimension reduction, performance improvement will be sought through tuning probability threshold below 0.5, with promising results. An ensemble model will be built with a view to more result stability. The whole modeling and analysis process will deliver numerous insights.

The ensemble model will then be validated on the validation set, producing the final results, which will be evaluated by comparing with the objective.

D. File and Document Organization

In this project, there are four files: Script.R, Report.Rmd, Report.pdf and bank.csv. They can all be accessed in the GitHub repository https://github.com/Dev-P-L/Bank-Marketing (https://github.com/Dev-P-L/Bank-Marketing) .

Code is included both in Script.R and in Report.Rmd. Code is very similar in both files (the exceptions being the chunk delimiters in Report.Rmd and the code generating the objective table in Report.Rmd). Script.R also contains comments about code. Results generated by both files are the same.

Report.Rmd describes data analysis, modeling, insights and results.

Knitting Report.Rmd (to HTML) generates an HTML document that contains data analysis, modeling, insights and results; that HTML document can be easily issued in PDF format as Report.pdf. For readability reasons, Report.pdf does not contain any code.

Dear Readers, you can run Script.R, knit Report.Rmd (to HTML) and issue it in PDF format. On my laptop, running Script.R takes approximately a quarter of an hour. Knitting Report.Rmd (to HTML) also takes more or less a quarter of an hour. Before running Script.R or knitting Report.Rmd to HTML, please read hereinabove “Important Foreword - Requirements”.

Some packages are required by both Script.R and Report.Rmd: *tidyverse*, *caret*, *kableExtra*, *MASS*, *randomForest* and *gbm*; the code contains instructions to download them if they are not available.

II. Downloading Data

I have downloaded data as follows.

1. Following the link: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing (http://archive.ics.uci.edu/ml/datasets/Bank+Marketing) .
2. “Clicking” on “Data Folder” to download “bank.zip”.
3. Unzipping “bank.zip”.
4. Extracting “bank.csv”.
5. Saving “bank.csv” to my GitHub repository.

Now, let’s retrieve “bank.csv” from my GitHub repository by using the read.csv() function and accessing https://raw.githubusercontent.com/Dev-P-L/Bank-Marketing/master/bank.csv (https://raw.githubusercontent.com/Dev-P-L/Bank-Marketing/master/bank.csv).

III. Preparing and Splitting Data

Let’s first adapt the symbols used as values in the target variable. The target or dependent variable is of binary categorical type *no/yes*. In order to have *yes* (subscriptions of deposits) as “positive” class in confusion matrices and performance measures (sensitivity and precision), I am going to rename the *no/yes* values as *no_deposit* and *deposit*.

Let’s create the required 30% validation set that will only be used at the very last step and only with the final model.

Let’s further split the data outside the validation set into a training set and a test set. Both will be used to train models.

There are no missing values.

IV. Description of Attributes

A. Description of Target or Dependent Variable

Let’s quickly remember the main characteristics of the target variable, y.

Variable_Name	Type	Meaning
y	Binary categorical - Response: deposit/no_deposit	Has the customer subscribed a term deposit?

B. Description of Predictors

Let’s build up a descriptive table of predictors. Which is the name of the predictors in bank.csv? Which is the type of variable? From a marketing point of view, which is the meaning of each predictor? Which type of information does each predictor bring, is it information about customer or about campaign action, is it financial or non financial?

Variable	R_Type	Meaning_in_Marketing_Campaign	Information_Type
age	integer	Age in years	Customer’s NON-FINANCIAL Profile
job	factor	Professional status	Customer’s NON-FINANCIAL Profile
marital	factor	Marital status	Customer’s NON-FINANCIAL Profile
education	factor	Education	Customer’s NON-FINANCIAL Profile
default	factor	Credit in default?	Customer’s FINANCIAL Profile
balance	integer	Account balance	Customer’s FINANCIAL Profile
housing	factor	Home loan?	Customer’s FINANCIAL Profile
loan	factor	Personal loan?	Customer’s FINANCIAL Profile

Variable	R_Type	Meaning_in_Marketing_Campaign	Information_Type
contact	factor	Medium of communication with this customer in this campaign	THIS Marketing Campaign
day	integer	Day (of the month) of last contact with this customer in this campaign	THIS Marketing Campaign
month	factor	Month of last contact with this customer in this campaign	THIS Marketing Campaign
duration	integer	Duration in seconds of last contact with this customer in this campaign	THIS Marketing Campaign
campaign	integer	Number of contacts with this customer in this campaign	THIS Marketing Campaign
pdays	integer	Number of days since last contact with this customer in previous campaign	PREVIOUS Marketing Campaign
previous	integer	Number of contacts with this customer before this campaign	PREVIOUS Marketing Campaign
poutcome	factor	Outcome from previous marketing campaign with this customer	PREVIOUS Marketing Campaign

V. Exploratory Analysis and Visualization

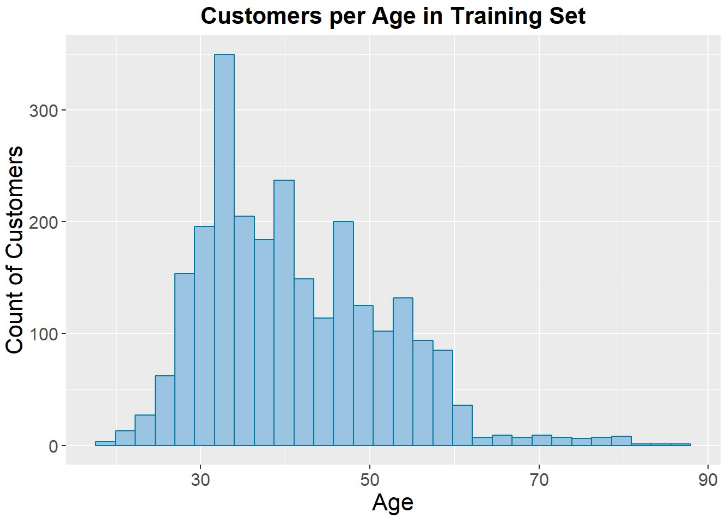
A. Target or Dependent Variable, y

y_value	Occurrences_in_Training_Set	Percentage
deposit	291	11.5
no_deposit	2240	88.5

Prevalence of “deposit” is very low, i.e. 11.5%. Consequently, a high accuracy of 0.885 would mean nothing since we would already reach 0.885 simply by predicting “no_deposit” for all customers! Other measurement tools have to be used, ensuring some levels of performance on the subgroup of subscribers, who represent 11.5% of the customers.

B. Age

Let’s build up a histogram of customers according to age.



The most populated subgroups are, in descending order, the 30-somethings, the 40-somethings and the 50-somethings. Will the same concentration be found in terms of subscription percentages? Let's have a look at a table where subscription percentages are sorted in descending order.

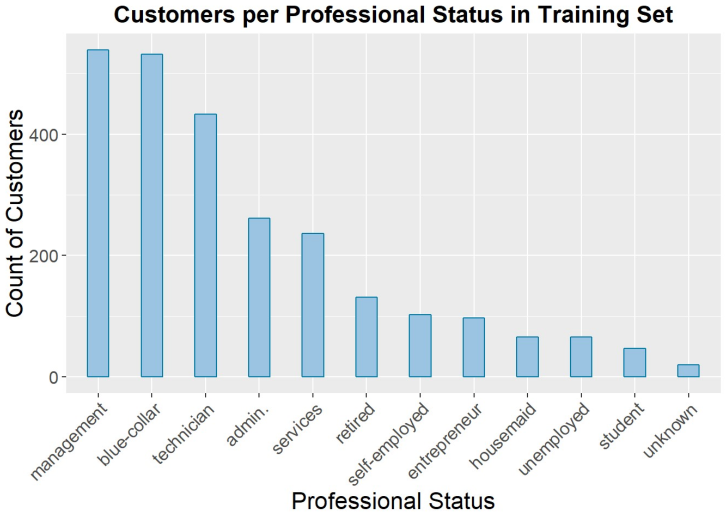
Age	Number_of_Customers	Subscription_Percentage
[61,88)	76	41
[19,25)	43	23
[25,30)	216	15
[45,50)	326	11
[55,61)	248	11
[30,35)	546	10
[35,40)	476	10
[40,45)	352	9
[50,55)	248	9

Actually, age categories that are more populated tend to have lower subscription percentages (but negative correlation is not causation!). The highest subscription percentages are noted below 30 years and above 60. This can be expected from people above 60.

Age can be an impactful predictor.

C. Job

The next predictor is job, or rather professional status since we also see retired people, students and unemployed people. Let's have a look at a barplot.



Some subgroups, such as blue-collars and managers, are much more populated than others. Let's build up a table, adding percentages of subscribers per professional status in descending order.

Professional_Status	Number_of_Customers	Subscription_Percentage
unknown	20	30
student	47	28
retired	131	26

Professional_Status	Number_of_Customers	Subscription_Percentage
management	539	14
self-employed	103	13
admin.	261	11
entrepreneur	97	11
housemaid	66	11
unemployed	66	11
technician	433	10
services	236	8
blue-collar	532	6

Retired people and students are above 25% and the subgroup with “unknown” professional status is first in subscription percentage with 30%. Blue-collars have the lowest subscription percentage, with 6%.

Partially, this intersects and corroborates insight from age: extremes in age have higher subscription percentages; this is similar information about retired people and students.

But age and professional status are no mere duplicates: the subgroup of retired people and the subgroup of customers over 60 are far from completely intersecting: there are 76 customers above 60 and 131 retired people! Moreover, there are many other subgroups!

This predictor could also be impactul in predicting.

D. Marital Status

The table hereunder gives the concentration of customers per marital status, distinguishing married people, single and divorced people, widows and widowers being considered as divorced people. It also gives the average subscription percentages per subgroup.

Marital_Status	Number_of_Customers	Subscription_Percentage
single	655	15
divorced	285	13
married	1591	10

The difference in subscription percentage is about 2-3 percentage points between groups, the lowest percentage being 10% for married people and the highest percentage being 15% for single people, divorced people being in the middle. This can indicate potential impactfulness from marital status.

E. Education

The last predictor from customers’ non-financial status is the education level, distinguishing primary, secondary, tertiary levels and unknown.

Education_Level	Number_of_Customers	Subscription_Percentage
tertiary	756	15
primary	391	10
secondary	1283	10
unknown	101	9

Customers with tertiary education have on average a subscription percentage of 15%. The other subgroups are around 9-10%. That dichotomy seems promising in terms of predictive power from that predictor.

Is there a link with professional status, e.g with managers? 756 customers with tertiary education have an average percentage of 15; 539 managers have an average percentage of 14 (see hereinabove). Is there a strong intersection between the two groups? Actually, 431 customers are managers with tertiary education with an average subscription percentage of 15, i.e. close to the average for managers and the average for people with tertiary education. So, most managers have a tertiary education level.

This does not mean that education and professional status are duplicate information. Indeed, the intersection is far from complete even for tertiary education and managers and, moreover, there are many other subgroups according to education and to professional status.

F. Credit Default

The first predictor from the financial status is credit default.

Credit_Default	Number_of_Customers	Subscription_Percentage
no	2492	12
yes	39	8

There is a difference in subscription percentage between customers in credit default and customers who are not. But the subgroup of customers in default is limited. Will this predictor be impactful?

G. Account Balance

The range between the minimum of -3,313 and the maximum of 71,188 is rather broad. Let’s have a look at a table with subgroups and average subscription percentages.

Account_Balance	Number_of_Customers	Subscription_Percentage
[-5000,-1000)	8	0
[-1000,0)	187	8
[0,0.1)	184	10
[0.1,1000)	1340	10
[1000,5000)	629	16
[5000,10000)	130	15
[10000,15000)	29	7
[15000,75000)	24	8

In the table above, tranches of account balances have not been sorted by descending order of subscription percentage because subscription percentages show a remarkable pattern, being rather centered: between 1,000 and 10,000, people are more prone to subscribe a term deposit than people with smaller or bigger account balances.

Average subscription percentages vary between 7 and 16, without taking into consideration a zero average percentage for a very small subgroup of people with negative imbalances below -1,000.

Account balance can be an effective predictor.

H. Home Loan

The third financial predictor is about the customer having a home loan or not having any.

Home_Loan	Number_of_Customers	Subscription_Percentage
no	1117	15
yes	1414	9

Clear percental difference between the two subgroups, i.e. customers with a home loan and customers without any. Moreover, both categories are substantively populated. This predictor seems a promising one.

I. Personal Loan

The fourth financial predictor is about the customer having a personal loan or not having any.

Personal_Loan	Number_of_Customers	Subscription_Percentage
no	2127	13
yes	404	6

There is a clear percental difference. This predictor seems promising even if the subgroup with a personal loan is quantitatively more limited.

J. Medium of Communication in this Campaign

Here is the first predictor linked directly to this marketing campaign, i.e. the marketing campaign whose results are to be predicted in this project.

Medium_of_Communication	Number_of_Customers	Subscription_Percentage
cellular	1607	14
telephone	174	14
unknown	750	5

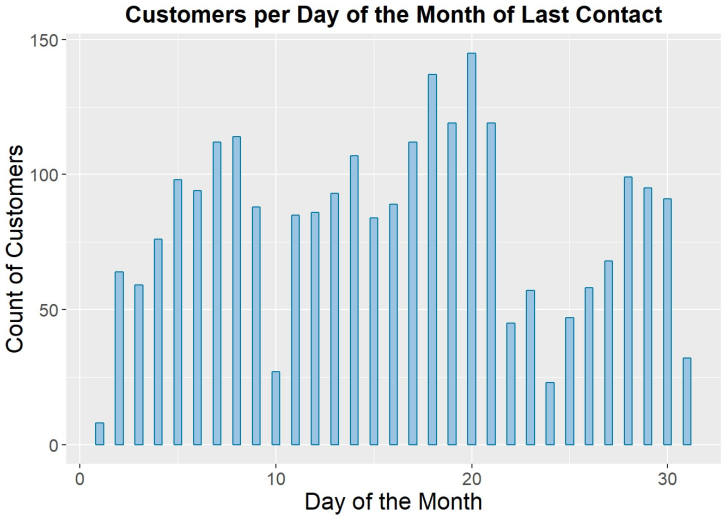
There are three subgroups: customers contacted on a cellular phone, customers contacted on a landline phone and people for whom that piece of information is unknown. In terms of average subscription percentage, there is a clear dichotomy between customers for whom that piece of information is available and customers for whom it is not. The respective levels are 14% and 5%.

This predictor seems very promising.

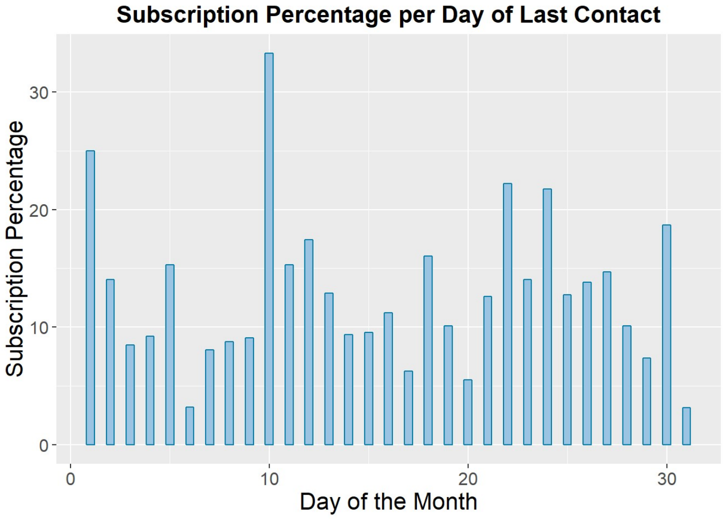
K. Day of the Month of Last Contact with Customer in this Campaign

The next predictor is the day of the month of the last contact with the customer. Maybe there are temporal patterns.

First, let’s see the distribution of the days of the month of the last contact.



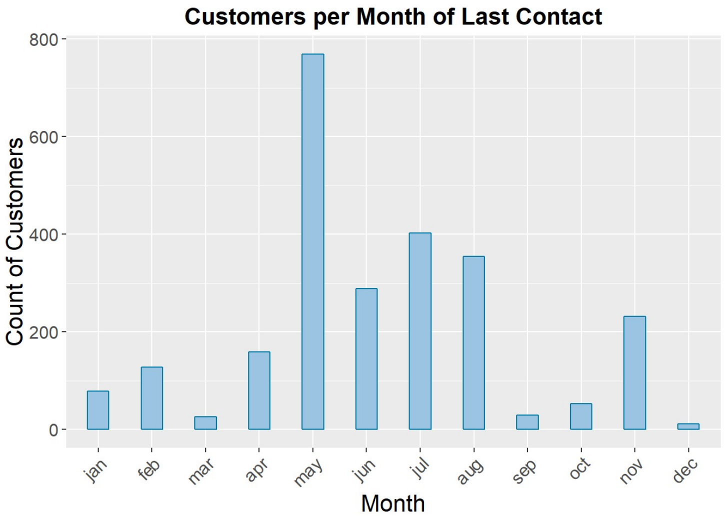
The number of customers per day of the month of the last contact is rather volatile. There are more contacts in the middle of months. Let’s have a look at the variation in subscription percentage between days of the month.



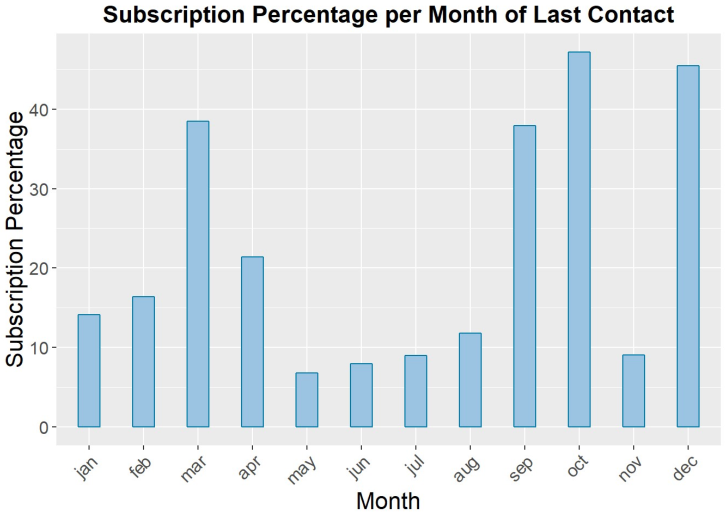
It also seems rather volatile. It looks partially counter-cyclical with respect to the number of customers per day. Will that predictor be influential?

L. Month of the Last Contact with Customer in this Campaign

The next predictor is also directly linked to this marketing campaign: it is the month of the last contact in this campaign. Is there a temporal pattern for the month of the last contact?



The last contact did often take place from May until August. But how is this related to the percentage of subscription?



The subscription percentage per month of last contact looks generally counter-cyclical with respect to the number of customers: there are several higher subscription percentages outside the period May-August.

M. Duration in Seconds of Last Contact with Customer

It is the duration of the contact whose results are to be predicted.

There is a strong relationship between contact duration and subscription probability. But using it here would be a bias. Indeed, the duration of the contact whose results are to be predicted is not known before that contact. Consequently, that piece of information cannot be considered as a predictor of subscription since it is known only after contact.

For that reason, and in complete agreement with the recommendations available on the site, this variable will not be taken into account. It will be discarded from the training set, from the test set and from the validation set.

O. Number of Contacts with Customer during this Campaign

The number of contacts per customer can vary very much as the table hereunder shows it. For most customers, the number of contacts in this campaign is one, two or three. But the number can attain 50! That range is rather broad.

Is the subscription percentage in a subgroup related to the number of contacts?

Number_of_Contacts_in_this_Campaign	Number_of_Customers	Subscription_Percentage
[1,2)	960	13
[2,3)	729	13
[3,5)	482	11
[5,10)	260	6
[10,51)	100	5

In the table above, there are five subgroups constituted on basis of the number of contacts in this campaign. In the subgroup of customers with one contact and in the subgroup with two contacts, the subscription percentage is 13%. The subscription percentage decreases somewhat in the subgroup with three or four contacts. But in the subgroups with more than four contacts, the subscription percentage drops dramatically, landing at 5% above 9 contacts.

Consequently, that table shows, at the level of subgroups, a clear negative link between number of contacts and subscription percentage, even if we perfectly know that correlation is not causation!

This might make sense, though! If there have already been e.g. seven or eight contacts, I can imagine that there is probably some hesitation from the customer!

P. Number of Days since Last Contact with Customer in Previous Campaign

This is the first predictor related to previous marketing action, thus before “this” campaign.

Days_since_Last_Contact_in_Previous_Campaign	Number_of_Customers	Subscription_Percentage
[-1,0)	2064	9
[0.1,50)	6	33
[50,100)	66	47
[100,250)	208	20
[250,500)	181	15
[500,1000)	6	33

In the table above, we can see that the number of days since the last contact before this campaign is very often - 1! Actually, the value - 1 means that there had been no contact before this campaign. It is so for the vast majority of customers, i.e. for 2,064 customers out of 2,531!

This is correlated with the average subscription percentages (we can state it even if we know that correlation is not causation!). There is a major dichotomy between the subgroup “no contact before this campaign” and the subgroup “one or more contacts before this campaign”. Indeed, customers without any contact in a previous marketing campaign respond favorably on average in 9% of cases against 15%, 20%, 33% or even 47% in case of previous contact(s). Is this so surprising? Selling on the phone to customers who have not been contacted in that way before can prove to be a tough challenge. A promising piece of information can be the dichotomy between “previous contact” and “no previous contact”.

In case of previous contact, less than 100 days seems more promising than more than 100 days, which also makes sense.

Q. Number of Contacts with Customer before this Campaign

This predictor will be partially redundant with respect to the previous one. Indeed, here, once again, we will find the scenario of customers not having been contacted before the campaign under review.

Number_of_Previous_Contacts	Number_of_Customers	Subscription_Percentage

Number_of_Previous_Contacts	Number_of_Customers	Subscription_Percentage
[0,0.1)	2064	9
[0.1,5)	385	21
[5,10)	67	27
[10,26)	15	27

In the table above, we can find back the 2,064 customers who had not been contacted before the campaign under analysis.

Zero previous contact is a very negative factor. This duplicates information from previous predictor, i.e. the number of days since the last contact before this campaign. In the case of the previous predictor (number of days), the average subscription percentages in subgroups with previous contact was rather "bumpy" from one subgroup to another. In the case of the number of previous contacts, there appears less variation among subgroups with previous contact as they are constituted.

R. Outcome from Previous Marketing Campaign

This is the last predictor. Of course, we will find once again the same piece of information about the 2,064 customers who had not been contacted before. But maybe there is some additional information.

Outcome_from_Previous_Campaign	Number_of_Customers	Subscription_Percentage
failure	286	13
other	103	19
success	78	60
unknown	2064	9

There is, indeed, some very interesting new information. Apart from the 2,064-customer subgroup, the subgroup of customers who were previously contacted and who previously subscribed term deposits shows an extraordinary subscription rate of 60%. That looks like a piece of information with very impactful predictive power.

VI. Insights from Exploratory Analysis and Data Preparation

A. Low Prevalence of "deposit" in the Target Variable y

There are 88.5% "no_deposit" responses in the training set. Predicting "no_deposit" for all customers would already deliver an accuracy level of 88.5%. But this would also deliver a sensitivity of zero!

This is at the opposite of the objective, which consists in reaching one of the equivalent alternatives from the objective table reprinted hereunder (equivalent from the point of view of reaching the objective).

Objective	Subscribers_Reached	Global_Coverage_Reduction
Alternative 1	>= 76 %	>= 49 %
Alternative 2	>= 75 %	>= 50 %
Alternative 3	>= 74 %	>= 51 %

Alternative 1 requires a sensitivity of 0.76, alternative 2 a sensitivity of 0.75 and alternative 3 a sensitivity of 0.74. Consequently, a sensitivity level of zero would be absolutely incompatible with the objective in this project.

Even an accuracy level of e.g. 0.92 would be compatible with any sensitivity value in a range from 0.304 to 1, consequently with a lot of sensitivity values smaller than 0.74. The value 0.304 is coming from (92 - 88.5) / 11.5.

In consequence, accuracy is not the right performance measure in this challenge. It is about percentage of subscribers reached and global coverage reduction (please see hereinabove the section “II. Objective of the Project and Terminology”).

B. Insights about Predictors

Each predictor provides a segmentation of customers into subgroups, each subgroup having its average subscription percentage. In the exploratory analysis, some remarkable subgroups have been noted from the point of view of subscription percentages:

- customers younger than 30 or older than 60 have on average higher subscription percentages than the other subgroups formed on the basis of age;
- the same holds for students, retired people and customers with “unknown” professional status while, on the contrary, the subgroup of blue-collars has the lowest subscription percentage with respect to other “professional” subgroups;
- the subgroup of married people has a lower subscription percentage than single and divorced people;
- on the contrary, customers with tertiary education form a subgroup with the highest subscription percentage with respect to other education levels;
- it is the same for customers with an account balance between 1,000 and 10,000;
- the subgroups of customers with home loans or personal loans have lower subscription percentages;
- the same holds for customers for whom the medium of communication is unknown!
- we can see higher subscription percentages than average when the last contact in this campaign is in March, September, October or December and lower from May until August;
- the subgroup of customers who have been contacted but less than five times in the campaign under review has a higher average subscription percentage;
- segmentation on the basis of previous marketing action indicates the following: customers who favorably responded in a previous marketing campaign have on average a much higher subscription rate ; the same holds, but to a lesser extent, for customers that had been contacted in a previous marketing campaign less than 100 days before the last contact in this campaign; but customers who had not been contacted in previous marketing action, which is a vast majority, have a much lower subscription percentage on average than the other subgroups.

Duration should not be present as a predictor. “this attribute highly affects the output target (e.g., if duration = 0 then y =”no_deposit“). 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.” (https://archive.ics.uci.edu/ml/datasets/Bank+Marketing (https://archive.ics.uci.edu/ml/datasets/Bank+Marketing) , retrieved on 2019-04-16)

Therefore, duration will be discarded from the training set, the test set and the validation set.

Let’s express a caveat about some subgroups and their subscription percentages: some subgroups are rather small and this might affect their predictive power on the test set or the validation set.

We are now ready to turn to modeling on the training set.

VII. Analysis, Modeling and Insights on the Training Set

Let’s use the training set to try several machine learning models available in the caret package. They will be tested later on the test set before a final model is chosen and then validated on the validation set.

A. Running 4 Machine Learning Models on the Training Set

1. Picking up Models and Getting Results

Numerous machine learning models have been tried, such as qda, knn, kknn, svmLinear, svmRadial, svmRadialCost, svmRadialSigma, rpart, ranger, wsrf, Rborist, monmlp, etc. After many trials, four models have been selected: glm, lda, rf and gbm.

The four selected models will be run with the train() function from the caret package.

For each model, two measures will be computed:

- the percentage of subscribers reached: sensitivity or recall multiplied by 100;

- the global coverage reduction: percentage of customers for whom the model prediction is “no_deposit”.

These are the two criteria to use to evaluate model performance.

Model	Subscribers_Reached	Global_Coverage_Reduction
gbm	14 %	97 %
glm	15 %	97 %
lda	24 %	95 %
rf	100 %	89 %

2. Analysis of Results on the Training Set, Insights and Ways Forward

There is a very broad range of “predictive” quality on the training set.

On the one hand, in percentage of subscribers reached, only rf meets the target, with 100% against 14%, 15% and 24% for the other ones. On the other hand, in global coverage reduction, all models largely meet the target. Globally, when taking both criteria into account, as required, **rf meets the target, the other 3 models are far from it**.

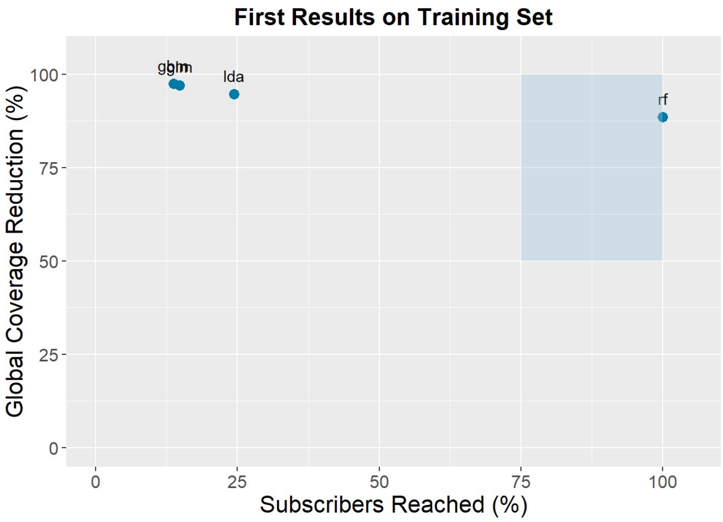
Actually, there are **two caveats**: three models do not reach the target; the model that does might be overfitting.

Let’s start with the second caveat. The performance from rf in percentage of subscribers reached is combined with high global coverage reduction! Let’s check up on the probability of model rf overfitting by building up a table with FP and FN.

Model	FP	FN
glm	30	248
lda	62	220
rf	0	0
gbm	24	251

For rf model, there is no false positive and no false negative! Consequently, accuracy, sensitivity, positive predictive value, etc. are all equal to 1! Isn’t it too perfect? Running rf on the test set will bring additional information about possible overfitting.

Let’s turn now to the first caveat, i.e. 3 models missing the target area. Performance from the various models can be visualized on the following graph.



The target area is approximately represented by the light blue rectangle in the upper right corner.

Each model is represented by a point with its label. The coordinates of the points are calculated with the default threshold 0.5 to decide whether the fitted value for a customer should be “deposit” or “no_deposit”.

Performance should be upgraded for the 3 models that do not at all meet the target.

B. Dimension Reduction

I have first tried dimension reduction. I have limited the number of predictors by using information from the glm and the rf models.

From the glm model, I have taken the 14 predictors that had some statistical significance (p-value < 0.1). From the rf model, I have taken the first 20 predictors delivered by the function varImp().

I have added a predictor.

This has given the following list: age, jobblue-collar, jobretired, jobstudent, jobtechnician, jobunknown, maritalmarried, educationsecondary, educationtertiary, balance, housingyes, loanyes, contactunknown, day, monthjan, monthmay, monthjul, monthaug, monthoct, monthnov, campaign, pdays, previous, poutcomesuccess, poutcomeother.

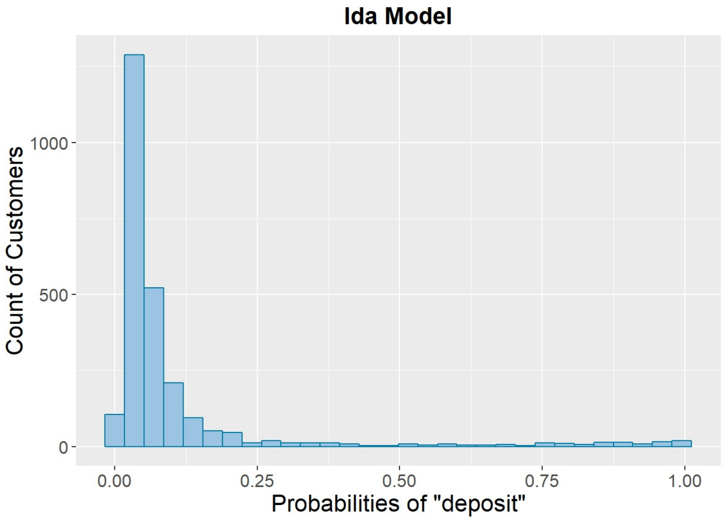
This list largely intersects the insights from the exploratory analysis.

In fact, with this dimension reduction, model performance has worsened. Let’s explore another avenue of research to improve performance.

C. Tuning the Probability Threshold

1. Retrieving and Analyzing Probabilities

Let’s retrieve probabilities for each model and analyze e.g. probabilities from the lda model.



On the graph above, there are very few probabilities larger than 0.5. To select more customers, and hopefully more subscribers, the probability threshold has to be lowered.

2. Optimizing the Probability Threshold on the Training Set

Cross-validation will be performed by tuning the probability threshold below 0.5. This will be done for all models, even if the rf model already meets the target on the training set. Let’s produce for each model and for each threshold the two performance measures that are applied in this challenge. The table hereunder shows performance measures for a limited number of thresholds values.

	Thresh	glm_s	glm_r	lda_s	lda_r	rf_s	rf_r	gbm_s	gbm_r
1	0.05	92	27	77	54	100	66	99	4

	Thresh	glm_s	glm_r	lda_s	lda_r	rf_s	rf_r	gbm_s	gbm_r
11	0.06	87	36	69	63	100	71	99	15
21	0.07	81	44	65	69	100	74	94	30
31	0.08	78	51	63	74	100	77	87	44
41	0.09	74	57	60	77	100	80	79	58
51	0.1	70	63	54	80	100	82	71	69
61	0.11	65	68	50	82	100	83	62	78
71	0.12	63	72	48	84	100	83	57	81
81	0.13	60	75	46	85	100	84	53	84
91	0.14	58	78	44	87	100	85	50	85
101	0.15	54	81	43	87	100	86	48	86
111	0.16	51	83	43	88	100	86	46	88
121	0.17	47	85	41	89	100	86	45	89
131	0.18	45	86	41	89	100	87	43	89
141	0.19	43	87	40	90	100	87	40	90
151	0.2	42	88	39	91	100	87	40	91
161	0.21	41	89	38	91	100	87	39	91
171	0.22	40	90	37	92	100	87	38	91
181	0.23	38	90	36	92	100	87	36	92
191	0.24	36	91	35	92	100	88	36	92
201	0.25	34	92	35	92	100	88	34	92
211	0.26	34	92	34	92	100	88	34	92
221	0.27	32	93	32	92	100	88	33	93
231	0.28	31	93	31	93	100	88	33	93
241	0.29	29	94	31	93	100	88	32	93
251	0.3	29	94	31	93	100	88	30	94
261	0.31	27	94	30	93	100	88	29	94
271	0.32	27	94	30	93	100	88	26	94
281	0.33	26	95	29	93	100	88	26	95
291	0.34	25	95	29	93	100	88	24	95
301	0.35	24	95	29	94	100	88	23	95
311	0.36	24	95	28	94	100	88	23	95
321	0.37	22	95	28	94	100	89	22	96
331	0.38	22	96	27	94	100	89	21	96
341	0.39	22	96	27	94	100	89	19	96
351	0.4	21	96	26	94	100	89	18	96
361	0.41	21	96	26	94	100	89	18	96
371	0.42	20	96	26	94	100	89	18	96
381	0.43	20	96	26	94	100	89	18	97
391	0.44	19	96	26	94	100	89	18	97
401	0.45	18	97	26	94	100	89	16	97
411	0.46	17	97	25	95	100	89	15	97
421	0.47	17	97	25	95	100	89	15	97

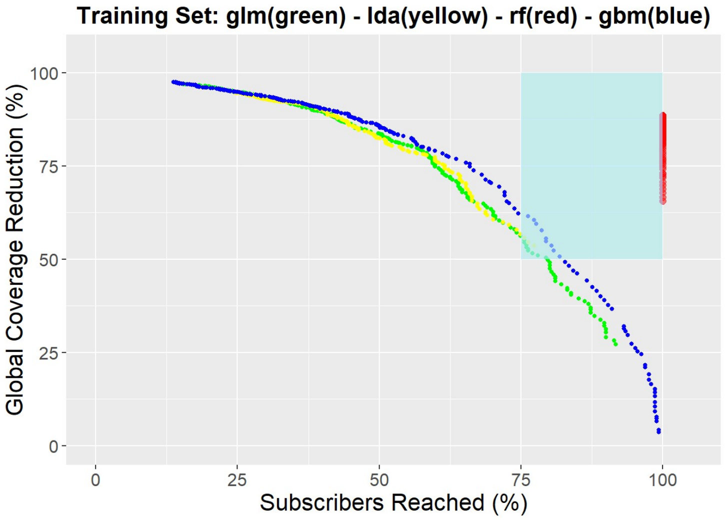
	Thresh	glm_s	glm_r	lda_s	lda_r	rf_s	rf_r	gbm_s	gbm_r
431	0.48	17	97	25	95	100	89	14	97
441	0.49	15	97	25	95	100	89	14	97
451	0.5	15	97	24	95	100	89	14	97

In the table above, “glm_s” heads the column of percentages of subscribers reached (“s” standing of course for “subscribers”) for the glm model and for some thresholds. The equivalent holds of course for lda, etc.

“glm_r” heads the column of global coverage reductions (“r” standing of course for “reduction”) for the glm model and for the same thresholds. The equivalent holds of course for lda, etc.

For the concepts of percentage of subscribers reached and of global coverage reduction, please see above the section “II. Objective of the Project and Terminology”.

Now, the 4 models meet the target: glm, lda, rf and gbm. Let’s graphically visualize.



By the way, in the graph above the diameter and transparency of points have been chosen for visual clarity and not for reasons linked to data.

3. Analysis of Results on the Training Set, Insights and Ways Forward

The graph above shows that the 4 individual models meet the target, in different ways.

The rf model, represented as a vertical red line on the right side of the graph, fully meets the objective: all points from the rf model are in the target area. The higher the threshold, the higher the points. There can be overfitting.

The gbm model is represented by bright blue points. The series of points moves up and left in a curvy way as the threshold increases. Points are rather close to each other at the top of the curve but are most distant from each other in lower parts of the series.

The glm model, represented in green, also passes through the target area, a little bit more “off-center” than gbm.

The lda model is represented in yellow. It slightly touches the target area.

The 4 models will be run on the test set.

By the way, I have not built up an ensemble model with the four successful models because of the patterns of the series representing the rf model.

VIII. Testing and Insights on the Test Set

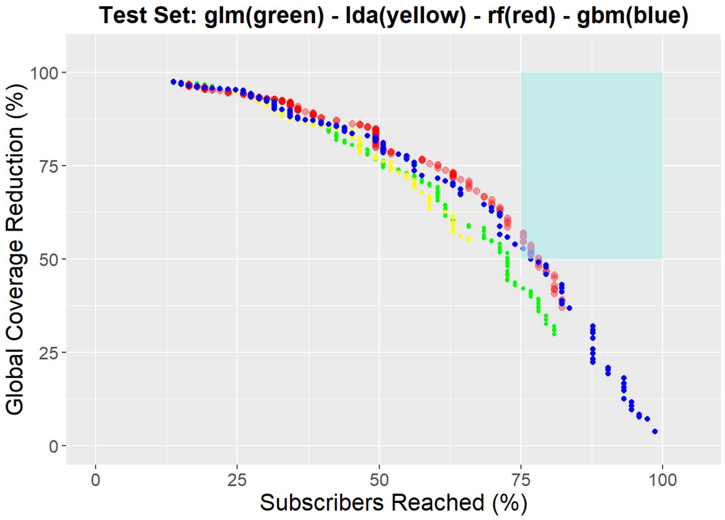
On the test set, let’s use the predict() function from caret to retrieve probabilities of “deposit”. Let’s then select a sequence of thresholds for probabilities, compute for each threshold and for each model the percentage of subscribers reached and the global coverage reduction, analyze results and possibly move forward.

A. Running the 4 Machine Learning Models on the Test Set - Insights

	Thresh	glm_s	glm_r	lda_s	lda_r	rf_s	rf_r	gbm_s	gbm_r
11	0.06	78	38	59	63	81	42	93	16
12	0.061	78	39	59	65	81	42	93	17
13	0.062	78	39	59	65	81	43	93	18
14	0.063	77	40	59	66	81	43	90	19
15	0.064	77	41	59	67	81	45	90	20
16	0.065	77	41	58	68	81	45	90	21
17	0.066	77	41	56	69	81	46	88	22
18	0.067	75	42	56	70	81	46	88	23
19	0.068	74	43	56	70	79	47	88	25
20	0.069	74	44	56	70	79	47	88	26
21	0.07	73	44	56	70	79	47	88	29
22	0.071	73	45	56	71	79	47	88	30
23	0.072	73	45	56	71	78	48	88	31
24	0.073	73	46	55	72	78	48	88	32
25	0.074	73	47	55	72	78	49	84	37
26	0.075	73	48	55	73	78	49	82	38
27	0.076	73	49	53	73	78	49	82	39
28	0.077	73	49	52	74	78	49	82	41
29	0.078	73	50	52	74	78	50	82	42
30	0.079	73	52	52	74	78	50	82	43
31	0.08	71	52	52	74	77	51	79	46
32	0.081	71	52	52	75	77	51	79	48
33	0.082	71	53	52	76	77	52	79	48
34	0.083	71	54	51	76	77	52	78	49
35	0.084	70	55	51	76	77	52	77	50
36	0.085	70	55	49	77	77	52	77	52
37	0.086	68	55	49	77	77	53	75	53
38	0.087	68	56	49	77	77	53	74	54
39	0.088	68	56	49	78	77	54	73	56
40	0.089	68	57	47	79	77	54	71	57
41	0.09	68	58	47	79	75	55	71	59
42	0.091	66	59	47	79	75	55	71	61
43	0.092	66	59	47	79	75	56	71	62

	Thresh	glm_s	glm_r	lda_s	lda_r	rf_s	rf_r	gbm_s	gbm_r
44	0.093	63	60	47	80	75	56	70	63
45	0.094	63	61	47	80	75	57	70	64
46	0.095	63	61	47	81	75	57	68	65
47	0.096	62	61	47	81	73	59	64	67
48	0.097	62	62	47	81	73	59	64	68
49	0.098	62	63	47	81	73	60	64	69
50	0.099	62	64	47	82	73	60	63	70
51	0.1	62	64	47	82	73	60	63	70

2 models meet the target: rf and gbm. Let’s graphically visualize.



2 models meet the target: rf and gbm.

gbm passes through the target area with 5 points.

rf passes through the target area with 22 points. But it has also shown important variation from the training set to the test set with seeming overfitting on the training set.

glm and lda no longer cross the target area. glm is closer to the target area than lda.

Let’s build an ensemble model to try to ensure more security in reaching the target area on the validation set. Indeed, rf and gbm only slightly scratch a corner of the target area. Moreover, rf shows large variation between training set and test set.

B. Ensemble Model and Insights

I have tried several combinations of individual models on the test set. I have opted for the combination of glm, rf and gbm.

For each threshold, for each observation from the test set, a majority vote is organized: for each threshold, the predicted value for a customer is “deposit” if at least two out of the three models give for that customer a probability of “deposit” larger than the threshold. Consequently, the code will generate 451 series of 633 predicted values since there are 451 thresholds values between 0.05 and 0.5 with increments of 0.001 and 633 rows in bank_test.

For each series of predicted values, two performance measurements will be calculated: the percentage of subscribers reached and the global coverage reduction.

Here is a table of results comparing performance from the best performing individual models (glm, rf and gbm) and from the ensemble model.

Thresh	glm_s	glm_r	rf_s	rf_r	gbm_s	gbm_r	ens_s	ens_r
0.05	81	30	82	37	99	4	89	18
0.051	81	31	82	37	99	4	89	18
0.052	81	31	82	39	97	7	89	19
0.053	81	32	82	39	96	8	89	19
0.054	79	33	82	39	96	8	89	20
0.055	79	34	82	39	95	10	89	21
0.056	79	35	81	41	95	11	88	23
0.057	78	36	81	41	95	12	88	24
0.058	78	37	81	42	93	12	88	25
0.059	78	37	81	42	93	15	88	25
0.06	78	38	81	42	93	16	88	26
0.061	78	39	81	42	93	17	88	26
0.062	78	39	81	43	93	18	88	27
0.063	77	40	81	43	90	19	85	28
0.064	77	41	81	45	90	20	85	29
0.065	77	41	81	45	90	21	85	29
0.066	77	41	81	46	88	22	84	30
0.067	75	42	81	46	88	23	84	30
0.068	74	43	79	47	88	25	82	32
0.069	74	44	79	47	88	26	82	32
0.07	73	44	79	47	88	29	82	34
0.071	73	45	79	47	88	30	82	35
0.072	73	45	78	48	88	31	81	36
0.073	73	46	78	48	88	32	81	37
0.074	73	47	78	49	84	37	81	40
0.075	73	48	78	49	82	38	79	41
0.076	73	49	78	49	82	39	79	42
0.077	73	49	78	49	82	41	79	44
0.078	73	50	78	50	82	42	79	44
0.079	73	52	78	50	82	43	79	45
0.08	71	52	77	51	79	46	79	47
0.081	71	52	77	51	79	48	79	48
0.082	71	53	77	52	79	48	79	49
0.083	71	54	77	52	78	49	78	50
0.084	70	55	77	52	77	50	78	51
0.085	70	55	77	52	77	52	78	52
0.086	68	55	77	53	75	53	77	53
0.087	68	56	77	53	74	54	77	54
0.088	68	56	77	54	73	56	75	55
0.089	68	57	77	54	71	57	75	57
0.09	68	58	75	55	71	59	75	58

Thresh	glm_s	glm_r	rf_s	rf_r	gbm_s	gbm_r	ens_s	ens_r
0.091	66	59	75	55	71	61	73	60
0.092	66	59	75	56	71	62	73	61
0.093	63	60	75	56	70	63	71	61
0.094	63	61	75	57	70	64	71	62
0.095	63	61	75	57	68	65	71	62
0.096	62	61	73	59	64	67	66	64
0.097	62	62	73	59	64	68	66	65
0.098	62	63	73	60	64	69	66	67
0.099	62	64	73	60	63	70	66	67
0.1	62	64	73	60	63	70	66	68

The **glm model** does not meet the target; it shows stability between results on training and test sets.

The **gbm model** meets the target for 5 probability thresholds between 0.083 and 0.087, with performance ranging between 78-49 and 74-54; it shows homogeneity between results on the training set and results on the test set.

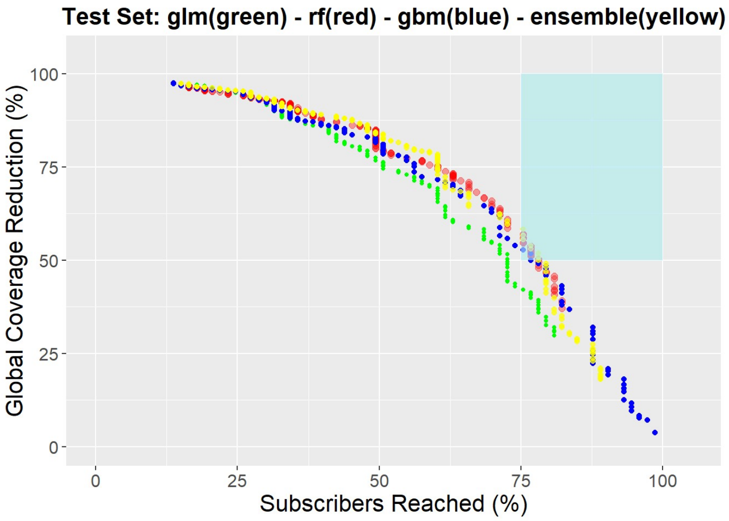
The **rf model** meets the target for 22 probability thresholds between 0.074 and 0.095 with performance ranging between 78-49 and 75-57; it shows huge variation between results on the training set and results on the test set; on the training set, sensitivity remained at 1 for all probability thresholds between 0.05 and 0.5. There was apparent overfitting.

The **ensemble model** meets the target for 9 probability thresholds between 0.082 and 0.090, with performance ranging between 79-49 and 75-58. It reaches slightly higher scores than rf but is successful on less probability thresholds. Since the ensemble model is supported by three models, of which two (glm and gbm) are rather stable in results on training set and test set, I have opted for the ensemble model as final model to be applied to the validation set.

For the ensemble model, the optimum threshold has been fixed at 0.086. How? It is the mean and the median of the thresholds that deliver performance that meet the target for the ensemble model. These probability thresholds range from 0.082 to 0.090, so the mean and the median are 0.086.

The glm model does not attain the target area but it improves the ensemble model performance.

Let’s graphically visualize results from the ensemble model and from the three subcomponents, i.e. the glm, the rf and the gbm models.



I had also developed another ensemble model. I had organized majority vote among three individual series of predictions already individually optimized with three different thresholds, one series for glm, one for rf and one for gbm. This alternative ensemble model underperformed with respect to the chosen ensemble model.

IX. Final Results - Validating the Final Model on the Validation Set

From the models trained on the training set, let's get probabilities of "deposit" on the validation set for glm, rf and gbm. Then, an ensemble model will be built on the validation set just as it was on the test set, with a majority vote and a threshold of 0.086 obtained on the test set (see above).

Let's remember the objective, quantified and summarized in the objective table, which is reprinted here.

Objective	Subscribers_Reached	Global_Coverage_Reduction
Alternative 1	>= 76 %	>= 49 %
Alternative 2	>= 75 %	>= 50 %
Alternative 3	>= 74 %	>= 51 %

Let's analyse the final results on the validation set. The challenge was to meet one out the three alternatives from the table above, alternatives considered as equally valid.

Model	Probability_Threshold	Subscribers_Reached	Global_Coverage_Reduction
Ensemble glm-rf-gbm	0.086	77 %	52 %

The final results fully meet all alternatives. They not only meet all alternatives but they do better than all alternatives. Indeed, in the final results, the percentage of the subscribers reached is 77, thus higher than the percentage of subscribers in any alternative. Moreover, in the final results, global coverage reduction is 52%, which is also higher than in any alternative.

So, the final results not only meet the objective but do even better.

In this scenario, by contacting only customers for whom the ensemble model has predicted "deposit", i.e. 48% (100 - 52) of the customers, the bank would reach 77% of the customers who have really subscribed term deposits.

In a nutshell, 48% of customers contacted, 77% of subscribers reached.

X. Conclusion

At the end of this project, there are **meaningful results both from a marketing and from a data science point of view**.

In direct bank marketing, it seems interesting: **sampling less than 50% of a population of customers and still liaising with more than three-fourths of customers who are interested in the product you are trying to promote**.

In data science, it has been an opportunity to get in touch with real data from the banking sector, which is by international standards a sector highly interested in data science.

From an analytical point of view, exploratory analysis has delivered statements that are sometimes truisms but also sometimes surprising, which means that nothing in data should be taken for granted.

Cross-validation by tuning the probability threshold has proved very powerful.

Using a test set, even of limited size, showed itself decisive in reaching, and even exceeding the target. Running models on the test set has allowed to get rid of overfitting in the case of one model and to deselect a less performing model. Furthermore, it has allowed to build up an ensemble model, which has been very productive, just as cooperation in a team can make a difference.

Combining models demonstrated that, just as in the case of data, there can be surprises: in a specific context, some computationally demanding models such as Random Forests can prove performing but it is also possible to get help from less demanding models such Logistic Regression, which in this project was individually less performing but nevertheless a decisive contributor to the ensemble model.

As a conclusion, this project has proved very stimulating and enriching and is a **powerful incentive towards business and data science**.

****