Predicting Effectiveness of Bank Marketing

Group 14: Mahima Mago, Ninad Khare, Shreyansi Jain, Subasish Behera

Contents

1 Introduction

A Portuguese banking institution launched directed marketing campaign to promote their products. These marketing campaigns were based on telephonic calls. It is important for the institution to know whether the campaign is effective in converting clients, thus data was collected about the subscription of the product by the clients contacted. In this report, we aim to apply different classification techniques to the data gathered to predict the success of this bank marketing campaign. The performance of the different classification techniques, like Support Vector Machines, Decision Trees, K nearest neighbors and so on, will be compared against each other to determine the model that produces the best results. This analysis can help banks optimize their marketing campaigns by targeting customers that are more likely to subscribe to the term deposit.

2 Description of the Dataset

The dataset under study relates to 17 campaigns that occurred between May 2008 and November 2010.

Table 1: Description of the Bank Marketing Dataset

Variable Name	Description
age	age at the contact date (Numeric)
job	type of job (11 Categories)
marital	marital status (3 categories)
education	education level (7 Categories)
default	has credit in default? ('no', 'yes')
housing	has housing loan? ('no', 'yes')
loan	has personal loan? ('no', 'yes')
contact	contact communication type ('cellular', 'telephone')
month	last contact month of year (10 Categories)
day_of_week	last contact day of the week (5 Categories)
duration	last contact duration, in seconds (Numeric)
campaign	number of contacts performed during this campaign and for this client (Numeric)
pdays	number of days that passed by after the client was last contacted from a previous campaign
previous	number of contacts performed before this campaign and for this client (Numeric)
poutcome	outcome of the previous marketing campaign (3 Categories)
emp.var.rate	employment variation rate – quarterly indicator (Numeric)
cons.price.idx	consumer price index – monthly indicator (Numeric)
cons.conf.idx	consumer confidence index – monthly indicator (Numeric)
euribor3m	euribor (Euro Interbank Offered Rate) 3 month rate – daily indicator (Numeric)
nr.employed	number of employees – quarterly indicator (Numeric)
y	has the client subscribed a term deposit? ('no','yes')

Table 1 gives the details about each variable in our dataset including the description, type of variable and the number of categories present. Note that that there are missing values for some variables which will be dealt with in the further sections and have not been mentioned in the table above. The target variable in this dataset is whether or not the customer responded positively to the bank's marketing campaign. This is indicated by the binary variable "y".

It is important to note that the original dataset gathered by the Portuguese researchers contains over 40,000 observations, however this analysis is based on a subset of the complete dataset containing randomly selected 10,000 observations. Thus, the analysis in the report has less predictive power and accuracy.

3 Exploratory Data Analysis

Descriptive Analysis has been performed to understand the overall structure and features of the dataset.

3.1 Data Cleaning

The first step is to the clean the data by looking into missing values and any anomalies in the dataset. Table 2 shows the number of missing values in the data for each variable and the percentage of the number of the missing values out of the total observations.

Table 2: Number of Missing values

Variable	no_missing	perc(%)
default	2151	21.51
education	402	4.02
housing	241	2.41
loan	241	2.41
job	84	0.84
marital	25	0.25

For the variables 'Education', 'Housing', 'Loan', 'Job' and 'Marital', the number of missing values are less than 5% of the total observations. The proportion of the response variables for these values is the same as the data as a whole. These observations are not expected to be very influential in the classification. Thus, these observations have been removed from the dataset.

Table 3: Split of the Default Variable

Value	Frequency	perc
missing	1931	20.76
no	7371	79.23
yes	1	0.01

Additionally, after the above adjustment, a few variables have been analysed to check their influence on the response variable.

- Table 3 shows the split of the 'Default' variable. It can be seen that 79.23% individuals answered "no" and 20.76% did not reply at all. Hence, this variable is not of much significance.
- A chi square test performed on the variable 'Loan' resulted in the p value of 15%. Thus, there is not significant relationship of this variable with the response variable.
- As our goal to to determine whether the client will subscribe to the term deposit or not and it is difficult to know the duration of the call before hand. Thus has little influence on the response variable

Based on the above, the variables 'Default', 'Loan' and 'Duration' have been removed from the dataset.

3.2 Data Splitting

The cleaned data contains 9303 observations. This is split into training and test datasets for the modelling and testing purposes. The training contains 7442 observations (80% of the cleaned data) and the test data contains 1861 observations (20% of the cleaned data).

The exploratory data analysis has been performed on the Training dataset.

3.3 Exploratory Analysis of Categorical Variables

Table 4 shows the summary of the categorical variables. The table shows the number of unique categories for each variable. It also shows that there are no remaining missing values.

Table 4: Summary Statistics of Categorical Variables

Variable	Missing	Complete_Rate	Unique
job	0	1	11
marital	0	1	3
education	0	1	7
housing	0	1	2
contact	0	1	2
month	0	1	10
day_of_week	0	1	5
poutcome	0	1	3
y	0	1	2

The frequency plots and the barplots showing the proportion of the 2 categories of the response variable have been analysed (*all variables have not been displayed*). Some inferences from the plots are given below:

- Variable 'Job'- The proportion of 'retired' individuals resulting in 'yes' is higher than the other categories
- Variable 'Education' - 'university degree' and 'professional course' have a higher proportion of resulting in 'yes'
- Variable 'Contact'- there have been more term deposits from cellular responders, 14.5% as compared to telephone responders which is just 5.6%
- Variable 'marital'- the proportion that resulted in 'yes' is not marked differently across marital status

3.4 Exploratory Analysis of Numeric Variables

Table 5 shows the summary statistics of the numeric variables.

Table 5: Summary statistics for numerical variables

Variable	n	Mean	SD	Min	Median	Max	IQR
age	7,442	39.86	10.21	17.00	38.00	95.00	9.00
emp.var.rate	7,442	0.08	1.56	-3.40	1.10	1.40	0.30
cons.price.idx	7,442	93.58	0.58	92.20	93.44	94.77	0.55
cons.conf.idx	7,442	-40.54	4.65	-50.80	-41.80	-26.90	5.40
euribor3m	7,442	3.61	1.74	0.63	4.86	5.04	0.10
nr.employed	7,442	5,166.16	72.86	4,963.60	5,191.00	5,228.10	37.10
campaign	7,442	2.55	2.73	1.00	2.00	41.00	1.00

The figure 1 shows the pairs plot of all numeric variables along with the categorical response variable 'y'. Some inferences from the plots are given below:

- There is high correlation between some variables. 'emp.var.rate' has a very strong correlation with multiple variables. Correlation of 89.99% with 'nr.employed', correlation of 97.14% and correlation of 48.41% with 'cons.price.idx'.

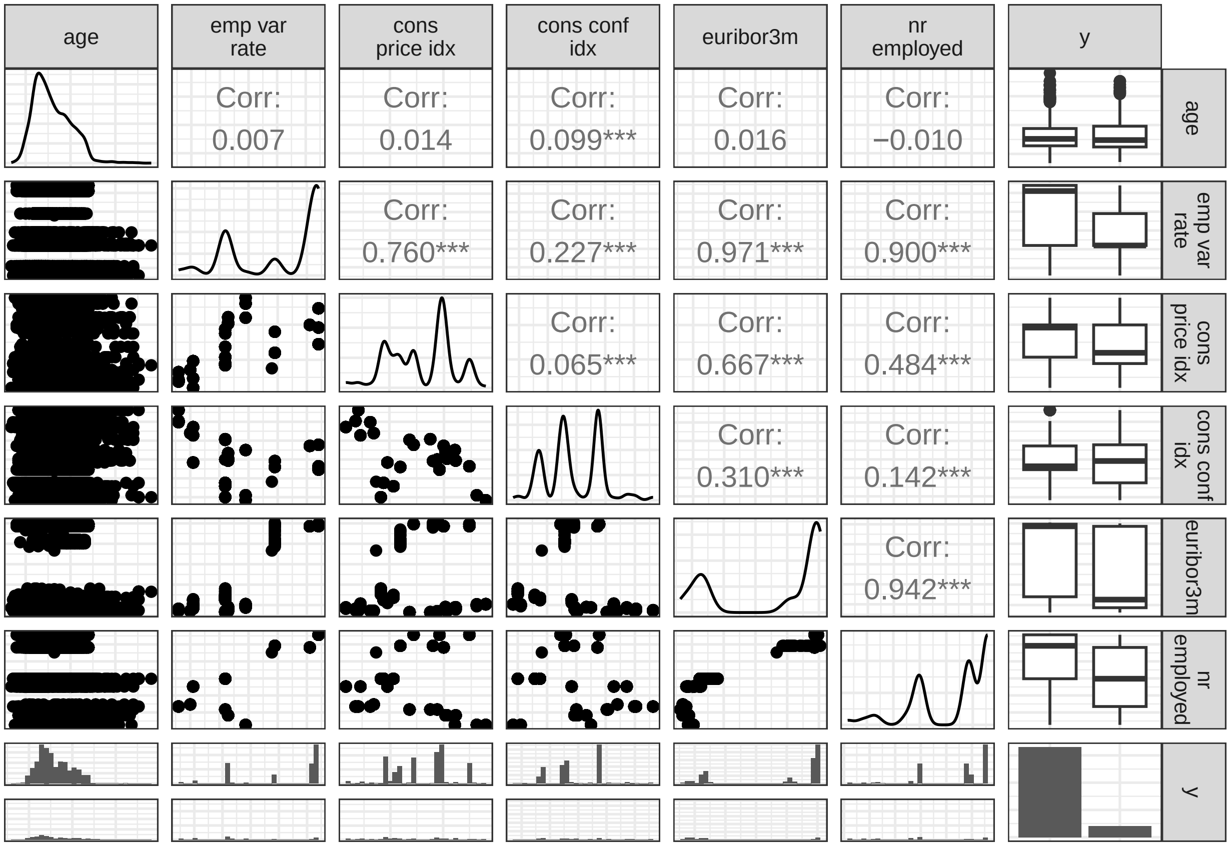

Figure 1: Correlation plots for numerical variables

- The boxplots for ‘emp.var.rate’ and ‘euribor3m’ show some inherent difference between the categories. Figure 2 shows the density plots for these variables which show a difference in the density of the 2 categories.
- For the variable ‘pdays’ 999 means that the client has not been contacted before, this constitues majority of the clients. In the training data 7174 observations have the values 999 for the variable ‘pdays’. Thus, this variable has been converted into a binary variable ‘n_contact’ which defines whether the individual was previously contacted or not. Additionally, a chi-squared test indicates that when a client is contacted before, they are more likely to say ‘yes’.
- The ‘campaign’ variable has been converted into a binary variable with the categories ‘less than 15’ and ’ more than 15’ considering the split of the data.

3.5 Data Scaling

The data being used for further processing and model building includes the following variables after the cleaning and some data transformation performed in the above sections:

1. Categorical Variables: job, marital, education, housing, contact, month, day_of_week, campaign, poutcome, n_contact, y
2. Numerical Variables: age, previous, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed
   y is the binary response variable for the study

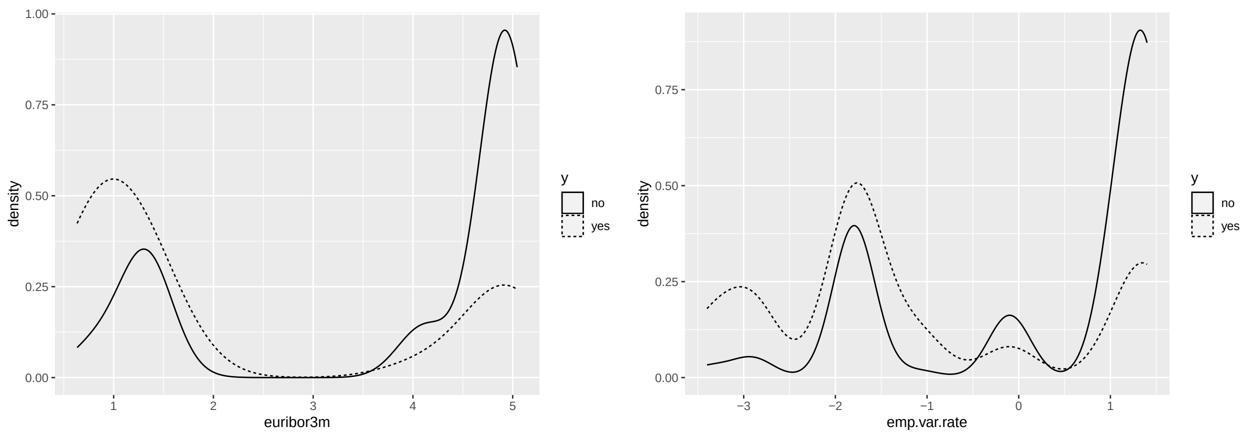

Figure 2: Density Plots of ’euribor3m’ and ’emp.var.rate

It is important to note that the numerical variables in the data are measured at different scales and thus do not contribute equally to the model fitting. This might create a bias. Thus, The numerical variables have been scaled using min-max normalization. The mathematical formulation of this is:

$$x_{scaled} = \frac{x - min(x)}{max(x) - min(x)}$$

4 Statistical Modelling

4.1 Model 1: Support Vector Machines

Support vector machine is a classification technique used for a binary output variable. In its basic form, SVM is a linear classifier which fits a hyperplane to the data that divides the two classes of the binary output into separate regions. Any new point in either region is classified as such.

We can use kernels to use non-linear classification with SVM.

If a linear classification is not possible, or does not give particularly good predictions, we can transform the input variables into a higher dimensional vector space which can be classified using a linear hyperplane in the higher dimensional space and then project the result back onto our original vector space. Projecting the decision boundary in a lower dimension vector space will lead to a non-linear classifier in the original vector space. We can make the calculations easier by transforming the variables with known functions which are called kernels.

Commonly used kernels are:

- Linear
- Polynomial
- Radial

All of the above 3 kernels have been considered in this analysis.

So, the training dataset was split into a training set and a validation set for hyperparameter tuning. After running the model for the 3 kernels and different values for the hyper parameter, the model with minimum error was chosen as well as maximum accuracy and sensitivity.

From the above results, the ‘linear’ kernel with the cost parameter set to ‘0.01’ turned out to gives us the best result for the SVM classifier with an accuracy of 90% and sensitivity of 23%.

Even though the accuracy of the model is 90%, the sensitivity is low. The reason could be an unbalanced dataset. But a different model might be a better fit for the given data.

4.2 Model 2: K-Nearest Neighbours

Under k-nearest neighbours classification technique, the k-nearest labelled points predict the class of this point to be the class that most of its neighbours share.

Here, the value of k has been chosen through Cross-validation approach. Cross validation has been done in two different ways - K-fold cross validation and leave-one-out cross validation. In K- fold cross validation, the data is divided into K roughly equal sized parts. Firstly, the validation data is taken as the first set and training data as all the other sets and the validation error rate/classification rate is estimated for this split. The process is then repeated K-1 more times, with a different part of the data set as the validation data. The final error rate is the average of the K error rates estimated.

Leave-one-out cross validation is performed entirely on the training data.

From the above results, the optimum k value from both the methods is coming out to be 6, however the Accuracy and Sensitivity rates are higher for leave-one-out cross validation. Thus, we would choose model 2 out of the two models.

4.3 Model 3: Decision Tree

4.4 Model 4: Random Forest

4.5 Model 5: Gradient Boosting

Boosting algorithm is a method used to create an ensemble of simple individual models that together create a better model. First, an initial model is fitted to the data. Then a second model is built that focuses on correctly predicting the cases that were incorrectly predicted by the first model. The combination of these two models is better than either of the two models alone. This process of boosting is repeated, with each successive model attempting to correct for the shortcomings of the combined boosted ensemble of all the previous models.

Gradient Boosting is a classification technique that uses this boosting algorithm. The word gradient is used because here the target outcomes for each case are set, based on the gradient (partial derivative of our loss function) of the error with respect to the prediction. This means that the target outcome for each case depends on how much changing that case's prediction impacts the prediction error. The gradient can be used to find the direction in which to change the model parameters to reduce the error in the next round of training.

When the target variable is continuous, we use Gradient Boosting Regressor and when it is a classification problem, we use Gradient Boosting Classifier. Since our target column is binary, we have used Gradient Boosting Classifier.

Here, we use log likelihood as our loss function. When we differentiate this function, we get log(odds) which is then used to find a value for which the loss function is minimum. This minimum value is the first prediction of our base model. Next, we calculate the pseudo residuals which are then used as target variables for our second model. Finally, we get new predictions by adding this model to our base model.

For our model, we have used XGBoost, which stands for Extreme Gradient Boosting. It is a specific implementation of the Gradient Boosting method which used more accurate approximations to find the best tree model. It computes second order gradients that provides more information about the direction of the gradients and how to get the minimum of our loss function.

The sensitivity for this model is 68% and the accuracy is almost 90%. Overall, this appears to be a good fit to the data.

4.6 Model 6: Linear Discriminant Analysis

5 Results

6 Conclusion