Prediction of the success of bank telemarketing and critical factors for it

Kunmei Li
li.kun@husky.neu.edu
College of Engineering, Northeastern University

1. Background

Marketing selling campaigns constitute a typical strategy to improve the business. Marketing campaigns of bank are based on the phone-calls, the objection of this research is to use the data mining method to predict the success of the bank telemarketing for selling the long-term deposit. The most successful telemarketing is to focus on the quality of prospect data, attempting to predict the expected customers that have a higher probability to use the service by using data mining technique. Based on the data mining result, the bank would classify the clients into potential clients and non-potential clients before they contact them. (Vajiramedhin, 2014)

If we can correctly predict the probability of a client to subscribe a term deposit based on their background (such as job, education and so on), we could minimize the cost by taking the method like decreasing the number of calls for the client or rearrange the number of employees. It is easy to see that it's a classification problem.

In this paper, this problem could be solved by two main objectives, one is for prediction, another is for explanation. The prediction is more focus on selecting the best prediction model based on the sensitivity and accuracy, while the explanation is more focus on the initial relationship among attributes and responses.

2. Data and method description

The dataset is a public resource, which comes from UCI Machine Learning Repository about the bank telemarketing of a Portuguese retail bank collected from 2008 to 2013. There are 45211 observations and 21 variables in the dataset, 11 of them are categorical variables while others are continuous variables. Related research has been done before for the same dataset, for example, there is a paper describes an implementation of a DM project based on the CRISP-DM methodology (Sérgio Moro, 2011) and a paper introduces analysis and applications of the most important techniques in data mining; multilayer perception neural network (MLPNN), tree augmented Naïve Bayes (TAN) known as Bayesian networks, Nominal regression or logistic regression (LR), and Ross Quinlan new decision tree model (C5.0) (Elsalamony, 2014). In another paper, the Logistic Regression, Decision Trees and Neural Network are compared. And the result show that the Neural Network gives the best performance. (Sérgio Moro, 2014)
Depends on the variables' types, few methods could be used to do the analysis. As there are two objectives for this analysis, prediction and explanation, different methods would be used to

deal with the variables. The true meaning of the variables for prediction is not as important as that for explanation. Therefore, for prediction, dimension reduction methods like PCA could be used to reduce the complex of the dataset before further analysis. However, for explanation, we could not ignore the initial meaning of each variables, which means it's better to put all 21 variables into analysis in their original forms.

2.1. Explanation

There are total 21 variables, and each of them has its own meaning, choosing the suitable one is important. Understanding the relationship among them is quite important. at first, the correlation results of all the numeric variables shows that:

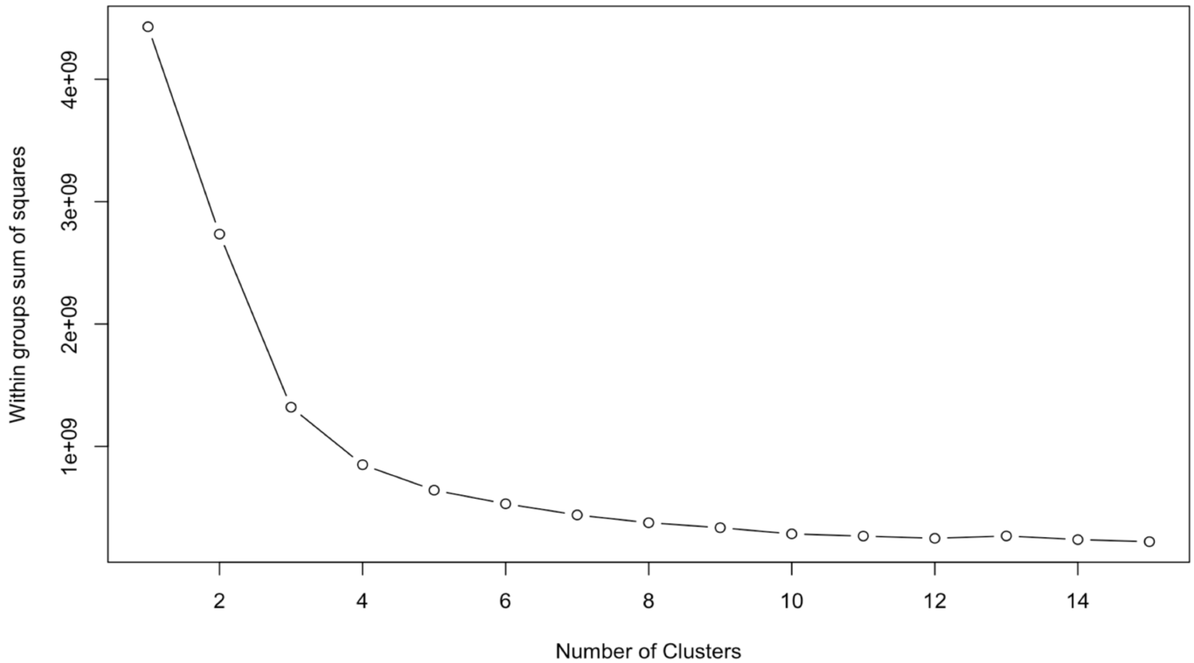
- There exists high correlation between number of employees and euribor 3-month rate.
- There exists high correlation between employment variation rate and consumer price index/ euribor 3-month rate/number of employees

Except that, clustering and association rules are used to get more information among variables. However, when we use clustering, it require all the variables are numeric which is opposite to association rules as association rules require categorical variables. Therefore, before we doing the analysis by these two method, transformation of variables is needed. However, as if we use the associate rules, it's hard to decide the breaks for the continuous variables. Therefore, let's only use associate rule for the original categorical variables.

2.1.1. Clustering

Before doing clustering, as there are 11 categorical variables, we need to transfer the multi-level categorical variables into numerical variables by adding dummy matrix.

The clustering analysis are done by R, first the graph shows how groups sum of squares changes with the number of number of groups are shown below:



From this graph, we could see that when number of clusters=10, it's a clear embow of the line when number of clusters are less than 15. Let's choose 10 as the number of clusters to do clustering. As the response only has two levels, so the clustering could not give better explanation for the problem.

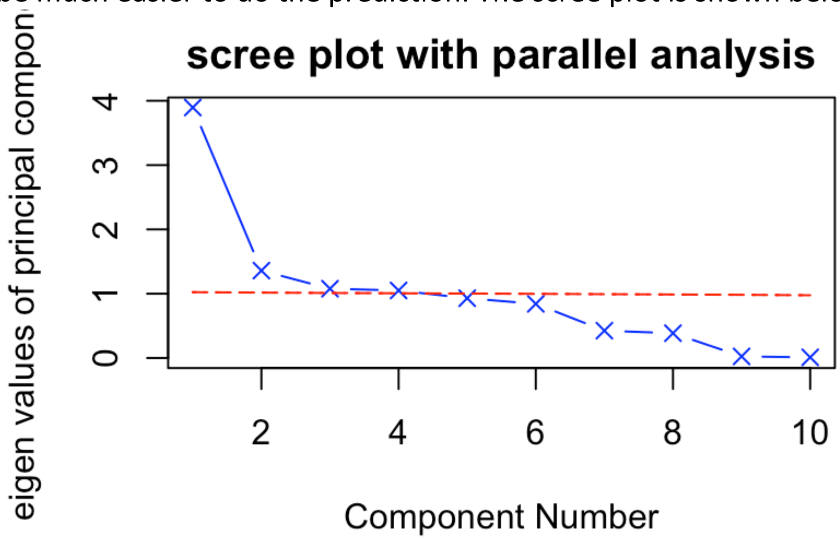
2.1.2. Association rules

For association rules, I set the support as the 0.4 and confidence=0.9, then there are total five rules which is shown below. And it is easy to see that the client with the previous campaign outcome is "nonexistent", the client would almost all refuse to subscribe a term deposit. This is the rule only comes from the categorical variables.

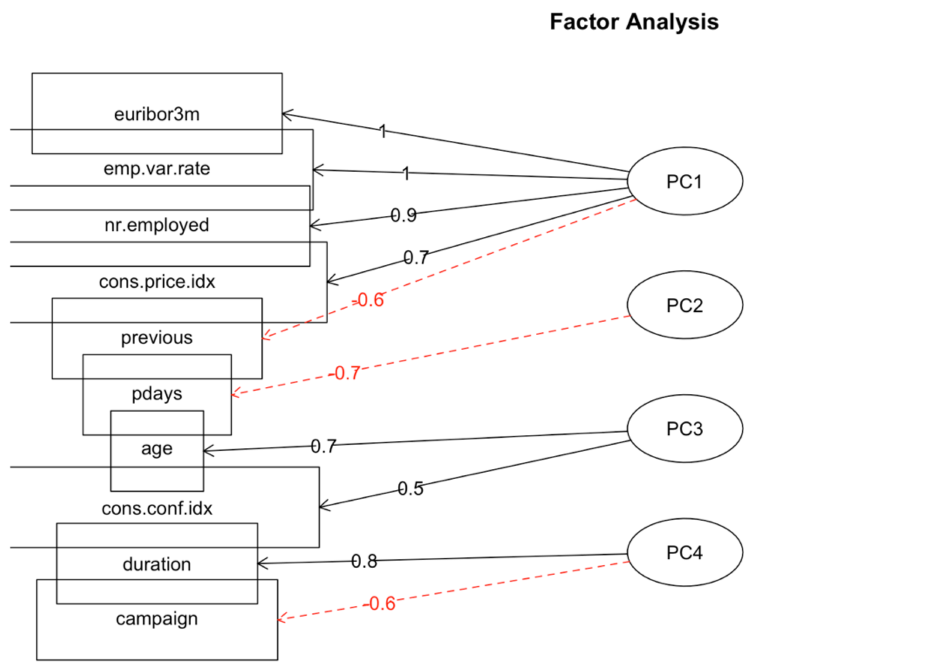
```
    lhs                          rhs
[1] {poutcome=nonexistent} => {y=0}
[2] {housing=yes,
     poutcome=nonexistent} => {y=0}
[3] {marital=married,
     poutcome=nonexistent} => {y=0}
[4] {loan=no,
     poutcome=nonexistent} => {y=0}
[5] {marital=married,
     loan=no,
     poutcome=nonexistent} => {y=0}
```

2.2. Prediction

For prediction, as there are lots of variables, if the objective is to predict, then we need to do dimension reduction first which could help us to reduce the calculation, and more important, is to in case of the influence on the results because of correlation among the independent variables. As PCA is quite hard to do with categorical variables, this is just used to decrease the number of the numerical variables. Here PCA (Principal Component Analysis) is used to do this. After drawing the scree plot, it shows that the number of PC should be 4, and we transfer 10 numerical variables into 4 PCs, it would be much easier to do the prediction. The scree plot is shown below:



For the prediction, there are three methods were used, Logistic Regression, Naïve Bayes, and Classification Tree. Before doing the analysis, we need to separate the dataset with the 4 PCs transferred from the numerical variables into two subsets, the training part and the test part with the ratio as 7:1. The details of 4 principal components are shown below. And from the graph showing the relationship between the variables and principal components, it is easy to see that the PC2 is negatively related to pdays, this means that the longer time since last contact to the customer, the smaller the PC2; PC4 is positively related to duration while negatively related to campaign, this means the longer the last contact duration and the less the contact this time, the larger the PC4; and PC3 are positively related to the age and cons.conf.idx, this means that the older the bank client and the higher the consumer confidence index, the larger the PC3; PC1 positively related to euribor3m, emp.var.rate, nr.employed, cons.price.index and negatively reated to previous, this means that the more the employees, the higher the employment variation, the higher the euibor 3 month rate and the higher the consuer price index and less experience of the client, the larger the PC1. From all those things we cold easily to see that PC1 stands for the condition of employees, PC2 stands for the time interval between two campaign, and PC3 stands for the condition of the client, PC4 stands for the contact information for the campaign. Each of the principal components has their own meaning.



	PC1	PC2	PC3	PC4	h2	u2	com
age	0.00	0.29	0.66	-0.26	0.59	0.411	1.7
duration	-0.05	0.09	0.04	0.79	0.63	0.367	1.0
campaign	0.20	-0.01	-0.34	-0.59	0.50	0.499	1.9
pdays	0.45	-0.73	0.26	-0.01	0.81	0.192	2.0
previous	-0.60	0.55	-0.29	-0.02	0.76	0.244	2.4
emp.var.rate	0.96	0.19	-0.09	0.05	0.97	0.025	1.1
cons.price.idx	0.72	0.33	-0.29	0.08	0.72	0.284	1.8
cons.conf.idx	0.20	0.50	0.53	-0.07	0.58	0.425	2.3
euribor3m	0.97	0.17	0.00	0.04	0.97	0.032	1.1
nr.employed	0.93	-0.02	-0.03	0.03	0.86	0.137	1.0

2.2.1. Definition of data mining methods

As the 10 variables are categorical variables with more two levels, so it is hard to use K-nearest neighbors as if transfer all the categorical variables into dummy variables, then the attributes are too much which would leads difficult calculation. And the neural Network is more like a black-box , so only Logistic regression, Naive Bayes, classification trees are selected to do the prediction. Here is some introduction for each method.

a. Logistic Regression

Logistic regression, also called logit model, is a regression model for the data with the binary categorical response. Logistic regression was developed by statistician David Cox, it is used to estimate the probability of a binary response based on one or more predictor variables. It allows one to say that the presence of a risk factor increases the probability of a given outcome by a specific percentage.

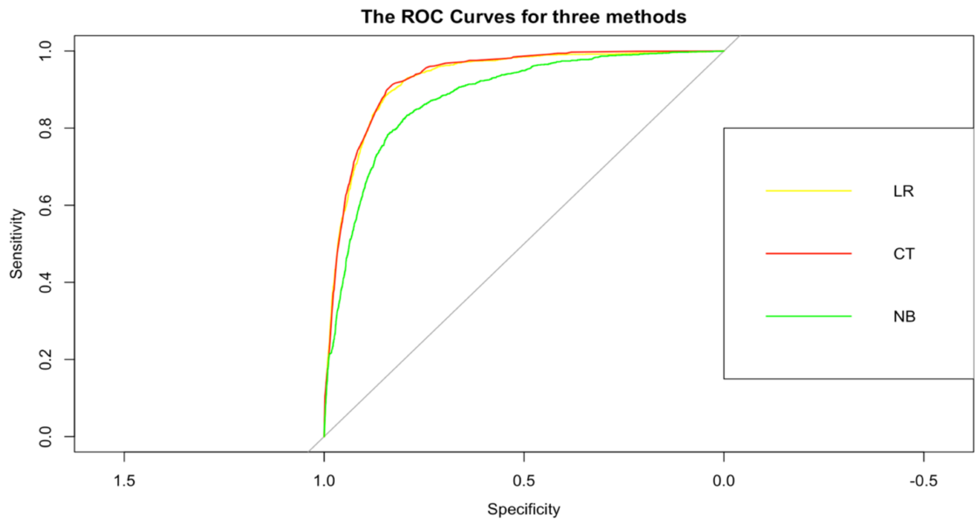
b. Naïve Bayes
Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. Naïve Bayes classifier is a simple classifier which is based on the Bayes' theorem with strong independence assumptions between variables. Naive Bayes classifiers are highly scalable, requiring a number of parameters which are linear in the number of variables (features/predictors) in a learning problem.

c. Classification Tree
A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. The decision tree learning is a method using decision tree as a predictive model which maps observations. It is one of the most popular predictive method for machine learning, and It is one of the predictive modelling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a finite set of values are called classification trees.

2.2.2. Comparison of the performance of three methods
For binary response, there always should exists one is much more important or meaningful than another one, we call this class as the class of the interest. Therefore, when we compare the performance of different methods. it could not only depend on the accuracy, but also depend on the sensitivity. From the results, we could know that the prediction results of three methods are shown in the table below:

Prediction results		True class					
		Logistic Regression		Naïve Bayes		Classification Tree	
		0	1	0	1	0	1
Predicted class	0	10734	264	10483	515	10499	499
	1	823	536	813	546	602	757
accuracy		0.9120		0.8900		0.9111	
sensitivity		0.3300		0.4854		0.2324	

From the table, we could see that they have their own advantages, like naïve Bayes gives the higher sensitivity but lower accuracy. And the logistic regression gives high accuracy with low sensitivity. And ROC curves are shown below:



From the ROC curve, it is easily to see that CT and LR give better performance than NB for this dataset. So we would like to choose LR as the best prediction method from the results shown in the above figure. From the model, it shows that only jobblue-collar, jobretired, defaultunknown, months except August and September, Poutcome, four principal components are significant for the model. As we all know that the larger the probability, the larger the log odds. From the results, it shows jobblue-collar and defaultunknown, monthmay, monthnov and PC1 lead the decrease of the probability to be yes, and jobretired, other months and Poutcome, PC2, PC3, PC4 lead increase of the probability to be yes.

Let's combine this with the analysis result of the principal analysis part. We find that:

- The blue-collar people are less probable to say "yes" to subscribe while the retired client are more probable to say "yes" to subscribe
- The clients last contacted month is in May or Nov are less probable to say "yes" to subscribe compared to other months
- The client with unknown default information are less probable to say "yes" to subscribe
- The clients who is success for campaign are more probable to say "yes" to subscribe
- The better quality of the employee, the client is less probable to say "yes" to subscribe
- The less try this time, the client is more probable to say "yes" to subscribe
- The less time interval since last campaign, the client is more probable to say "yes" to subscribe
- The better financial condition the client has, the client is more probable to say "yes" to subscribe

Discussion

From this paper, it is easy to find that for the prediction, the best model for prediction is Logistic regression based on the ROC graph. There are some relationship exists between the independent variables and dependent variables. For example, the blue-collar people are less probable to subscribe than the retired people. A people with better final condition, has already subscribe are more probable to subscribe. The employee with lots of experience, try to contact less times and try not to contact in May and Nov would more probable to get success.

All the information could give the strategies for the bank when do the telemarketing campaign. The bank should try not do the telemarketing campaign in May and Nov, try to contact the people with better financial condition or who has subscribed before, and also control the times contacting the clients and decrease time interval between the campaigns. These methods would increase the probability to success of campaign, and avoid unnecessary cost.

Reference

Elsalamony, H. A. (2014). Bank Direct Marketing Analysis of Data Mining Techniques. *Hany A. Elsalamony, 85*(7).

Sérgio Moro, P. C. (2011). Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. *European Simulation and Modelling Conference.*

Sérgio Moro, P. C. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 22-31.

Vajiramedhin, C. (2014). Feature Selection with Data Balancing for Prediction of Bank Telemarketing. *Applied Mathematical Sciences, 8*(114), 5667-5672.