Bank Marketing Campaign Prediction

1. Introduction

1.1 Abstract

Nowadays, marketing campaign strategies such as phone calls are important ways to increase banks' long-term deposit subscriptions. In order to improve the efficiency of campaigns, it is essential to locate the potential subscribers.

In this project, we will use the "Bank Marketing Campaign" data set. It contains the variables of bank clients' information, social and economic context attributes and some others. The dependent variable is "whether the client will subscribe a term deposit or not". We build classification models such as Logistic Regression and Decision Tree to predict the telemarketing call's results. Finally, we are going to compare all these models according to different criterion like the precision and sensitivity.

With the implementation of this project, banks can better understand what kinds of clients are more likely to subscribe a term deposit, so that they can select a high quality set of potential customers and reduce the phone calls.

1.2 Questions of Interest

1. Which set of variables should we use for building the logistic regression model?
2. Compare different models' performance according to certain criterion(like the accuracy and sensitivity).

2. Data Investigation

2.1 Imbalance and Method of Downsampling

In the data set, the clients subscribing the deposit after phone calls only take up a percentage of 12.7%. So it is a imbalanced data set. In a logistic model, the imbalance of the data not only affects the statistical inference (e.g., hypothesis testing) but also influences the prediction capabilities **(Christian Salas-Eljatib, 2018)**.

In order to deal with this problem, we apply the resampling method. Specifically, we apply the method of downsampling which means we randomly downsample the majority class to equate the number of minority of majority class samples. It is because in this project we are more interested in the minority class(those who are potential subscribers of the deposit). Logics behind this interest will be explained in the model comparison part.

2.2 Data Set Selection

One disadvantage of downsampling is that part of information could be folded up which means it could result in poorer performance for the majority class. So we decide to choose the data set with the largest number of observations("bank-additional-full") to minimize the information loss.

2.3 Exploratory Data Analysis

There are 41188 records of clients and 20 predictor variables in our data set.

For the input variables, there are 0.08%, 4.20%, 2.40%, 2.40%, 20.9% "unknown" values in job, education, default, housing, loan respectively. We will use "unknown" as a class label instead of deleting them in order to help draw more insight and predict the outcome.

- **Bank client data**

Duration: The duration is not known before a call is performed. Thus, this variable should be discarded since our intention is to have a realistic predictive model.

Age: According to the left panel of Figure 2.3, we can see that when age is smaller than 20 or larger than 60, the proportion of subscribing a term deposit is significantly larger than that when age is between 20 and

60. This is reasonable because teenagers and the seniors are more likely to be influenced by marketing than middle-aged.

Housing and loan: According to Table 2.3, the clients whose housing is unknown also have unknown loans. These two variables have high correlation. We will drop loan to avoid multicollinearity.

Job: According to the right panel of Figure 2.3, we can see the proportion of subscribing a term deposit of people who are retirees or students is much higher than that of others. So we concluded that job has a significant association with y.

The percentage plots of other client's categorical variables are in Appendix (Figure 2 and figure 3). All of them seems have significant relationship with y.

- **Other Attributes**

96.3% of pdays are 999 and 86.34% of previous are 0, which means the client was not previously contacted. We will create a binary variable Indicating whether the client has been contacted before to replace pdays and previous.

According to the Figure x in Appendix, there exists collinearity among emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed. In fact, when these five variables were put in the logistic regression model, four of them have vif greater than 7. We also notice that the vif of emp.var.rate greatly decreases with the absence of the other four. So we decide to drop the other four variables.

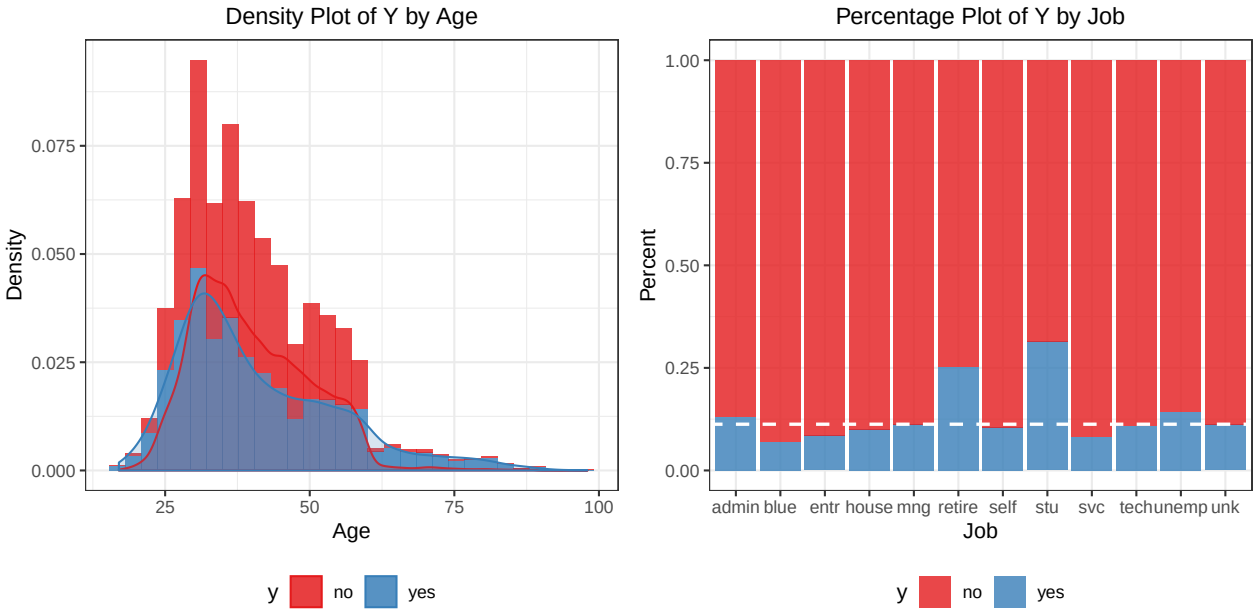

Figure 2.3 Left Panel: Density plot of y by age; Right Panel: Percentage plot of y by job. The white dotted line denotes the overall percentage of 'yes' in y.

	loan		
housing	no	unknown	yes
no	16065	0	2557
unknown	0	990	0
yes	17885	0	3691

Table 2.3 Frequency table of loan and housing

3. Modelling

3.1 Data Preparation

We divide the data set into a training data set and a validation data set which separately takes up 70% and 30% of the observations in the downsampled data set.

3.2 Logistic Model

The response variable is a binary variable. So we use logistic model to fit the data and to make predictions.

$$log(p_i/1-p_i)=\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_k x_{ki},\ i=1,2,...,n$$

$x_1,\cdots,x_k$ are the independent variables we use for predicting the response variable and are selected based on the our results of EDA. The left side of the function is called "log odds". The "odds" measures the ratio of the probability of response variable being 1 over the probability of response variable being 0. With each observation of the predictor variable, we will get an estimation of the probability of response variable being 1.

3.2.1 Model Assumptions

1.Binary logistic regression requires the dependent variable to be binary. Here it is satisfied.

2.The error term needs to be independent.

3.No severe multicollinearity.

4.The linear relation between the independent variables and log odds.

3.2.2 Diagnostic

1.We check the VIF index for multicollinearity. Except those listed above, all predictor variables' vif are smaller than 2 which is quite a safe value. The comparatively higher vif for age and age^2 is reasonable. Usually a vif > 10 will be treated as an obvious sign for severe multicollinearity. In our results, the assumption is not violated.

2.Figure 3.2.2 shows the residuals of logistic regression, there is no obvious violation of error's independence.

3.For linearity assumption, we conduct chi_square test:

H_0: There is linearity between the predictor variables and the log odds.

H_α: No linearity.

The p-value for the test is 0.995 which means the null hypothesis is rejected under 0.01 level. One of the reasons could be that there are too many categorical predictor variables.

	GVIF	Df	VIF
age	49.90	1	7.06
ContPrev	13.26	1	3.64
age2	51.84	1	7.20

Table 3.2.2 Three highest VIFs among the predictor variables

```
## Warning: About 89% of the residuals are inside the error bounds (~95% or higher would be good).
```

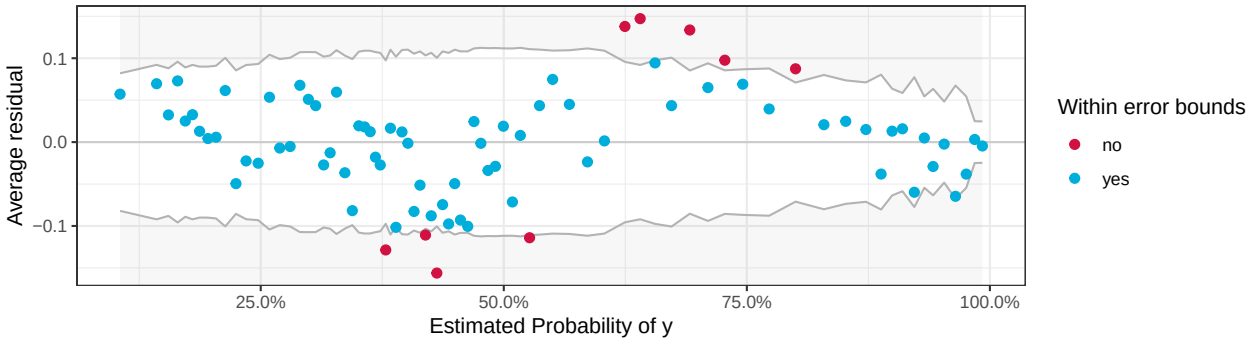

Figure 3.2.2: Residuals' plot of the logistic regression

3.2.3 The fitted result of the logistic model:(Those significant under 0.05 level are shown)

	Estimation	Std	P-value
age	-0.077	0.019	5.54e-05***
age2	0.001	0.0002	5.63e-05***
edu.degree	0.282	0.136	0.038*
contacttelephone	-0.292	0.090	0.001**
monthdec	0.986	0.451	0.029*
monthmar	1.477	0.282	1.68e-07***
monthmay	-0.645	0.114	1.61e-08***
monthnov	-0.507	0.133	0.0001***
monthoct	1.225	0.244	5.21e-07***
campaign	-0.051	0.013	0.0001***
poutcomenonexistent	0.572	0.095	1.84e-09***
emp.var.rate	-0.445	0.027	<2e-16***
ContPrev1	2.176	0.535	4.75e-05***

Table 3.2.3 Fitted result of the logistic model

3.3 Random Forest

Another model we apply here for prediction is Random Forest. The algorithm of Random Forest is based on decision trees. Random Forest consists of a large number of independent models. There are two important assumptions for Random Forest:

1.We need features that have at least some predictive power. It means the community of all the models should be at least better than random classification. It is reasonable to assume that this property holds because we have selected some potential important variables in the EDA process. Also, this property could be checked in the following model comparison part where we have drawn the plot to demonstrate the prediction performance of the models. For example, we could compare the ROC line of Random Forest with the straight line from left bottom to right top(the line for a fair coin flip).

2.The trees of the forest and more importantly their predictions need to be uncorrelated(or at least have low correlation). This property is the base of the "advantage of population". Only if every individual makes their decision independently will the whole population be less likely to make a biased decision. This assumption needs further check but here we believe it holds.

4. Model Comparisons

We need certain criterion to comment on each model's prediction performance. In the confusion table, the "sensitivity" is a better criterion. We try to explain the logic behind: the sensitivity measures how much percentage of actual subscribers will be predicted to be positive. Higher sensitivity means a greater chance to catch the potential subscribers, so sensitivity is more related to banks' profits.

In order to make our conclusion more robust, we repeat the prediction process separately for the Logistic model and the Random Forest model by repeating the process of downsampling. We plan to repeat each prediction model for 100 times and make a summary of them. However, because of the huge amount of calculation for Random Forest, it is not realistic to run it for 100 times. Here we reduce the repeated times to 10.

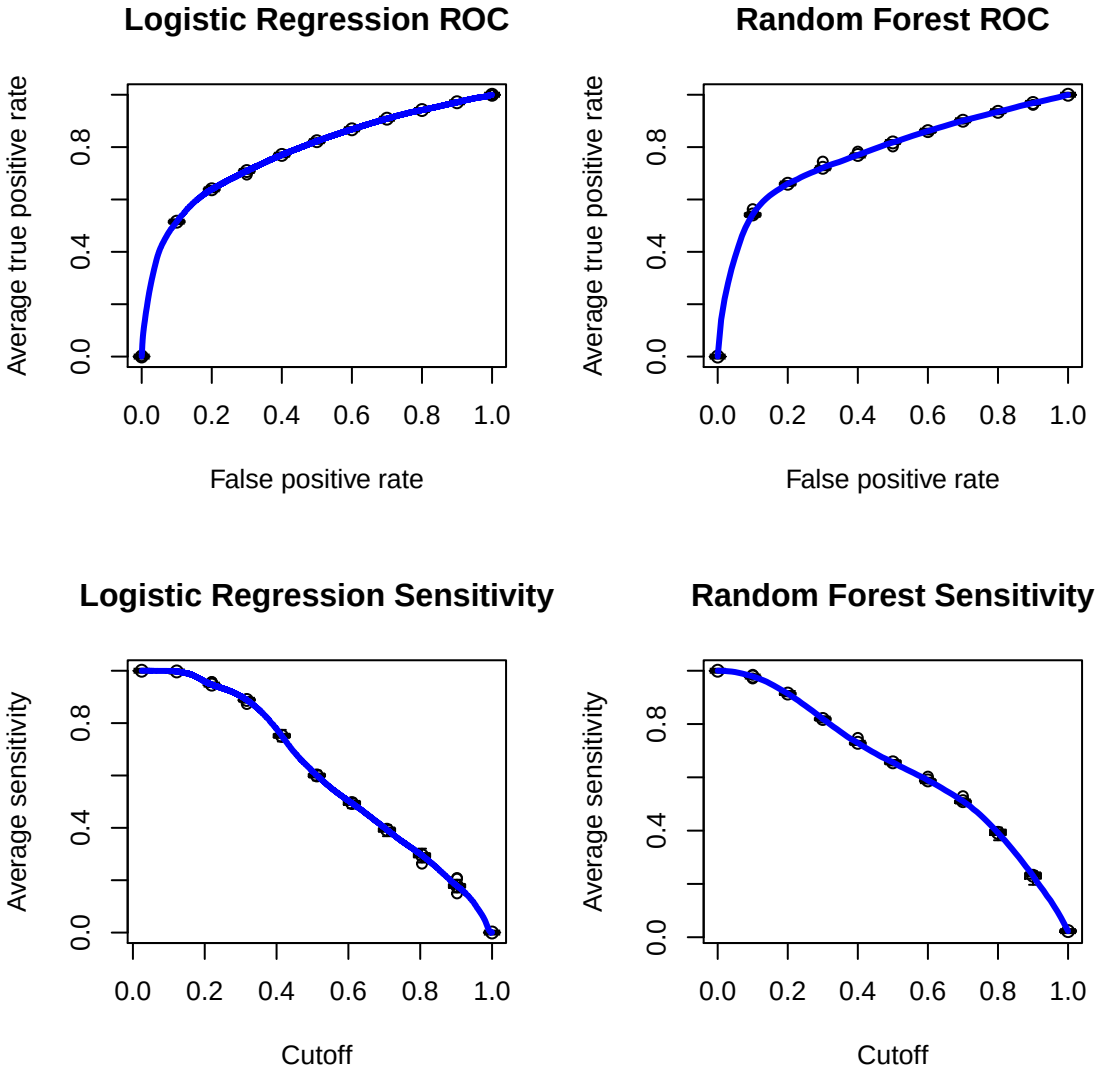

Figure 4. Up-left Panel: sensitivity plot of logistic regression with the change of cutoff; Up-right Panel: sensitivity plot of random forest with the change of cutoff; Down-Left Panel: ROC plot of logistic regression; Down-right Panel: ROC plot of random forest.

The ROC(receiver operating characteristic) curve plots (Up Panel of Figure 4)the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The larger space under the line, the better sensitivity the model has. We can see that the performance of the logistic model is not significantly better than a coin flip(TPR=0.5, FPR=0.5). From this plot we can conclude that the random forest is a better model.

The down panel of Figure 4 shows how the sensitivity changes with the cutoff rate and the confidence intervals for each model. When the cut off rises, the sensitivity will drop because the model becomes stricter on making positive decisions. From the plot we can see that the random forest does better than the logistic model. First, random forest has smaller variance in sensitivity which means the prediction is more robust. Secondly, when we compare the results at cutoff 0.5, we can find that random forest has better sensitivity

which is approximately 0.6. It means the random forest model has a significant ability to identify the actual subscribers.

To confirm the existence of the gap statistically, we collect the mean sensitivity at different cutoff levels and conduct a t test to test the difference between them.
H_0:The true mean sensitivity of each model are the same.
H_α: The true mean sensitivity of each model are not equal.

	t.value	df	p.value
value	15.53	9.204	6.497e-08

Table 4 Result of two models' mean sensitivity

The extremely small p-value confirms our conclusion before. There is a gap between two models' prediction performance under 0.01 significance level.

5. Explanation for the Gap
In our analysis, the random forest model performs better in the sensitivity than the logistic model. Possible reasons are as follows:
1. The linearity between the logit and the predictor variables is not significant. In the model diagnostic part, we conduct the chi_square test to test the linearity and the result indicates this assumption is violated.
2. More noise variables than predictor variables. In our logistic model, there are more than 15 variables. It is highly possible that a large portion of them have little effect in predicting.
3. Too many categorical variables in our data set.

Reference

[1] Christian Salas Eljatib, Andres Fuentes-Ramirez, Timothy G. Gregoire, Adison Altamirano, Valeska Yaitul. Study on the effects of unbalanced data when fitting logistic regression models in ecology. Ecological Indicators, 85, 502-508.

[2] Georges Dupret, Masato Koda. Bootstrap re-sampling for unbalanced data in supervised learning. European Journal of Operational Research, 134(2001), 141-156.

[3] Yang Liu, Nitesh V. Chawla, Mary P. Harper, Elizabeth Shriberg, Andreas Stolcke. A study in machine learning from imbalanced data for sentence boundary detection in speech. Computer Speech and Language, 20(2006), 468–494.

[4] Mr.Rushi Longadge, Ms. Snehlata S. Dongre, Dr. Latesh Malik. Class Imbalance Problem in Data Mining: Review. International Journal of Computer Science and Network (IJCSN), Volume 2, Issue 1, February 2013.

[5] Tobias Sing, Oliver Sander, Niko Beerenwinkel and Thomas Lengauer. ROCR: visualizing classifier performance in R. BIOINFORMATICS APPLICATIONS NOTE, Vol. 21 no. 20 2005, 3940–3941.

[6] Anne Ruiz-Gazen, Nathalie Villa. STORMS PREDICTION: Logistic regression vs random forest for unbalanced data.

Appendix

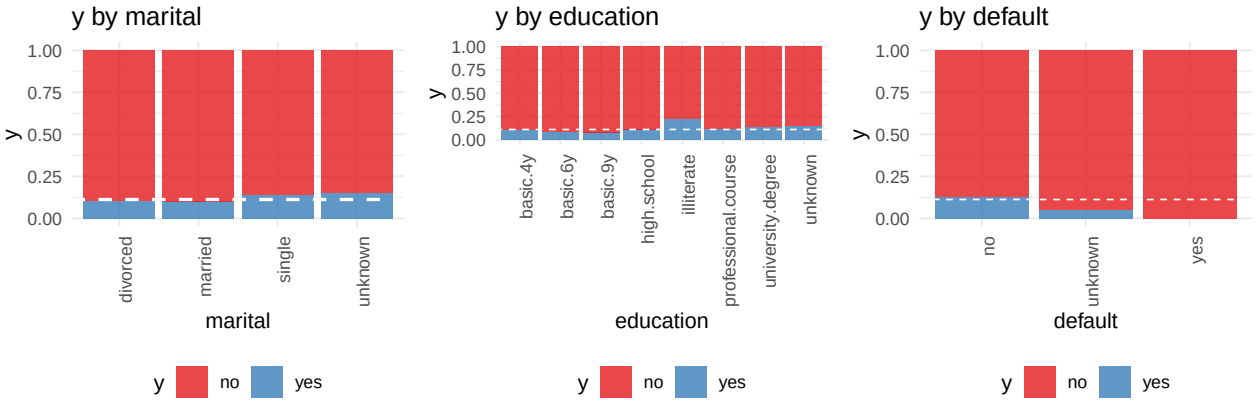

Figure 1 Left Panel: Percentage plot of y by age; Cental Panel: Percentage plot of y by education; Right Panel: Percentage plot of y by default. The white dotted line denotes the overall percentage of 'yes' in y.

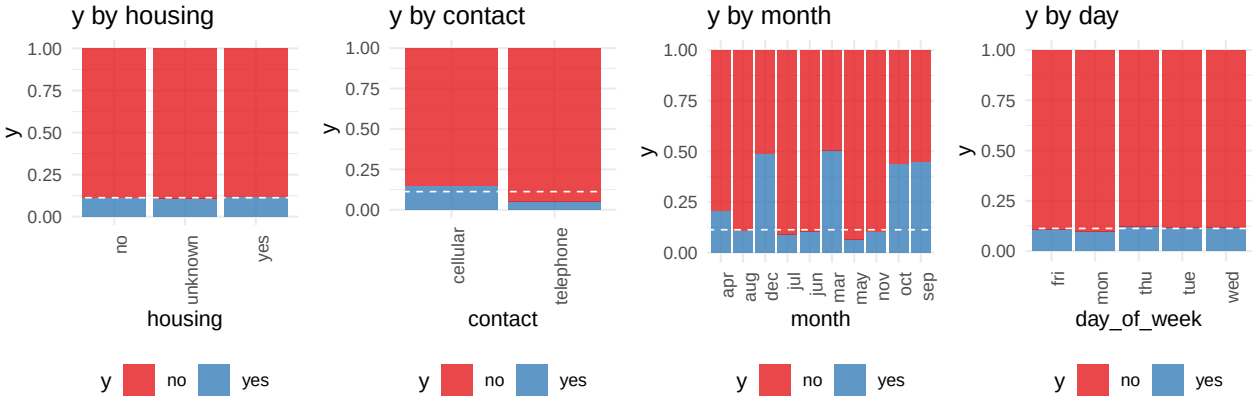

Figure 2 Percentage plots of y by housing, contact, month, and day_of_week. The white dotted line denotes the overall percentage of 'yes' in y.

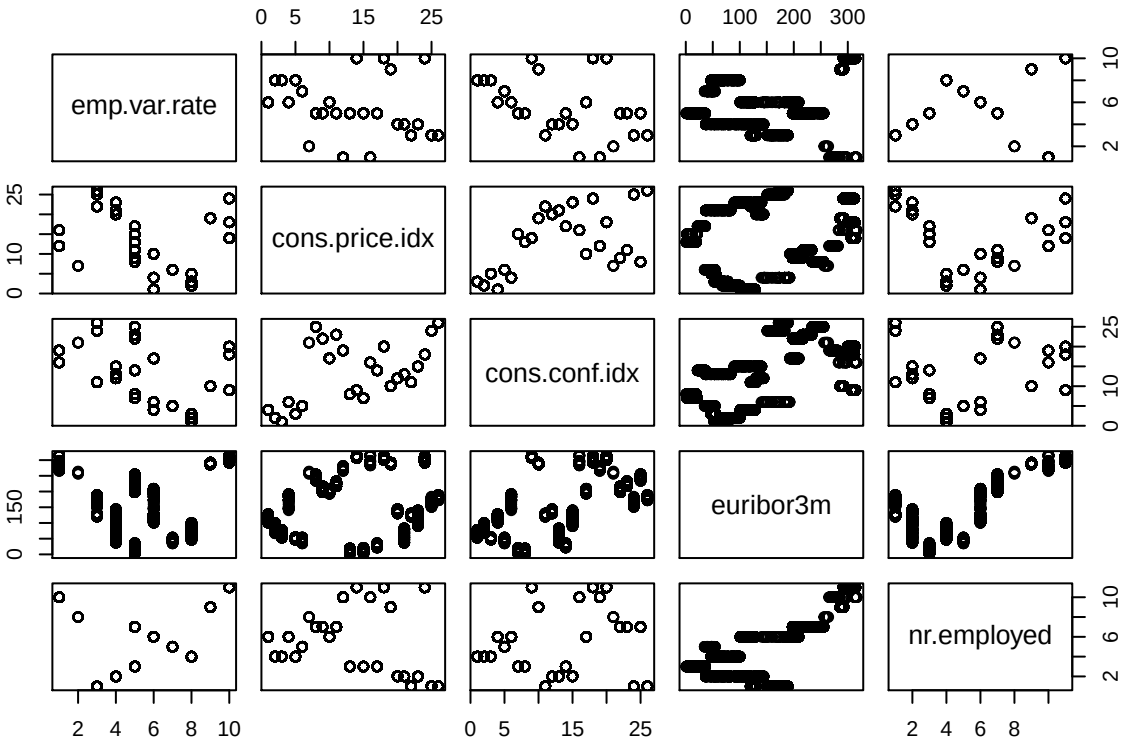

Figure 3 Pairwise plots of four variables.