Predict if client will subscribe to direct marketing campaign for a banking institution

By
Nikita Jindal
Zoheb Shaikh
Kushal Shah
Gaurav Gargi
Vinit Dhamale

What is Problem?

Data Set Information:

- Direct marketing campaigns of a Portuguese banking institution. 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.
- There are four datasets:
- 1. bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]
- 2. bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.
- 3. bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less
- inputs).
- 4. bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs). The smallest datasets are provided to test more computationally demanding machine

- Goal :- The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

Agenda

- EDA
- Basic Model
- Model with under and oversampling
- Model with Treating missing values with simple methods
- Model with best feature
- Model with imputed values from other dataset
- Realistic Model- Without duration, week, Month

EDA

- For this we are using ‘Bank additional full’ data set.
- Originally 21 columns and 41,118 rows.
- Numerical columns: 10
- Categorical columns: 11
- Missing values: in 6 columns. Out of which 3 are numerical and 3 are categorical.

Columns with missing values

```
age                  0
job                330
marital             80
education         1731
default           8597
housing            990
loan               990
contact              0
month                0
day_of_week          0
duration             0
campaign             0
pdays                0
previous             0
poutcome             0
emp.var.rate         0
cons.price.idx       0
cons.conf.idx        0
euribor3m            0
nr.employed          0
y                    0
dtype: int64
```

Positive and Negative correlation

y_yes	1.000000
duration	0.405274
poutcome_success	0.316269
previous	0.230181
month_mar	0.144014
month_oct	0.137366
month_sep	0.126067

month_may	-0.108271
cons.price.idx	-0.136211
contact_telephone	-0.144773
poutcome_nonexistent	-0.193507
emp.var.rate	-0.298334
euribor3m	-0.307771
pdays	-0.324914
nr.employed	-0.354678

Which job groups to target?

1. Student
2. Retired
3. Unemployed

y	no	yes
job		
admin.	87.027442	12.972558
blue-collar	93.105684	6.894316
entrepreneur	91.483516	8.516484
housemaid	90.000000	10.000000
management	88.782490	11.217510
retired	74.767442	25.232558
self-employed	89.514426	10.485574
services	91.861930	8.138070
student	68.571429	31.428571
technician	89.173958	10.826042
unemployed	85.798817	14.201183
unknown	88.787879	11.212121

Person took loan or not?

There is no impact of this variable.

y	no	yes
loan		
no	88.659794	11.340206
unknown	89.191919	10.808081
yes	89.068502	10.931498

Person have his/her home on loan?

y	no	yes
housing		
no	89.120395	10.879605
unknown	89.191919	10.808081
yes	88.380608	11.619392

Campaign

For single campaign 10 calls can be a benchmark

campaign	
1	13.037071
2	11.456954
3	10.747051
4	9.392682
5	7.504690
6	7.660878
7	6.041335
8	4.250000
9	6.007067
10	5.333333

Before campaign how many calls should be there?

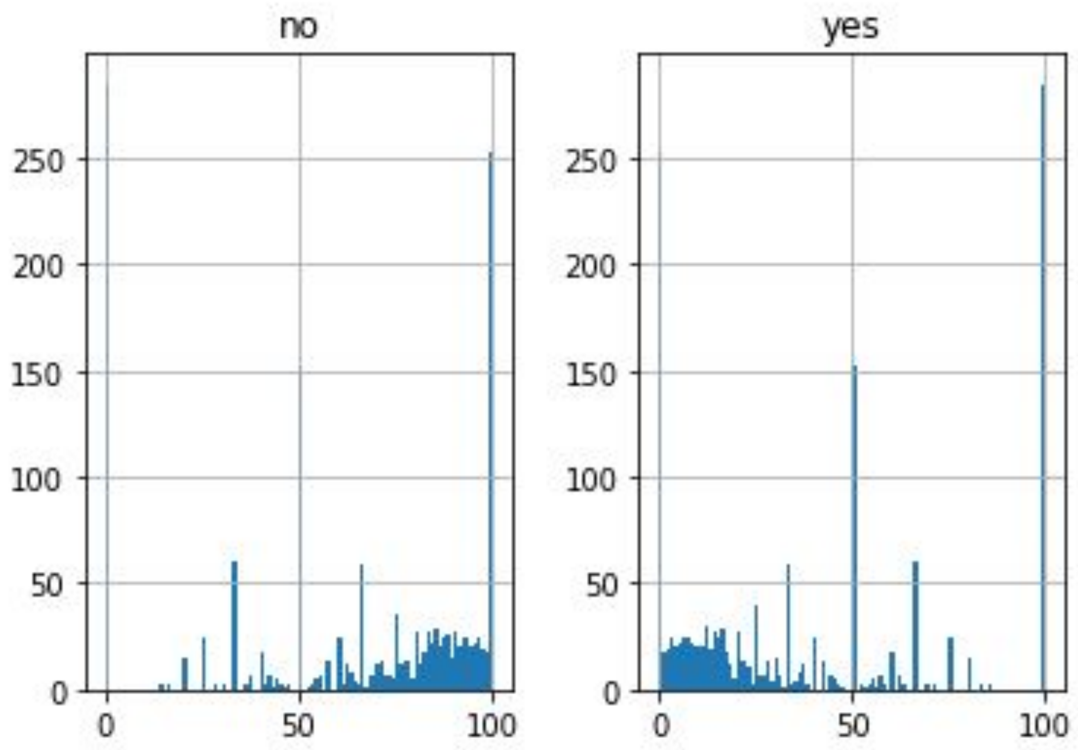
5-6 is optimum number of calls.

But even 2 single calls can change scenario

y previous	no	yes
0	91.167787	8.832213
1	78.798509	21.201491
2	53.580902	46.419098
3	40.740741	59.259259
4	45.714286	54.285714
5	27.777778	72.222222
6	40.000000	60.000000
7	100.000000	0.000000

Duration

Customer conversion will happen in first 30 seconds.



Previous campaign's outcomes

Previous successful converted customers have 65% higher chance to get conversion again

y	no	yes
poutcome		
failure	85.771402	14.228598
nonexistent	91.167787	8.832213
success	34.887109	65.112891

Seasonal Factor

1. March
2. Dec
3. Oct

y	no	yes
month		
apr	79.521277	20.478723
aug	89.397863	10.602137
dec	51.098901	48.901099
jul	90.953443	9.046557
jun	89.488530	10.511470
mar	49.450549	50.549451
may	93.565255	6.434745
nov	89.856133	10.143867
oct	56.128134	43.871866
sep	55.087719	44.912281

Which way is better?

Calling on cell phones is 3 times better than calling on Telephone

y	no	yes
contact		
cellular	85.262393	14.737607
telephone	94.768679	5.231321

Education

Call Illiterate more !!!!

y	no	yes
education		
basic.4y	89.750958	10.249042
basic.6y	91.797557	8.202443
basic.9y	92.175352	7.824648
high.school	89.164477	10.835523
illiterate	77.777778	22.222222
professional.course	88.651535	11.348465
university.degree	86.275477	13.724523
unknown	85.499711	14.500289

Target audience: Suggestion to Company

Illiterate, unemployed, Educated ,on cell phones, before campaign at least 2 times and during campaign 10 times maximum can ensure higher conversion.

Don’t spend more than 30-40 seconds with customers if don’t feel it’s going to convert.

Basic Model

1. In Data set ‘y’ is imbalanced.
   10% of values are only ‘Yes’
2. Basic model built without treating
   missing values and by doing hot encoding.
3. We used Random forest, Logistic,Hard-soft voting, bagging.
4. Among basic ‘Soft voting’ gave us highest TN and low FP.

	Soft
accuracy	0.895017966
F1 score	f1_score 0.528153644697
Precision	0.532102023
Recall	0.524263432
AUC	0.73303842
Confusion matrix	[[8611 532]
	[549 605]]

Using Under and OverSampling

1. As ‘y’ is imbalanced so we went for Under and oversampling.

2. Oversampling had better output in both trials.

	Oversampling
accuracy	0.88
F1 score	f1_score 0.643456790123
Precision	0.489849624
Recall	0.937410072
AUC	0.906837616
Confusion matrix	[[9610 1357]
	[87 1303]]

Treating missing values

1. Missing values replaced by simple mode values for categorical variables.

2. Applied same previous approach for this new database.

3. This model is better at capturing people who have higher tendencies to get conversion

accuracy	0.88290038
F1 score	f1_score 0.648713060057
Precision	0.490273775
Recall	0.958450704
AUC	0.914536681
Confusion matrix	[[9522 1415]
	[59 1361]]

Based on best features

1. 'age'
2. 'duration'
3. 'campaign'
4. 'pdays'
5. 'cons.conf.idx'
6. 'euribor3m'
7. 'nr.employed'

This is close to using all parameters but little less in accuracy and precision

accuracy	0.878854091
F1 score	f1_score 0.638841978287
Precision	0.480058013
Recall	0.954578226
AUC	0.91192904
Confusion matrix	[[9536 1434]
	[63 1324]]

Imputation using same and other data set

1. Using other dataset ‘bank full’ we replaced missing values in ‘bank-additional’.

2. We used random forest for getting missing values.

3. Education imputed using same data set and default imputed using other data set.

4. Hyper parameter tuning and after that did soft voting, which is having higher precision,recall.

Assumption: These both datasets consists of same set of customers.

accuracy	0.8977414846
F1 score	f1_score 0.644137224782386
Precision	0.5310257493
Recall	0.8184775537
AUC	0.8631621345
Confusion matrix	[[10945 1111]
	[279 1258]]

Practical Model

1.Dropped all time related data.
2. we want to apply model even before calling.

accuracy	accuracy score 0.878025454277937 2
F1 score	f1_score 0.414134275618374 57
Precision	0.4532095901
Recall	0.3812621991
AUC	0.6613095999
Confusion matrix	[[11349 707]
	[951 586]]

Best Model

1. Imputed using random forest,

soft voting, oversampling

2. Second result is for separate

'Test file'

	Training set	Testing set
accuracy	0.8977414846	accuracy score 0.972566156834183
F1 score	f1_score 0.644137224782386	f1_score 0.8745837957824639
Precision	0.5310257493	0.8755555556
Recall	0.8184775537	0.8736141907
AUC	0.8631621345	0.9291735076
Confusion matrix	[[10945 1111]	[[3612 56]
	[279 1258]]	[57 394]]

Github link

https://github.com/Nikita1993/GreyAtom-Hackathon