Machine Learning on Retail Bank Marketing Data

Connie Zhang

Introduction

Retail banking is the provision of services by a bank to individual consumers. Services offered include savings and transactional accounts, mortgages, personal loans and credit/debit cards.

One of big challenges of the business is how to identify consumer and to target those with suitable needs to sign up its services.

In this study, we applied Classification algorithms to the historical marketing data from a European bank to:

Understand important factors for deposit account sign-up.

Make prediction on the sign-up.

Evaluate multiple algorithms based on prediction sensitivity

The Data

The marketing data contains:

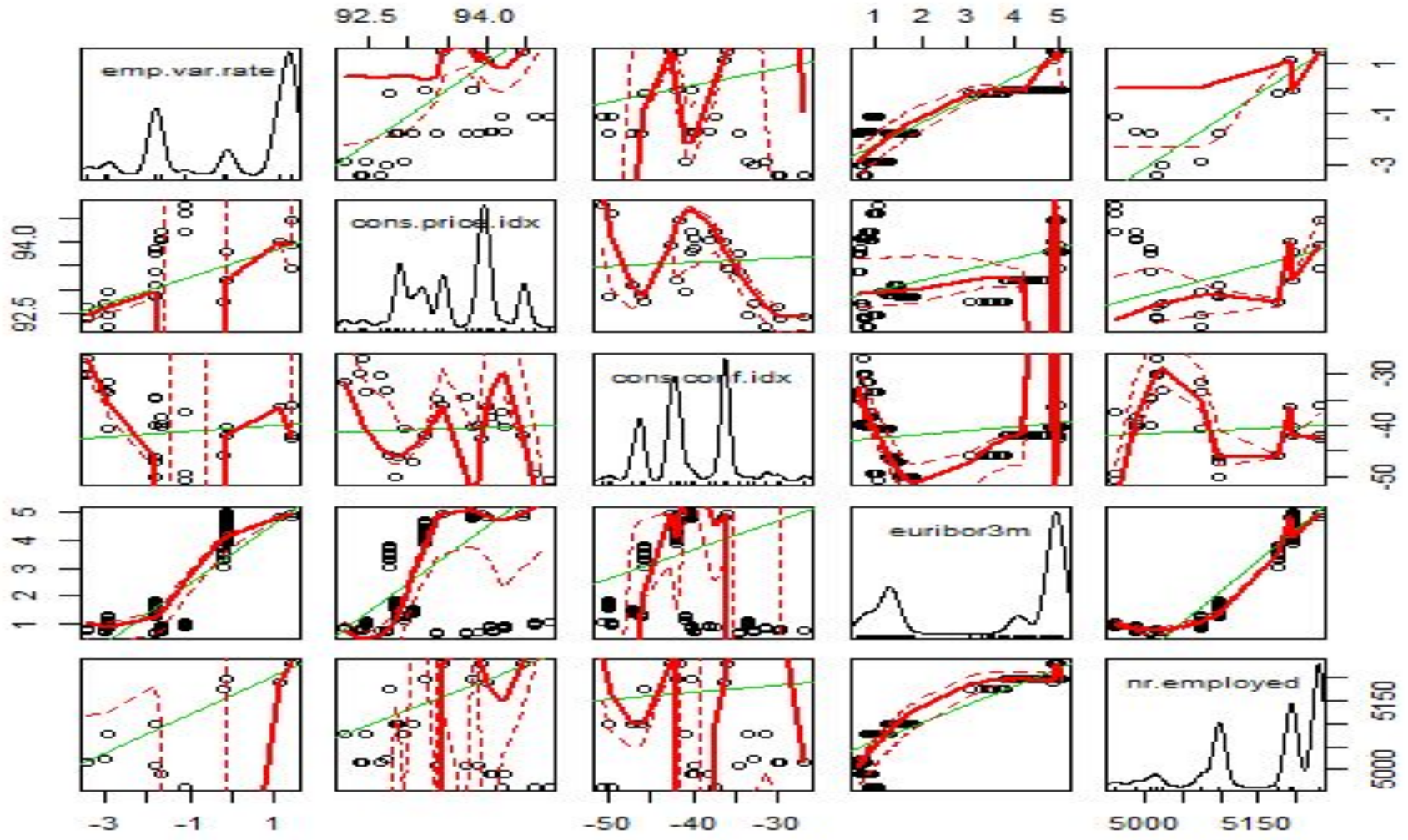
Consumer information: age, sex, marital and job status, etc.

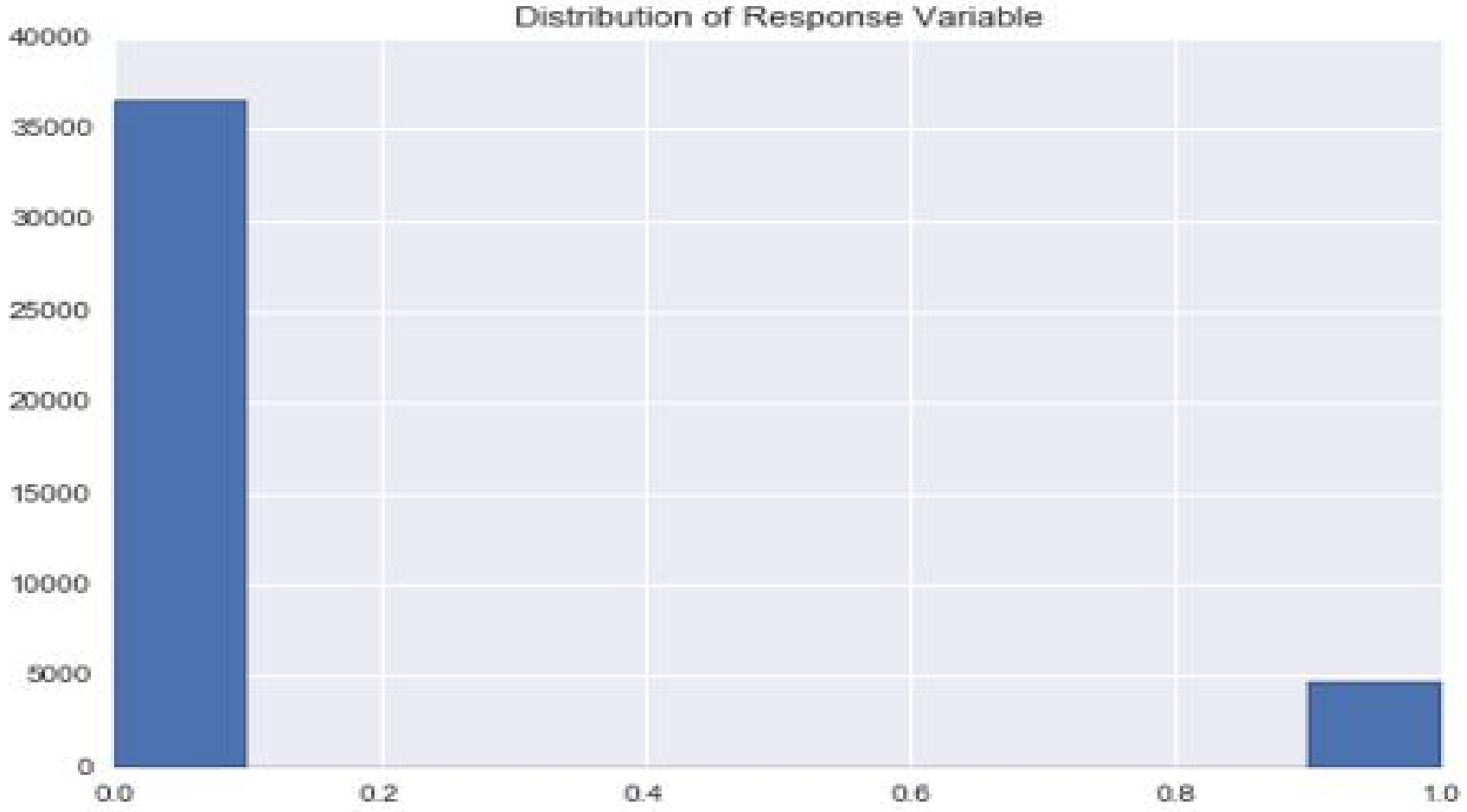
Campaign activity: when and how to contact, times to contact

Social and economic environment data; short-term in nature

Campaign outcome: sign up or not sign up on

Characteristic of the data: highly unbalanced


emp.var.rate
cons.price.idx
cons.conf.idx
euribor3m
nr.employed
92.5
94.0
1
2
3
4
5
-3
-1
1
-50
-40
-30
5000
5150


Distribution of Response Variable
40000
35000
30000
25000
20000
15000
10000
5000
0
0.0
0.2
0.4
0.6
0.8
1.0

Data Pre-processing

Encoding 12 categorical variables and response variable into numerical levels

Imputing missing data on column pdays through four approaches:

1. Leave it as it is (999)
2. Imputing missing data to the mean of the column
3. Imputing missing data as 0
4. Remove the variable from data set

Logistic Regression is used to evaluate the approaches above.

The outcome shows best result to the first approach.

Implementation of Classification Algorithms

The goal:
Find the most suitable classification algorithm based on sensitivity and score

The approach:
Rebalance data: Oversampling/Undersampling on training dataset

Multiple Algorithms: Logistic Regression, RandomForest, Gradient Boosting,
Support Vector Classifier

Model Evaluations: gridsearchCV to generate best parameters
Validation Curve used to evaluate overfitting
Classification Report to evaluate sensitivity

The Results

1. **Logistic Regression**:

On train set:

	precision	recall	f1-score	support
class 0	0.91	0.99	0.95	29235
class 1	0.68	0.21	0.32	3715
avg / total	0.88	0.90	0.88	32950

On test set:

	precision	recall	f1-score	support
class 0	0.91	0.99	0.95	7313
class 1	0.68	0.23	0.34	925
avg / total	0.88	0.90	0.88	8238

2. Random Forest Model

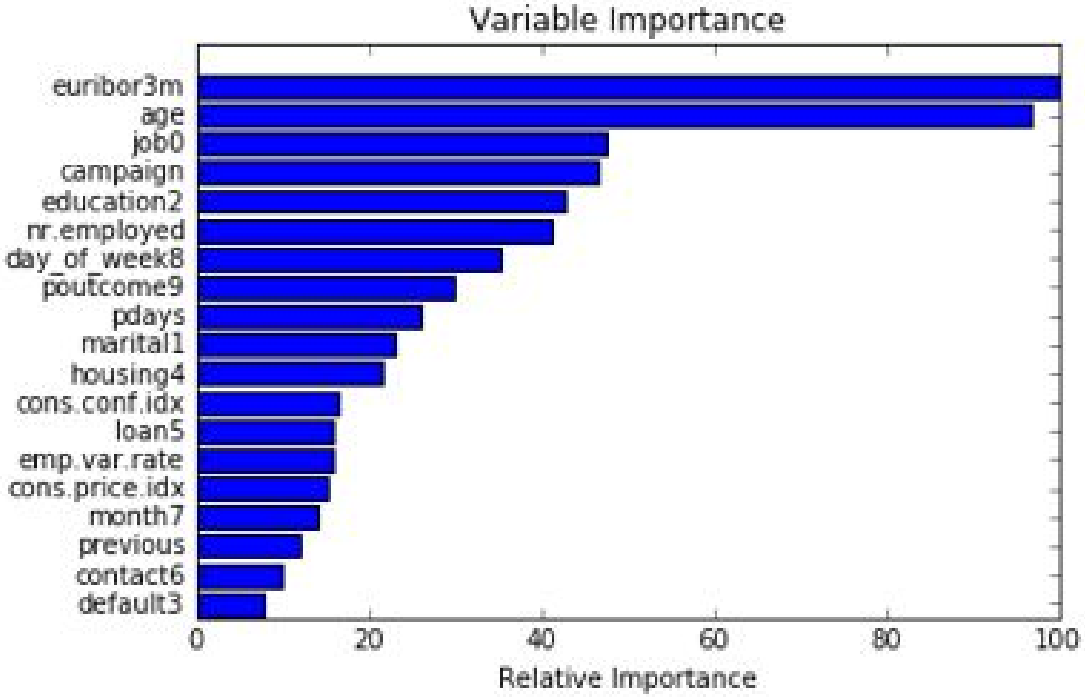
The parameter setting to give the highest sensitivity:

n_estimators=1000, max_depth=4,max_features='log2',class_weight={1:6}

Train set :	precision	recall	f1-score	support
class 0	0.95	0.87	0.91	29224
class 1	0.37	0.62	0.47	3726
avg / total	0.88	0.84	0.86	32950

Test set:	precision	recall	f1-score	support
class 0	0.95	0.87	0.91	7324
class 1	0.37	0.61	0.46	914
avg / total	0.88	0.84	0.86	8238

Relative Importance of Features:



3. Gradient Boosting model

The parameter setting: max_depth = 4, n_estimators= 60,learning_rate= 0.1,
Sample_weight is 1:7 in which 1 corresponds to class 0

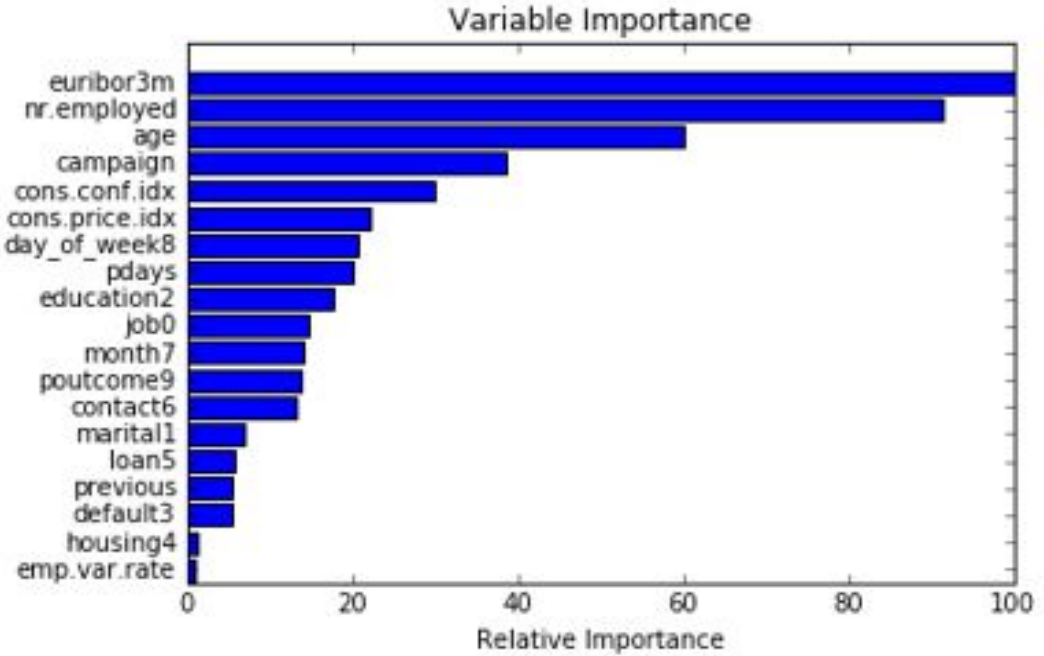
Train set:

	precision	recall	f1-score	support
class 0	0.95	0.88	0.91	29224
class 1	0.40	0.63	0.49	3726
avg / total	0.89	0.85	0.87	32950

Test set:

	precision	recall	f1-score	support
class 0	0.95	0.88	0.91	7324
class 1	0.39	0.63	0.48	914
avg / total	0.89	0.85	0.86	8238

The important features are:



4 Support Vector Classifier

Parameter setting: Kernel = “rbf’, class_weight= 1:8

Train set:

	precision	recall	f1-score	support
class 0	0.93	1.00	0.96	29224
class 1	0.94	0.37	0.53	3726
avg / total	0.93	0.93	0.91	32950

Test set:

	precision	recall	f1-score	support
class 0	0.91	0.98	0.94	7324
class 1	0.57	0.18	0.27	914
avg / total	0.87	0.89	0.87	8238

Conclusion

1. Multiple Algorithms have been evaluated
2. Oversampling/undersampling for unbalanced data
3. Cross Validation used for parameter selection
4. Gradient Boosting is the best model for the sensitivity