Bank Marketing Analysis Project

--Insights into attributes significant in clients' subscribing a term deposit

Zhihao Guo, Sherry Nie, Jiahui Ji, Yanmeng Song

Data Description

The data set is from UCI Machine Learning Repository. (link:http://archive.ics.uci.edu/ml/datasets/Bank+Marketing) It is a collection of client information from a Portuguese banking institution. 41188 unique clients have been recorded. Each client observation has 20 relevant variables recorded, and the response is a binary variable indicating whether the client subscribed a term deposit or not. I can divide these 20 variables into 4 categories: bank client attributes, last contact information attributes, other attributes, and social and economic context attributes. The overview is as follows.

Variables	Description	Type
# bank client data		
1 - age	Range: [17, 98]	numeric
2 - job	Type of job (11 levels + unknown)	categorical
3 - marital	Marital status (3 levels + unknown)	categorical
4 - education	Education level (7 levels + unknown)	categorical
5 - default	Credit in default (Yes/No + unknown)	categorical
6 - housing	Housing loan (Yes/No + unknown)	categorical
7 - loan	Personal loan (Yes/No + unknown)	categorical
# related with the last contact of the current campaign		
8 - contact	Contact communication type—cellular/telephone	categorical
9 - month	Last contact month of year (10 levels)	categorical
10 - day_of_week	Last contact day of the week (5 levels)	categorical
11 - duration	Last contact duration, in seconds	numeric
# other attributes		
12 - campaign	Number of contacts performed during this campaign for this client	numeric
13 - pdays	Number of days passed by after the client was last contacted from a previous campaign	numeric
14 - previous	Number of contacts performed before this campaign for this client	numeric
15 - poutcome	Outcome of the previous marketing campaign—success/failure/nonexistent	categorical
# 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

Overview

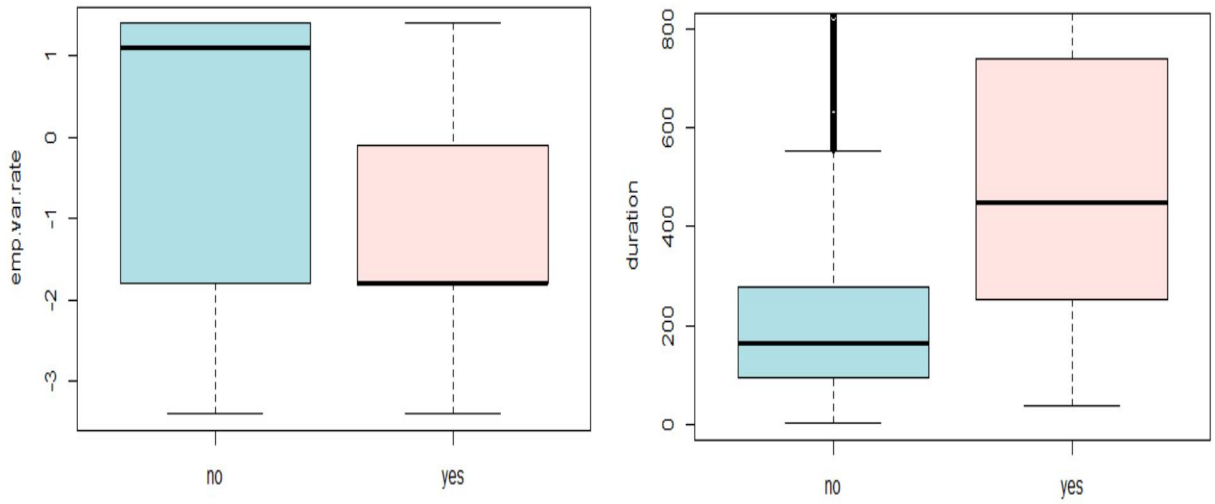
In this project, I have three goals. First, I want to investigate which variables are significant in predicting whether a client will subscribe a bank term deposit or not, which is my variable selection step. Second, based on a smaller set of variables, I would like to apply multiple classification methods to predict the subscription result. By comparing their performance, I can see which model fits best for my data set and thus I can predict the subscription result more accurately. Third, since the first 15 variables are client related, and the last 5 variables are social and economic attributes, I assume the social and economic context will influence personal decision. Thus I would like to compare prediction results before and after adding these five variables. Finally, based on my results, I can provide some useful suggestions for banks on their future marketing campaign.

Data Pre-Processing

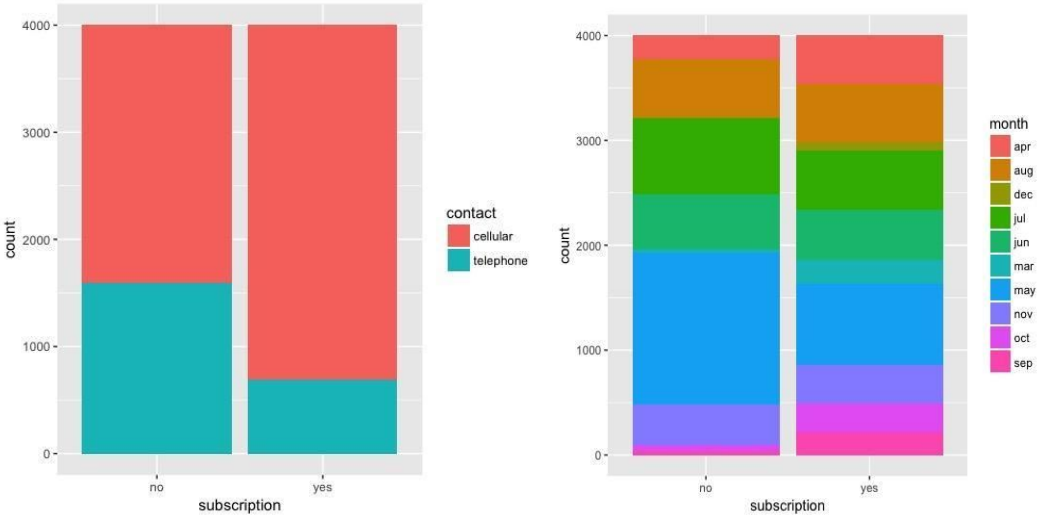
1. There are several missing values in some categorical attributes (job, marital, education, default, housing and loan), all coded with the "unknown" label. I treat these missing values as a possible class label.
2. The variable “duration” represents last contact duration in seconds. If duration = 0, then y = “no”. Since there is no contact with this client, he/she will not subscribe the term deposit through this campaign. Therefore I delete the observations whose duration is 0.
3. The response variable has 4640 “Yes” and 36548 “No”. The number of “No” is almost 8 times of the number of “Yes”. Even if a model randomly predicts all the responses to be “No”, it will still yield a high prediction accuracy, which is not meaningful. In order to deal with the imbalance, I construct the training and test data such that the number of “Yes” and “No” in each set are the same. From all the “Yes” data, I sample 4000 of them as training data, and use the remaining 640 observations for test. And I sample the same number of “No” data for training and test respectively.

Table of Contents

1. Visualizations



For numeric variables such as emp.var.rate and duration, I can see different distributions of the variables in the "no" and "yes" response group.

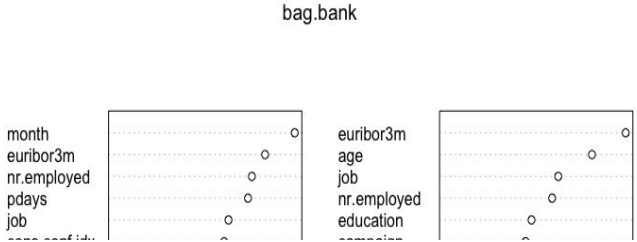


For the plots of categorical variables, I find that some levels are important for subscription result but not all levels. This motivates me to apply variable selection and classification methods to inference on their relationships as Ill as make predictions.

2. Variable Selection

I used four variable selection methods: random forest, lasso, backward selection and forward selection. If a variable is selected by at least three methods among these four, I consider it as important and use it to perform my analysis. The result of variable selection is as follows:

Random Forest:



By “MeanDecreaseAccuracy” criterion, I choose the top ten variables from “months” to “poutcome”. Then, by “MeanDecreaseGini” criterion, I pick the top eight variables from “euribor3m” to “emp.var.rate”.

Lasso:

From Lasso method, I see that most coefficients are shrunk to zero. Lasso method leaves variables: job, education, default, contact, default, day_of_Iek, month, duration, poutcome, emp.var.rate, nr.employed, cons.conf,idx.

Backward and Forward Selection:

Forward and Backward selection give similar result. They select the following variables: job, education, default, month, campaign, day_of_Iek, previous, poutcome, emp.var.rate, cons.price.idx, nr.employed, housing, loan, euribor3m.

Variable Selection Summary:

My final selected variables are: job, duration, education, contact, month, day_of_Iek, campaign, poutcome, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed.

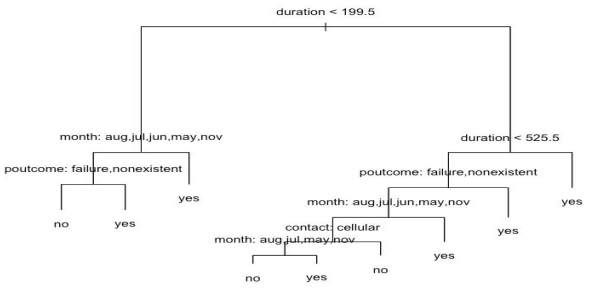
For the banks, clients jobs are important since people with Ill-paid jobs tend to spare more money to deposit. For durations, the more time banks spent to persuade, the more likely the client will subscribe. For education, higher education level is associated with higher salary, which may lead to deposit. For month, probably people get paid at certain month. For day_of_Iek, it is possible that at the beginning and the middle of the Iek, people are busy working so they are unwilling to be disturbed, while on Iekends they are able to consider other things such as make some deposit. For campaign, clients are more likely to deposit if they are contacted very often during the last campaign. For poutcome, if last time client subscribed, it is likely that he/she will also do this time since it shows that he/she is a valuable client. With respect to social economical attributes, clearly when the economy is good, people have free money to deposit.

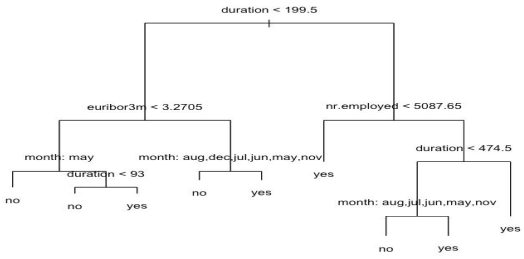
3. Classification

Since my response is the binary categorical variable, I only perform classification among my data set. In this section, I will describe my performance of RandomForest, Logistic Regression, SVM classification, and KNN, and compare their performance on this data set.

Single Tree and Random Forest :

Single Tree Random Forest





From this graph, I can see that single tree has much more branches than random forest. However, random forest performs much better than single tree. Furthermore, I can see from both single tree and random forest, duration is the most important criteria.

Logistic Regression :

I run the Logistic Regression to obtain a formula between response and other variables. Based on the train and test data, I see that the error is smaller than most of the other methods. Thus, I conclude that Logistics Regression is relative a good approach.

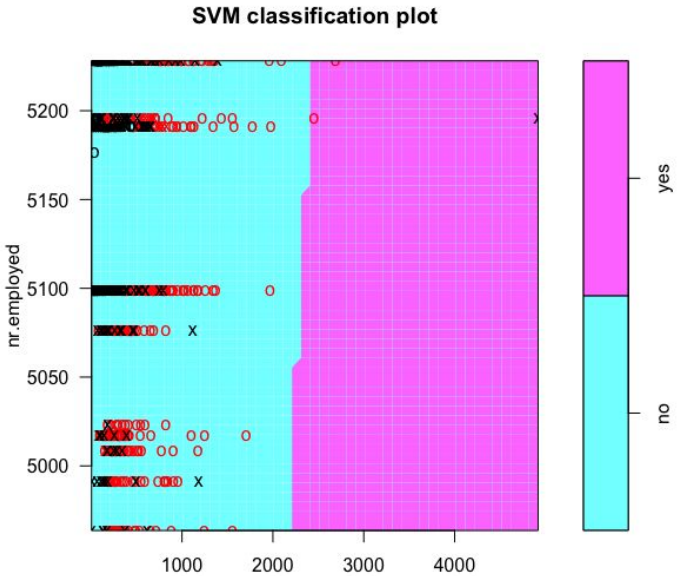
Model 1:

$$log(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)}) = \beta_0 + \beta_1 * job + \beta_2 * duration + \beta_3 * education + \beta_4 * contact + \beta 5 * month + \beta_6 * day\_of\_week + \beta_7 * campaign + \beta_8 * poutcome$$

Model 2:

$$log(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)}) = \beta_0 + \beta_1 * job + \beta_2 * duration + \beta_3 * education + \beta_4 * contact + \beta 5 * month + \beta_6 * day\_of\_week + \beta_7 * campaign + \beta_8 * poutcome + \beta_9 * emp.var.rate + \beta_{10} * cons.price.idx + \beta_{11} * cons.conf.idx + \beta_{12} * euribor3m + \beta_{13} * nr.employed$$

Where Y=1 indicates the response is “yes”, x is the vector of predictors.



SVM :

I compared linear, polynomial, and radial kernels for svm model. I first scale the data to avoid domination of some large-value variables. Since the training data still include 8000 observations, which increases the difficulty of cross validation. So I sample 800 observations from the training data set to cross-validate the

tuning parameters: cost, degree, and gamma. For each kernel, I also fit two versions, one without the five social and economic variables and the other including the five ones. Then I apply each best model from each kernel to test data set to calculate test classification error. The result indicates that linear performs the best with cross-validated choice of cost 0.1. The best polynomial SVM according to 10-fold CV has degree 1. This produces a classifier that is identical to the linear support vector classifier. By comparison, the radial SVM performs not as Ill as linear. One plot of SVM classification with linear kernel corresponding to nr.employed and duration is shown above.

KNN:

1. Self-Defined Distance

Among the selected variables, duration, campaign and the five social & economic variables are numeric, the rest 6 (job, education, contact, month, day_of_Iek, poutcome) variables are all categorical. In this case, I cannot put the data into the "knn" function in R directly, because the default method "Euclidean" is not applicable. Therefore, I created a way to calculate the distance between two data points if they contain both numeric and categorical variables.

Given two observations, I look at their numeric and categorical variables separately. For each numeric variable, I do the same thing as calculating Euclidean distance: take the difference of them and square it. For each categorical variable, I simply check if they have the same value. If so, the distance is 0; if not, the distance is 1. Then I sum up all the (squared) distances and take the square root. That is the distance obtained between the two observations.

2. Distance Matrix

Since the range of numeric variables differ, it is possible that some numeric distances dominate the final distance. After getting n observations to compute pairwise distance, I first scale the numeric columns to make each variable have mean 0 and standard deviation 1. Next I compute an nxn pairwise distance matrix M, where each entry M(i, j) is the distance between the i'th and j'th observation. Therefore the matrix is symmetric and has diagonal entries 0.

3. KNN prediction for training data

Based on the distance matrix obtained, I am able to perform knn prediction by majority vote, as this is a classification problem. I go through each row of the matrix, and pick the least k values (excluding 0), then the corresponding column index are the index of the closest k points to the current observation. I next find the responses for these k points, take the majority vote as my prediction for this observation. After getting all the predicted results, I compare with their true responses and compute an error. So I get a training error for this particular k.

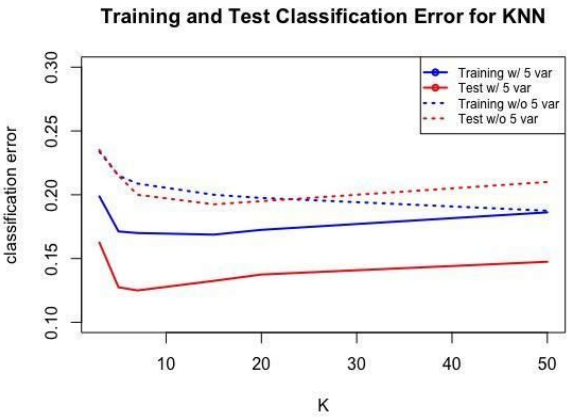
4. KNN prediction for test data

To perform KNN on test data, I need the distance matrix between training and testing data, which is a little different from what was described earlier. Say there are n training points and m test points. Each row of the matrix contains distances from a single test point to each of

the training point. Therefore the matrix is in dimension mxn. Next I still go through each row of the matrix, find the least k values, which correspond to the closest k training points to this test point. Then find their responses and take the majority vote as my prediction. By comparing with the true responses of test data, I can get a test error for this k.

5. Procedures

Since I have 16000 training data and 1280 test data, if I want to compute the distance matrix on the full data, it will be a huge matrix and costs too much computation. For convenience I sampled 800 training data and 400 test data to perform KNN. I chose a range of k (3, 5, 7, 15, 20, 50), run KNN for each of them, and select the k that achieves minimum test error. This is done for both with and without the last 5 social/economic variables.



KNN Classification Summary:

From the plot I can see that the overall training and test errors with the 5 variables are smaller than the ones without 5 variables. The test error with those 5 variables are also smaller than the training error, which is a sign of a good model. The optimal k chosen when including the 5 variables is 7, and the optimal k when excluding the 5 variables is 15.

Overall Classification Comparison:

Train and Test Error without five social & economic variables			Train and Test Error with five social & economic variables	
Method	**Training error**	**Test error**	**Training error**	**Test error**
Single Tree	0.1833	0.1578	0.1314	0.1164
Random Forest	0.1564	0.1422	0.1156	0.0992
Logistic Regression	0.1645	0.1508	0.1341	0.1211
SVM (kernel = linear)	0.1712 (cost = 0.1)	0.1563 (cost = 0.1)	0.1375 (cost = 0.1)	0.1273 (cost = 0.1)
SVM (kernel = polynomial)	0.17125 (cost = 10, degree = 1)	0.1578 (cost = 10, degree = 1)	0.1338 (cost = 1, degree = 1)	0.1320 (cost = 1, degree = 1)
SVM (kernel = radial)	0.2063 (cost = 1, gamma = 0.5)	0.1710 (cost = 1, gamma = 0.5)	0.1738 (cost = 1, gamma = 0.5)	0.1555 (cost = 1, gamma = 0.5)
KNN	0.2000(K=15)	0.1925(K=15)	0.1700(K=7)	0.1250(K=7)

From this table, I can conclude that random forest performs the best, since it has the lowest test error. This makes sense since random forest is considered as one of the most competitive classifiers. My data has both categorical and quantitative variables, and I set number of trees as 1000 when run random forest. Meanwhile, I can clearly see for each method, the model with five social & economic variables performs better than the one without those five variables. Therefore, I can conclude that five social & economic variables improve the accuracy in prediction. Overall, the test error is low, even the worst one only has 0.1925 test error, this means my models perform pretty Ill.

4. Clustering

I use 5 social and economic attributes to do clustering to verify my hypothesis that those five variables are significant in predicting whether or not a client subscribes the deposit. Firstly, I perform PCA on those five variables. PC1 and PC2 covers almost 80% of this data. Then, I perform K-means on the 2 PCs. After that, I compare the actual cluster (no as 0 and yes as 1) with the cluster I get from K-means clustering to get the accuracy. Since the majority of observations in cluster 1 has response y=0, I assume that all observations in this cluster has response 0 as they are in the same cluster. Then, I compare the number of observations in cluster 1 with the actual number of observations in y=0, and I use 1-number of difference to determine the cluster agreement for training data.

The table below is the comparison between the true cluster (no as cluster 1, yes and cluster 2) and the clusters I get by K-means using the training data (8000 observations in total). From this table, I can calculate the cluster agreement for cluster=2 is 0.711.

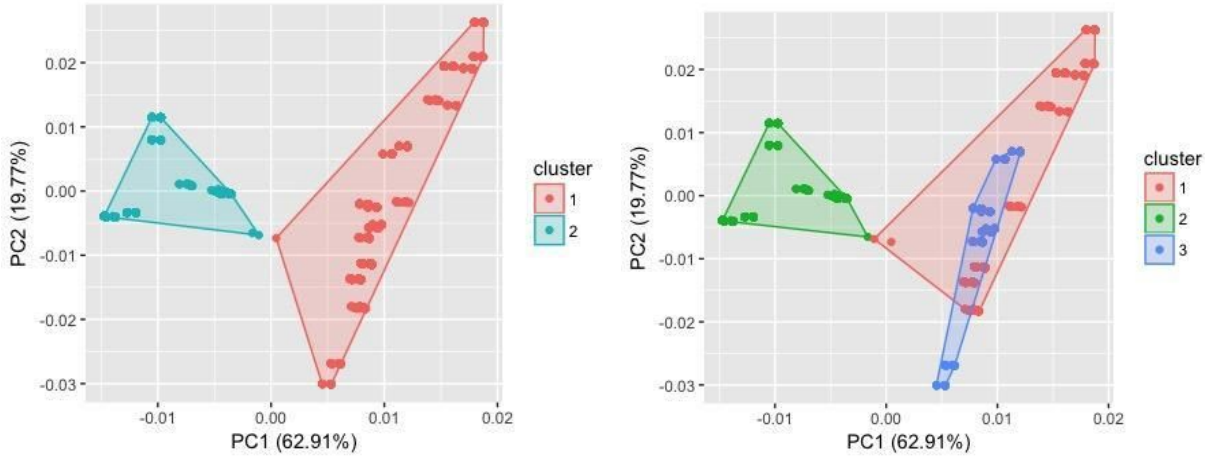
	true cluster 1	true cluster 2
predict cluster 1	2838	1149
predict cluster 2	1162	2851

The table below is the comparison between cluster=2 and cluster=3 using the training data

cluster	cluster agreement for training data
cluster = 2	0.711
cluster = 3	0.336

From this table, I can conclude that cluster=2 performs Ill for this data set, while cluster=3 does not perform Ill. From the graph below, it is easy to see that cluster=2 separates two clusters neatly. However, cluster=3 could not separate those three clusters, which results in low accuracy.

Furthermore, since cluster=2 has relatively high accuracy, I can conclude that my hypothesis on five variables are significant in predicting whether or not subscribe the deposit is correct.



5. Limitations

1.Data limitation: In my data set, there are almost fifty percent variables that Ire categorical. Therefore, it is hard for me to perform KNN and clustering. However, I develop my own method to determine the distance between two categorical variables, and create my own KNN function. Some regression methods still cannot be applied.

2.Interaction: I did not include interaction in this project. Now, my logistic regression already has relatively low train and test error. But in the future, adding interaction terms in regression model can be considered.

3.Level of categorical variables: There is a certain portion of "unknown" level in some categorical variables, which creates difficulty for interpretation. Some categorical variables have many levels and some levels are more distinguishable than others. In further study, I can investigate deeper to these levels and try to figure out its effect in prediction.

6. Summary

Based on my analysis, I have three conclusions:

1. The following variables are significant in predicting whether a client will subscribe a bank term deposit or not: job, duration, education, contact, month, day_of_Iek, campaign, poutcome, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed. Social and economic attributes play an important role and bring improvement to my model performance.
2. Random Forest performs the best among my data set, which has the lowest test error. And this was probably because my data is mixed with categorical and numeric variables.

3. Suggestions for bank on future campaign: firstly, banks should make more contacts to clients, especially those with higher degrees. Secondly, they should prefer cellular call communication type. Lastly, banks should prefer to choosing March, April and October to make contact calls.

Citation:

[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, In press, http://dx.doi.org/10.1016/j.dss.2014.03.001