Who is more likely to subscribe term deposit??

Author: Betty Soo Ying Lim (s3414351)   
Contact Details: [s3414351@student.rmit.edu.au](mailto:s3414351@student.rmit.edu.au)   
Date of Report: 16/05/2017

# Table of Content

**Executive Summary 3**

**Introduction 4**

**Methodology 5**

**Results 7**

**Discussion 39**

**Conclusion 41**

**Reference 42**

# Executive Summary

The aim of this report is to build a classification model to predict whether customers will subscribe a term deposit. The data is collected during telemarketing campaign of a Portuguese bank where product (term deposit) is promoted to clients via phone calls made. Overall, the results show that there is certain type of customers with certain characteristics have higher probability to subscribe the term deposit compared to others. The report concludes that bank can utilise these findings and the classification model built to develop an effective marketing strategies that target on potential customers only without wasting resources on customers that might not be interested in subscribing term deposit

# Introduction

Financial Institution such as banks, offer term deposits as one of their cash investment services available for customers. Term deposit is a cash investment deposited by customers with bank for an agreed interest rate over a fixed term (Ryan & Smith). Customers are only allowed to withdrawal the cash when the term ends to earn interest (Ryan & Smith). Banks can use these cash deposited to grow their economic profits by lending it out in a higher interest rate or invest in a high return investment. Thus, marketing campaigns are frequently launched by banks to attract more customers to deposit cash with them. In order to have a time and cost- efficiency marketing strategies within these campaigns, data mining can be incorporated where information can be generated based on historical data available (Rust, Moorman & Bhalla, 2010). Banks can then design customised marketing strategies by making use of this meaningful information to specifically target customers who are determined to have higher likelihood to subscribe a term deposit when contacted on phone. Hence, the purpose of this report is to build a model to determine customers who are more likely to subscribe term deposit by classifying them accordingly. In this report, several classification models are utilised to solve this classification problem.

# Methodology

The data is collected from a Portuguese banking institution, from year 2008 to year 2010, retrieved from <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>. Under this telemarketing campaign, the product (term deposit) was promoted via either phone call made to a list of clients or when clients called to call centre for other reasons (Moro, Paulo & Rita, 2014). Consequently, the outcome is a binary successful or unsuccessful contact (Moro, Paulo & Rita, 2014). The dataset used contains only 10% (4119) of clients’ records, which are randomly selected from the full dataset that contains 41188 rows, with around 150 potential useful attributes. These 150 attributes undergo feature selection to include only the more relevant attributes (Moro, Paulo & Rita, 2014). Hence, this narrow down to 20 features being chosen in the dataset, among which 10 are categorical attributes and the rest are real-valued attributes, along with a class attribute (contact outcome) denoted as y in the dataset. Each row in the dataset, representing each record contains the contact result, together with client’s input features. These features consist of attributes regarding information of bank clients, attributes associated with last contact of current campaign with clients, attributes about previous campaigns with clients and also social and economic context attributes. For the purpose of data modelling, all categorical values have been transformed to meaningless values. Also, the attribute Duration significantly influences the class attribute (contact outcome). When duration is 0, then contact outcome is unsuccessful (y=0), but duration is usually equal to 0 before a call is made and will only know after the call is made. Thus, it was removed in order to have a realistic model.

In this report, the classification models used include K-nearest Neighbour model, Decision Tree model as well as Logistic Regression model.

K-Nearest Neighbor is a simple algorithm that is often used in solving classification problems. It stores the training data and a new record is classified by majority vote of its neighbours, which are the similar records in the training set (Larose, 2014). The similarity between new record and records in training set are measured using distance metrics. (Larose, 2014). There are many different distance functions that can be used for continuous variables and categorical variables respectively (Larose, 2014). There is no explicit way to determine suitable value of k (Larose, 2014). However, if value of k is too small, this might eventually result in model overfitting, where classification may be affected by many outliers (Larose, 2014). If high value of k is used, this leads to exclusion any particular behaviour learned from training set (Larose, 2014). Therefore, it is required to try different values of k and select the one with lowest classification error rate (Larose, 2014).

Besides that, one of the commonly used classifier is Decision Tree Classifier which builds classification models in a tree structure form. The Decision Tree Classifier splits dataset into 2 or more smaller sets according to differentiator in input variables, meanwhile, establishing an associated decision tree (Tan, Steinbach & Kumar, 2006). It is constructed by asking a list of questions regarding the attributes of records (Tan, Steinbach & Kumar, 2006). Follow-up questions are asked continuously until a conclusion about the class label of record is attained (Tan, Steinbach & Kumar, 2006). Thus, this leads to a decision tree with non-terminal nodes (root node or decision nodes) that represent test conditions on attribute and leaf nodes   
  
  
indicating the class label (Tan, Steinbach & Kumar, 2006). After decision tree has been constructed, it is fairly easy to classify a test record (Tan, Steinbach & Kumar, 2006). The test conditions were applied to the test record which begin from root node and then follow the correct branch based on outcome of the test (Tan, Steinbach & Kumar, 2006). This result in either another non-terminal node where a new test condition is applied or a leaf node stating the class label of the record (Tan, Steinbach & Kumar, 2006).

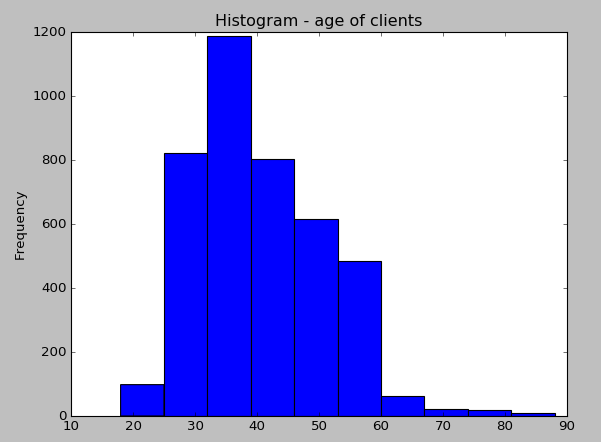
Moreover, another algorithm used for classification problems includes Logistic Regression. It is a linear classifier that assume linear relationship between input variables and output variable, where input variables are usually transformed to better represent this linear relationship, resulting in a more precise model (Brownless, 2016). It is a type of probabilistic classification model where the dependent variable (output variable) need be binary, meaning a new record can only be assigned to either 1 of the 2 classes of output variable (Brownless, 2016). According to the Logistic Regression equation, the coefficients value (denoted as beta) of the equation represent the model where each column in input data has a related coefficient values learned during training process (Brownless, 2016). Logistic Regression models the probability of output variable based on input features, combine them linearly and then apply the logistic function to this combination (Brownless, 2016). Logistic function is a S-shaped curve that transform actual numerical value by taking in numbers and map them into value of range from 0 to 1. The advantage of using this model is that it does not have significant impact from class imbalance (Brownless, 2016). However, model overfitting might occur when the inputs variables are highly-correlated among each other (Brownless, 2016).

The accuracy of these models built can then be tested and the performance of the model can be evaluated in terms of confusion matrix, classification error rate, precision, recall and f1-score. Confusion matrix is a matrix that summarise the total number of correct and incorrect predictions, broken down by each class where the main diagonals in the matrix represent the correct prediction in each class and all others are incorrectly classified (Santra & Christy, 2012). The classification error rate is fraction of incorrect classification by the model, calculated by dividing the total number of wrong predictions by total number of instances (Krzanowski, 2005). The precision, also known as Confidence, measures the fraction of predicted positive cases that are correctly real positive (Powers, 2007). The recall, also known as Sensitivity, indicates the fraction of Real Positive cases that are correctly predicted positive (Powers, 2007). For both precision and recall, a low value signifies a large number of false positive predicted by the mode, indicating a poor performance of the model. The F1-score is the weighted average of precision and recall score (Powers, 2007)

# Results

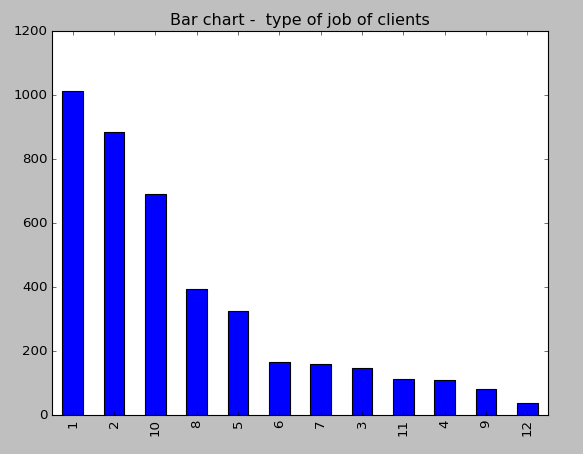
Each attribute in the dataset is explored using appropriate descriptive statistic and graphs respectively, as showed below.

**age**



The histogram above shows the age of clients in the dataset. According to the histogram above, the age of client in this dataset range from around 20 years old to 90 years old. The mean of age is approximately 40 years old.

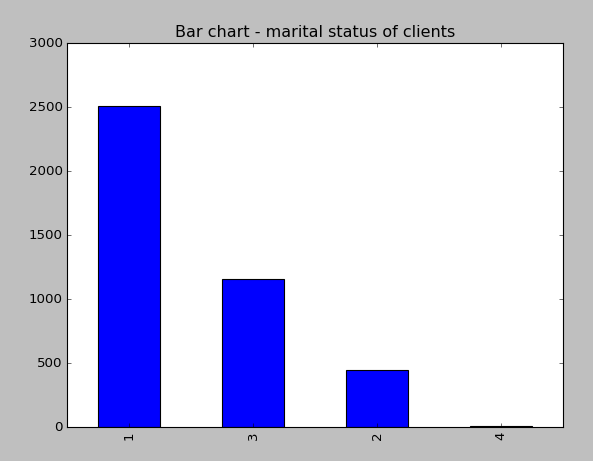
**job**



The bar chart above shows the type of jobs that clients from this dataset work as, where value 1 represents ‘admin.’, 2 represents ‘blue-collar’, 3 represents ‘entrepreneur’, 4 represents ‘housemaid’, 5 represents ‘management’, 6 represents ‘retired’, value 7 represents ‘self-employed’, 8 represents ‘services’, 9 represents ‘student’, 10 represents ‘technician’, 11 represents ‘unemployed’ and value 12 represents ‘unknown’.

According to the bar chart, out of the 4119 clients in the dataset, the type of job that 1012 of them work as is admin, 884 of them work as blue-collar, 691 of them are technician, 393 work in services industry, 324 in management, 166 retired from work, 159 of the clients are self-employed, 148 are entrepreneur, 111 are unemployed, 110 work as housemaid, 82 of them are still students and 39 remain unknown.

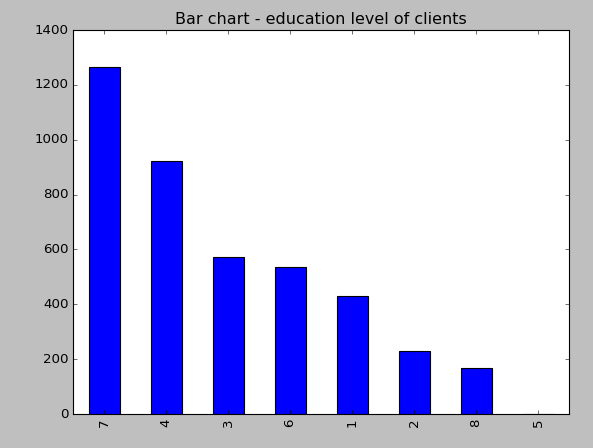
**marital**



The bar chart above shows marital status of the 4119 clients, where value 1 represents ‘married’, value 2 represents ‘divorced’, value 3 represents ‘single’ and value 4 represents ‘unknown’.

Based on the bar chart, 2509 clients in the dataset are married, while 1153 of them are single, 446 are divorced or widowed, and 11 clients where their marital status are unknown.

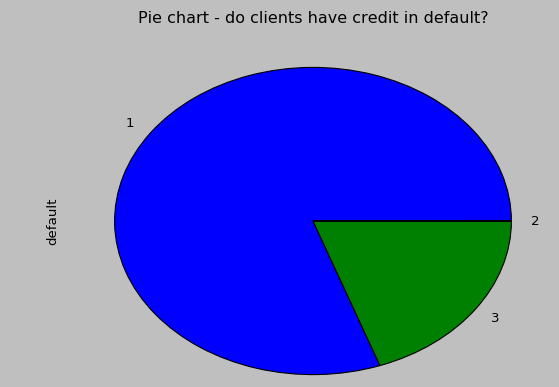
**education**



The bar chart above shows the education level of clients, where value 1 represents ‘basic.4y’, 2 represents ‘basic.6y’, 3 represents ‘basic.9y’, 4 represents ‘high.school’, value 5 represents ‘illiterate’, 6 represents ‘professional.course’, 7 represents ‘univeristy.degree’ and value 8 represents ‘unknown’.

By looking at the graph, it was found that 1264 out of 4119 clients graduated from university degree, 921 graduated from high school, 574 of them graduated from basic.9y, 535 graduated from professional course, 429 graduated from basic.4y, 228 of the clients graduated from basic.6y, 167 of their education level remain unknown and 1 of them is illiterate.

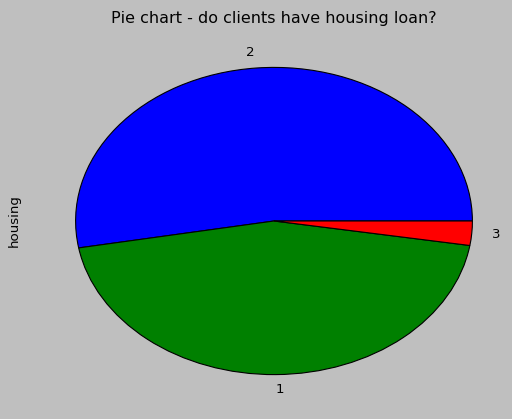
**default**



The pie chart above visualise the proportion of clients with credit in default, those without and those unknown, where value1 represents ‘no’, value 2 represents ‘yes’, value 3 represents ‘unknown’.

As seen in the pie chart, a large proportion of clients (3315) from this dataset do not have any credit in default with only 1 client has credit in default and 803 of clients where their status of credit in default is unknown

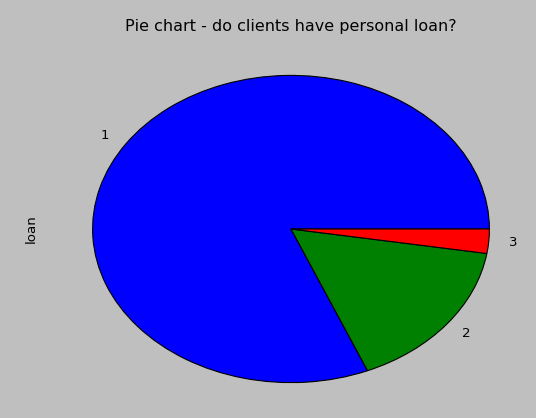
**housing**



The pie chart above illustates the proportion of clients with housing loan, those without and those unknown, where value1 represents ‘no’, value 2 represents ‘yes’, value 3 represents ‘unknown’.

According to the pie chart, from this dataset, 2175 clients have housing loan, which is slightly greater than those that do not have housing loan (1839) with the rest (105) remain unknown.

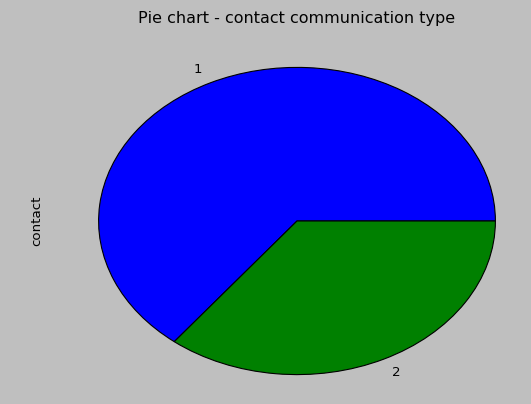
**loan**



The above pie chart shows the proportion of clients that have personal loan and those don’t have, where value1 represents ‘no’, value 2 represents ‘yes’, value 3 represents ‘unknown’.

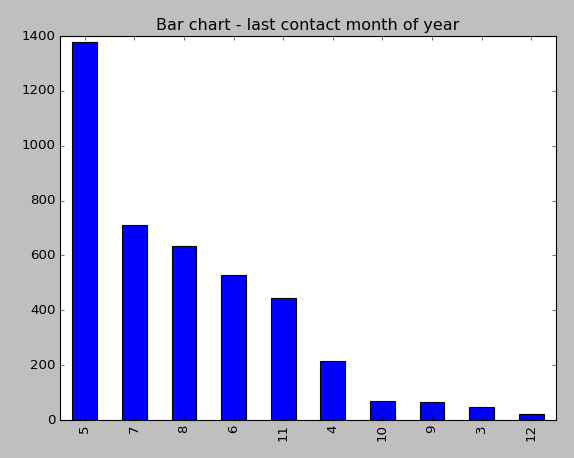
According to the pie chart, most of the clients (3349) do not have personal loan. There are only 665 clients in this dataset that have personal loan and 105 clients remain unknown.

**contact**

****

The pie chart above indicates the proportion of clients that are being contacted by cellular and those being contacted by telephone, with value 1 representing ‘cellular’ and 2 representing ‘telephone’.   
Out of 4119 clients, a larger proportion of them (2652) were contacted via cellular while 1467 of them were contacted via telephone, as seen in the graph.

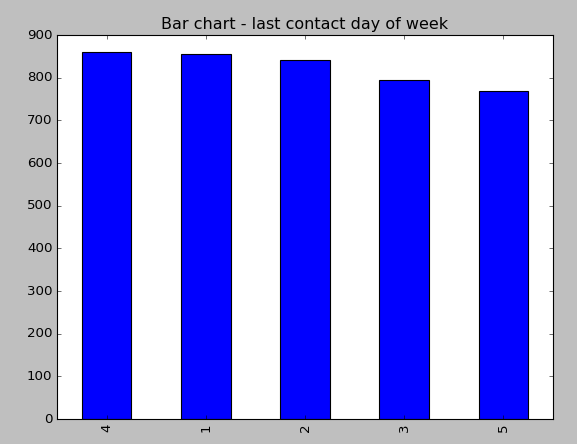
**month**



The bar chart above shows the last contact month of year for a client, where value 1 represents ‘jan’, 2 represents ‘feb’, 3 represents ‘mar’, 4 represents ‘apr’, 5 represents ‘may’, value 6 represents ‘jun’,7 represents ‘jul’, 8 represents ‘aug’, value 9 represents ‘sep’, 10 represents ‘oct’, 11– represents ‘nov’ and value 12 represents ‘dec’.

From the dataset, the last contact month of year for 1378 clients is May, 711 of them are last contacted in July, 636 are last contacted in August, 530 are last contacted in June, 446 of the clients are last contacted in November, 215 are last contacted in April, 69 of them are last contacted in October, 64 are last contacted in the month of September, 48 are last contacted in March and 22 are last contacted in December.

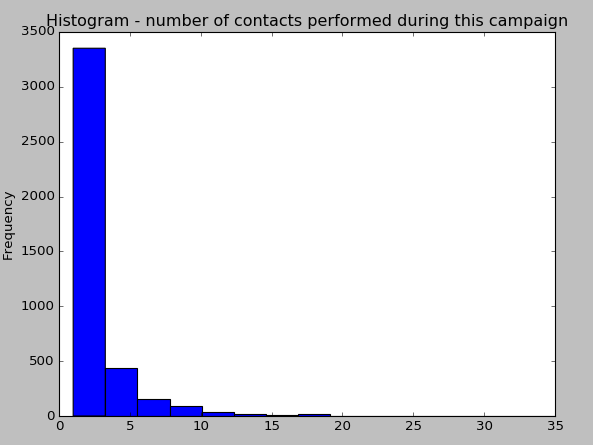
**day\_of\_week**



The bar chart above shows the last contact day of week of a client, where value 1 represents ‘mon’, 2 represents ‘tue’, 3 represents ‘wed’, 4 represents ‘thu’ and value 5 represents ‘fri’.

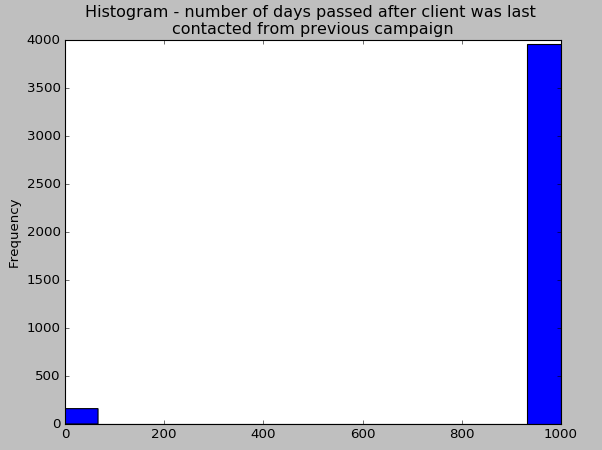
The frequency of last contact day of the week with client is highest on Thursday, which is 860 clients, followed by 855 clients last contacted on Monday and then 841 clients last contacted on Tuesday. 765 clients were contacted on Wednesday and the lowest frequency is on Friday, with only 768 of them being contacted.

**campaign**



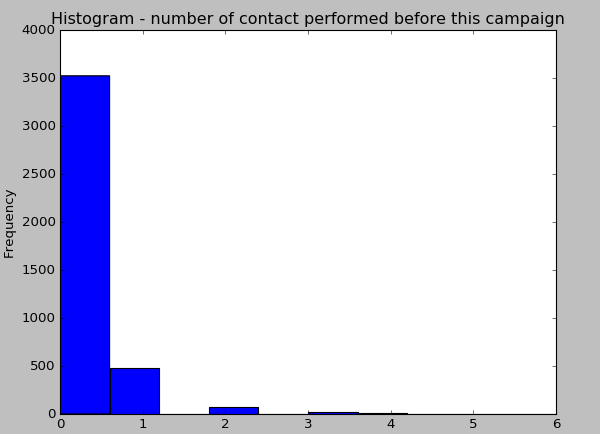
In this dataset, as seen in the graph, the number of contact performed during the campaign for each client range from 1 to 35 with a large proportion of the clients being contact around 1-5 times. The mean of number of contacted performed during the campaign is 3 times

**pdays**



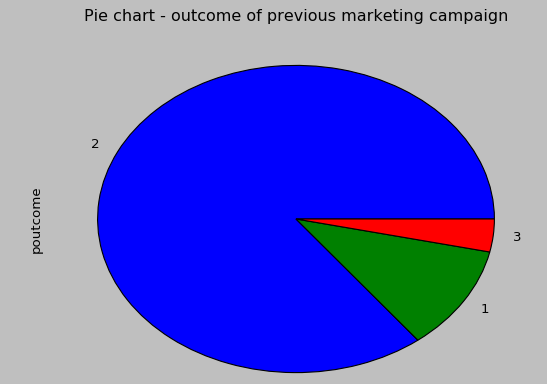
The number of days that passed by after client was last contacted from a previous campaign is within the range of 0 to 21. However, the weird shape of histogram in the graph above was due to a large amount (3959) of value 999, which indicates that clients were not previously contacted. Hence, this leads to mean of number of days to be 960days. However, if all values of 999 are removed, the mean of number of days passed after client was last contact from previous campaign will be 5.91.

**previous**



The number of contacted performed before this campaign for each client range from 0 to 6 with larger proportion of clients have 0 contact made with the bank before this campaign. The mean here is 0.19.

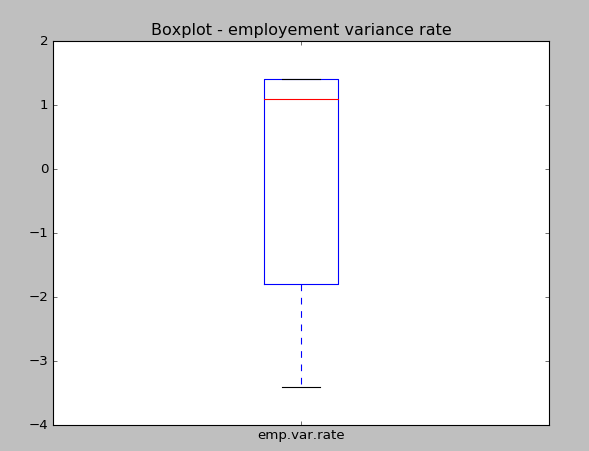
**poutcome**



The pie chart above visualises the proportion of clients where the outcome of previous marketing campaign is either success, failure or non-existance, where value 1 represents ‘failure’, 2 represents ‘nonexistance and value 3 represents ‘success’.

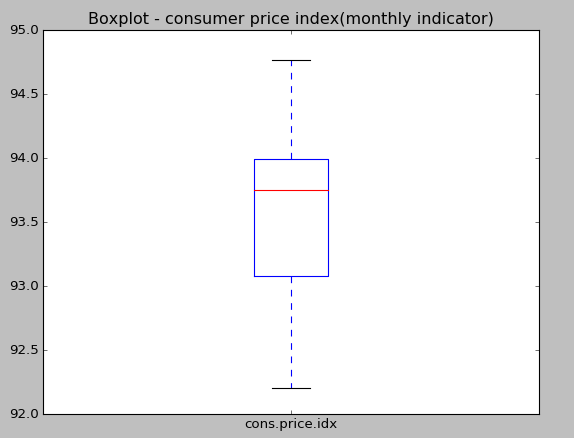
Out of 4119 clients, a large proportion (3523) of clients' outcome of previous marketing campaign do not exists. Previous marketing campaign was success in only 142 clients in this dataset while outcome of previous marketing campaign of the remaining 454 clients is fail.

**emp.var.rate**



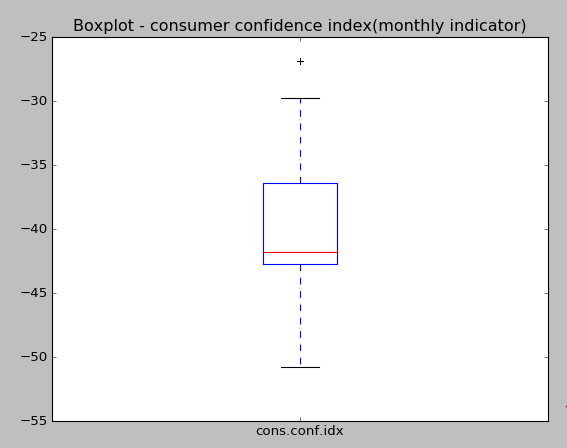
The employment variation rate ranges from -3 to 1.4 during this data collection process. The mean of employment variance rate here is 0.085.

**cons.price.idx**



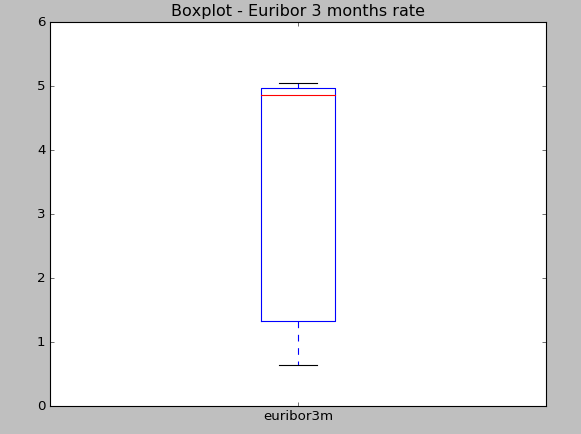
The consumer price index ranges from 92.20 to 94.77 during this dataset collection process. The mean of consumer price index here is 93.58.

**cons.conf.idx**



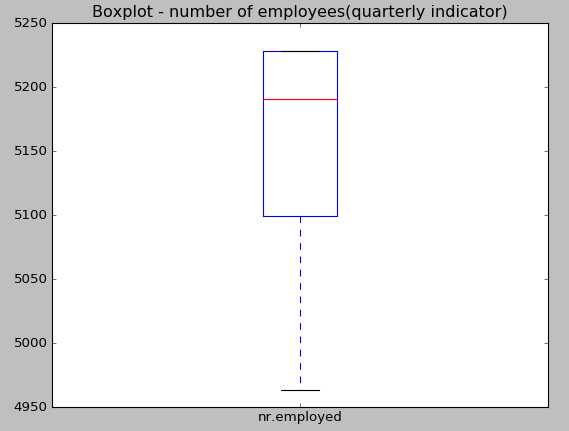
The value of consumer confidence index range from -29.8 to -50 with an outliers of -26.9 during this data collection process. The mean of the consumer confidence index is approximately -40.5.

**euribor3m**



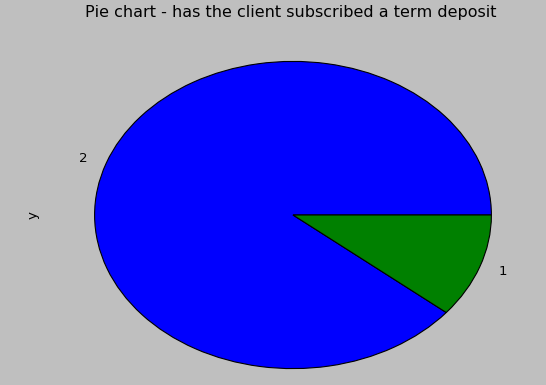
The euribior 3 months rate range from 0.7 to 5 during the whole dataset collection process. The mean of the rate is close to 5, as shown in the graph.

**nr.employed**



The number of employees varies between 4950 to 5228 when this data set was being collected. The mean of number of employees is 5166 employees.

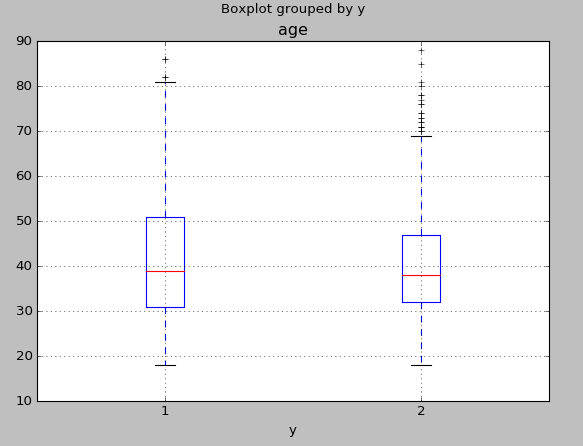
**y (output variable)**



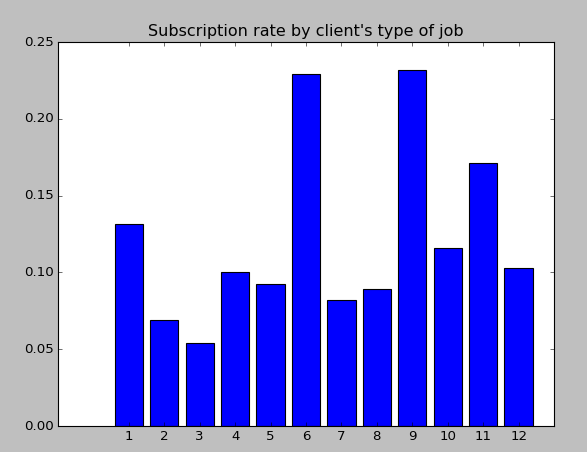
The pie charts above shows the proportion of clients who subscribed the term deposit versus those who didn’t, where value 1 represents ‘yes’(subscribe) and value 2 represents ‘no’ (didn’t subscribe).

In this dataset, out of 4119 clients, only 451(11%) respond positively to subscription of term deposit while the remaining clients (3668) respond no to subscription of term deposit.

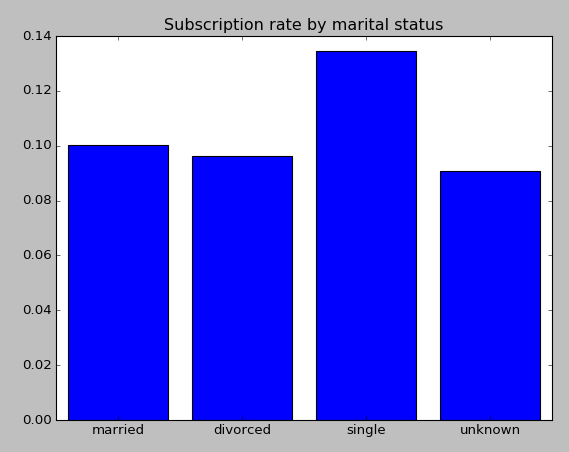
Attributes or columns in this dataset are also plotted against each other using either side-to-side boxplots or bar charts to indicate the relationships between attributes, which are shown below.



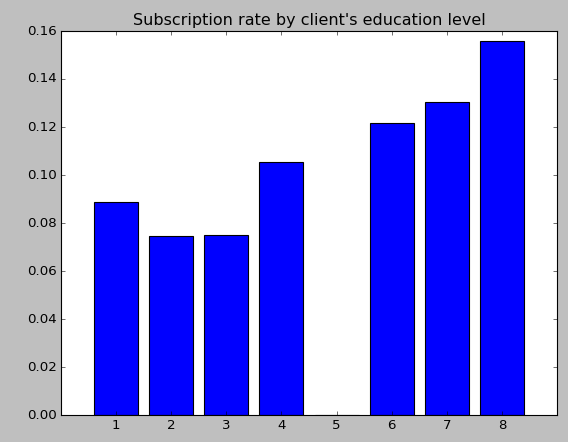
By looking at the side-to-side boxplot, the age of clients that subscribed term deposit(y=1) had wider range and less outliers than those who did not subscribe term deposit(y=2). The age of clients that subscribed term deposit has slightly higher mean when compared to those who did not subscribed.



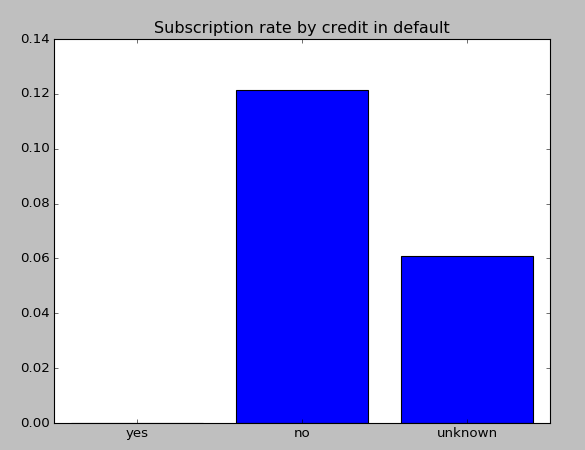
Clients who are found to be students (23.17%) and who already retired (22.89%) have higher probability to subscribe term deposit. Beside that, 17.12% of them who are unemployed subscribed, 13.14% who are admins subscribed, 11.57% of them that work as technician subscribed, 10.26% of clients’ job are unknown subscribed the term deposit, 10% of them that work as housemaid subscribed, 9.91% of clients who work in services subscribed the term deposit, 9.26% subcribed are those are in management, 8.18% who are self-employed subscribed and 6.9% of clients work as blue-collar subscribed term deposit. Clients who work as enterpreter are least likely (5.41%) to subscribe term deposit.



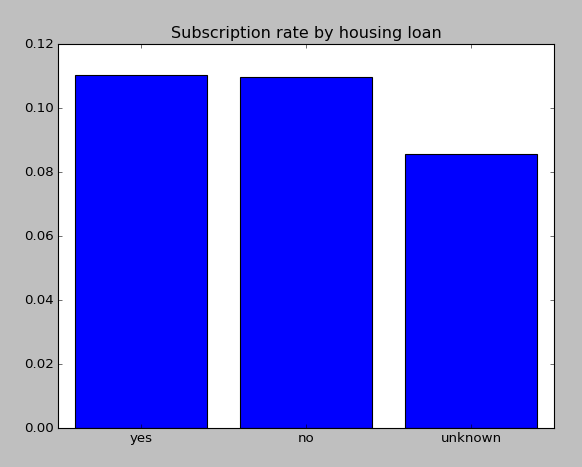
A higher percentage (13.4%) of clients who are single subscribed the term deposit, followed by clients who are married (10.04%) and then only 9.6% of them who are divoced or widowed subscribed the term deposit. Clients who have unknown marital status are least likely(9.1%) to subscribed the term deposit.



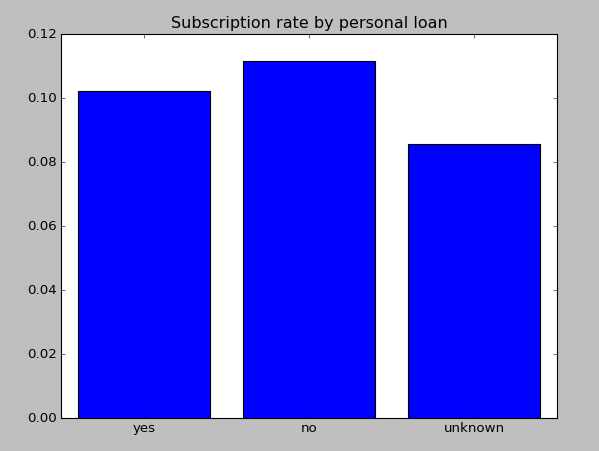
Clients who have unknown education level are found to have highest proability(15.57%) of subscribing the term depsoit. It was followed by those who graduated from univerisity degree (13.05%) and then those who gradutaed from professional course (12.15%). 10.53% of clients who graduated from high school subscribed the term deposit. The probability of clients that graduated from basic.4y subscribing term deposit is 8.86% while clients who graduated from basic.6y and basic.9y have similar probability of 7.50% of subscribing term deposit. However, there is only 1 client from the dataset that is illeterate that do not susbcribe term deposit, and thus, the probability of a client that is illeterate that subsribed term deposit is 0.



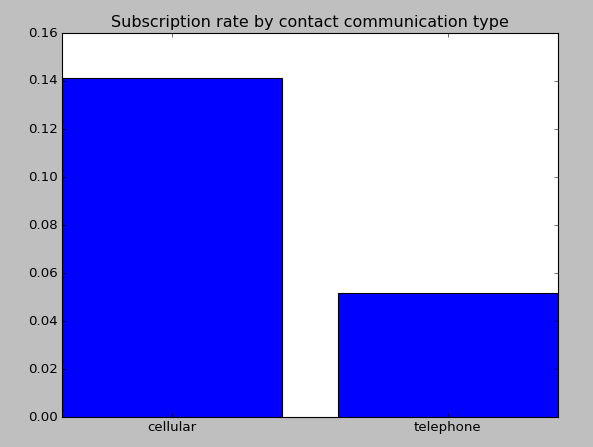
A higher percentage of 12.13% of clients in the dataset without credit in default subscribed the term deposit while 6.10% of them where the credit in default is unknown subcribe term deposit. It was found that 0 clients will subscribe term deposit if he or she has credit in default.



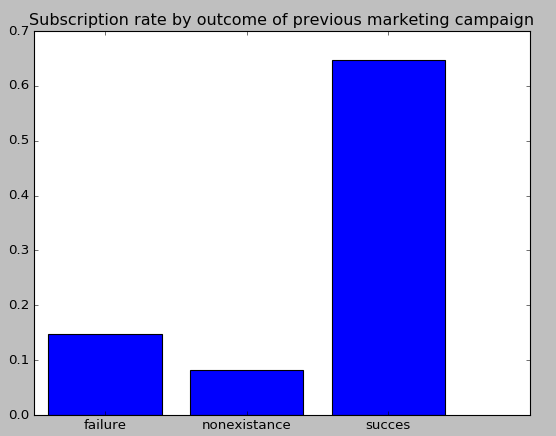
11.03% of clients with housing loan subscribe the term deposit, which is slightly higher than percentage of those who don’t have housing loan and subscribe term deposit (10.98%). 8.57% of client in the dataset where whether their have housing or not is unknown subscribed term deposit.



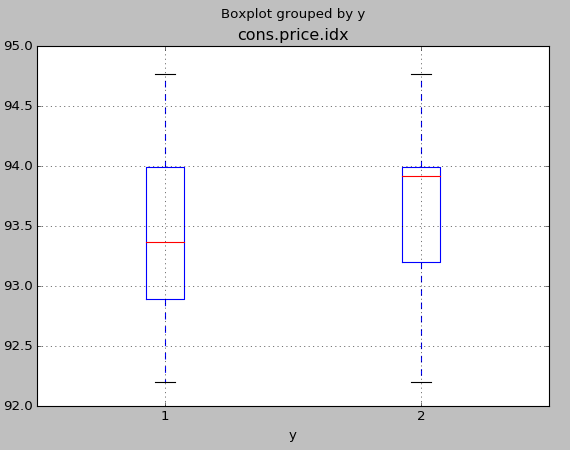
11.17% of clients that do not have personal loan subscribe the term deposit, which is higher than those who have personal loan and subscribe term deposit (10.23%). The probability of clients that subscribe term deposit where whether they have personal loan is unknown is 8.57%.



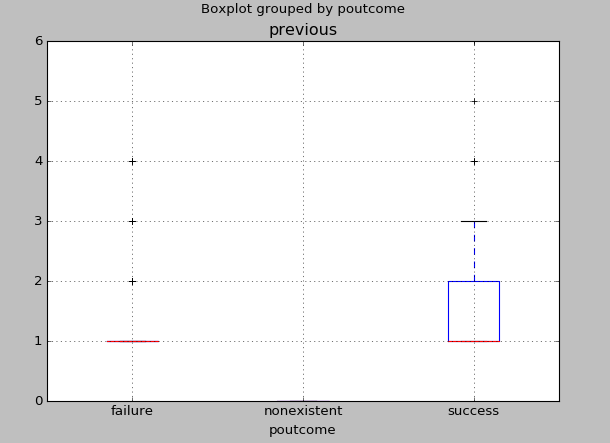
14.14% of clients where the contact communication type is cellular subscribed term deposit while only a small percentage,. 5.18% of clients that are being contacted by telephone subscribed term deposit.



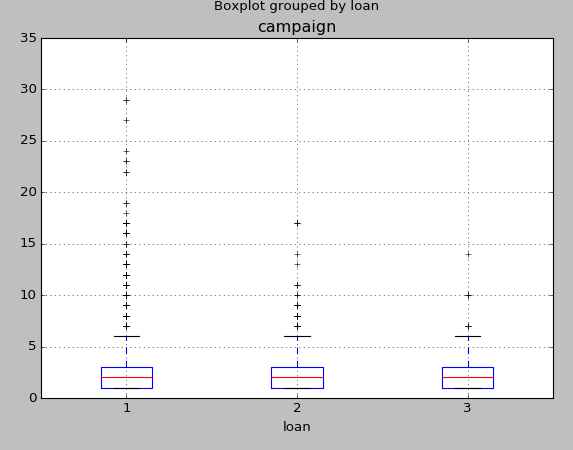
When clients’ outcome of previous marketing campaign is failure, 14.76% of them subscribed the term deposit. But, when thier outcome of previous marketing campaign is success, a higher probability, 64.79% of them subscribed the term deposit. When clients do not have outocome of previous marketing campaign, the probability of them subscribing term deposit is much lower (8.29%) than those who didn’t subscribe.



The boxplot above shows that the range of consumer price index value when clients subscribe(y=1) or didn’t subscribe(y=2) are similar. However, the mean of consumer price index is higher when client did not subscribe term deposit.

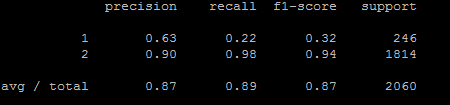


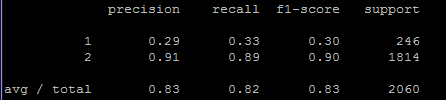
As seen from the boxplot, if one contact or more than 1 contact were made between bank and client previously, then the outcome of previous marketing campaign is most probably success while the outcome of previous marketing campaign is more likely to fail if the number of contact performed before this campaign is only one.



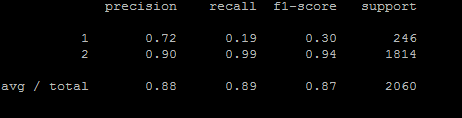
Based on the graph above, the number of contact performed during this campaign have similar range but more outliers when clients have no personal loan, when compared with those with personal loan and those where status of personal loan is unknown. However, the mean of number of contacted performed for clients with personal loan, without personal loans and unknown are similar.

For data modelling, the dataset which consists of 4119 instances are split into two, 2059 of it are for training purpose while 2060 are for testing purpose. The confusion matrixes and classification performance tables for the 3 different classifiers used to solve classification problems are shown below.

   
In the k-nearest neighbor model, the value of k is chosen to be 12. According to confusion matrix shown above, out of the 2059 instances, 53 of it were predicted correctly to be under class 1(subscribe) while 31 are predicted wrongly to be class 2(did not subscribe). On the other hand, 1783 instances were predicted correctly to be classified under class 2(not subscribe) while 193 were predicted wrongly. The classification error rate is (31+193)/2060 = 10.87. Also, based on the table above, the precision for class 1 is 0.63, recall score is 0.22 and f1-score is 0.32 while the precision, recall and f1-score for class 2 is 0.90, 0.98, 0.94 respectively. On average, the precision score is 0.87, the recall score is 0.89 and the f1-score is 0.87.

In the decision tree model, according to the confusion matrix above, 80 of instances were predicted correctly to be under class 1(subscribe) while 199 are predicted wrongly to be class 2. On the other hand, 1615 instances were predicted correctly to be classified under class 2(not subscribe) while 166 were predicted wrongly. The classification error rate is (166+199)/2060 = 17.72. Also, based on the table above, the precision for class 1 is 0.29, recall score is 0.33 and f1-score is 0.30 while the precision, recall and f1-score for class 2 is 0.91, 0.89, 0.90 respectively. On average, the precision score is 0.83, the recall score is 0.82 and the f1-score is 0.83.

In the logistic regression mode, according to the confusion matrix above, 47 of instances were predicted correctly to be under class 1(subscribe) while 18 are predicted wrongly to be class 2. On the other hand, 1796 instances were predicted correctly to be classified under class 2(not subscribe) while 199 were predicted wrongly. The classification error rate is (18+199)/2060 = 10.53 Also, based on the table above, the precision for class 1 is 0.72, recall score is 0.19 and f1-score is 0.30 while the precision, recall and f1-score for class 2 is 0.90, 0.99, 0.94 respectively. On average, the precision score is 0.88, the recall score is 0.89 and the f1-score is 0.87.

# Discussion

Based on the graphs above, from the dataset, it was found that in term of age, older people are more likely to subscribe the term deposit. With regards to education, students and those already retired are more likely to subscribe term deposit. In relations to marital status, clients who are single are found to be more likely to subscribe term deposit. Regarding education level, the term deposit is more likely to be subscribed by clients graduated from university degree, while clients who are illiterate have 0 probability to subscribe term deposit. Client without credit in default are more likely to subscribe while probability of those with credit in default to subscribe is 0. It was also found that whether a client has housing loan, does not really impact their decision in subscribing term deposit. The probability of clients without personal loan are more likely to subscribe. The probability of clients subscribing term deposit is higher when they are being contacted via cellular. Clients are more likely to subscribe if their outcome of previous marketing campaign is success. The outcome of previous marketing campaign is generally fail if previous number of contact made between clients and bank is only one. In relation to consumer price index, clients have higher likelihood to subscribe term deposit when the consumer price index is low.

The value of k used in k-nearest neighbor is determined using the precision value and classification error rate. The table below shows the value of precision and classification error rate for each value of k ranging from 3 to 13.

|  |  |  |
| --- | --- | --- |
| Value of k | Precision | Classification Error rate |
| 3 | 0.84 | 12.43 |
| 4 | 0.85 | 12.72 |
| 5 | 0.85 | 11.99 |
| 6 | 0.86 | 11.89 |
| 7 | 0.85 | 11.75 |
| 8 | 0.86 | 11.65 |
| 9 | 0.86 | 11.55 |
| 10 | 0.86 | 11.26 |
| 11 | 0.87 | 11.12 |
| **12** | **0.87** | **10.87** |
| 13 | 0.86 | 11.21 |

High value of precision with lowest classification error rate is desired as it indicates a superior performance of the model in predicting the output variable. According to the table above, the value of precision was high (0.87) when value of k is equal to 11 and 12. But, the value of 12 is chosen as it has the lower classification error rate of 10.87, meaning only 10.87% are misclassified by the model.

The performance evaluation of the 3 classification models can be summarised in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Model* | *Precision* | *Classification error rate* | *Recall* | *F1-score* |
| K-nearest Neighbor | 0.87 | 10.87 | 0.89 | 0.87 |
| Decision Tree | 0.83 | 17.72 | 0.82 | 0.83 |
| **Logistic Regression** | **0.88** | **10.53** | **0.89** | **087** |

From the table above, it can be observed that logistic regression model has the highest value of precision and lowest value of classification error rate, followed by K-nearest neighbor with a slightly lower value of precision and slightly higher value of error rate. The value of precision is lowest for decision tree and highest for classification error rate value, indicating a poor performance of the model in solving the classification problem. Hence, logistic regression model is recommended to solve the classification problem in this context of subscription of term deposit.

# Conclusion

To conclude, based on findings found from dataset and logistic regression model built, it enables bank to utilise these findings and model to predict potential customers and create customised marketing strategies to promote their product in order to increase their revenue.

However, some of the issues should be considered, as stated below:

* the number of instances in each class from this dataset are unequal, resulting in potential misleading classification accuracy. The classification model built might not be accurate, as there are only 11% records within the dataset that are related with success contact while the rest are related to unsuccessful contact.
* Attributes within the dataset such as marital, education, default, loan, housing contains missing values where some of clients’ information are unknown. For example, for attribute default, 803 clients where their status of credit in default are marked as unknown in the dataset. This might have impact in model performance of predicting.
* Instead of splitting the dataset into 2, namely, training set and test set, other validation strategy such as k-fold cross validation or leave-1 out can be apply to determine any difference between model built using different validation strategies and compare the performance of model to find out if one outperforms others.

# References

Brownless, J. (2016). Logistic Regression for Machine Learning. Retrieved form <http://machinelearningmastery.com/logistic-regression-for-machine-learning/>.

Krzanowski, W. J. (2005). Misclassification Rates. *Encyclopedia of Statistics in Behavioral Science*. doi=10.1002/0470013192.bsa197.

Larose, D. T. (2014). Discovering Knowledge in Data: An introduction to Data Mining. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/0471687545.ch5/pdf>.

Moro, S., Cortez, P. & Rita P. (2014). A Data-Driven Approach to Predict the Success of bank Telemarketing. Retrieved from <https://pdfs.semanticscholar.org/4a27/709545cfa225d8983fb4df8061fb205b9116.pdf>.

Powers, D. M. W. (2007). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Retrieved from <http://www.flinders.edu.au/science_engineering/fms/School-CSEM/publications/tech_reps-research_artfcts/TRRA_2007.pdf>

Rust, R. T., Moorman, C. & Bhalla, G. (2010). Rethinking Marketing. Retrieved from <https://hbr.org/2010/01/rethinking-marketing>.

Ryan, T. J. & Smith, T. (n.d.). What is a term deposit? Retrieved from <https://www.canstar.com.au/term-deposits/>.

Santra, A. K. & Christy, C. J. (2012). Genetic Algorithm and Confusion Matrix for Document Clustering. International Journal of Computer Science Issues, 9(1).

Tan, P. N., Steinbach, M. & Kumar, V. (2006). Classification: Basic Concepts, Decision Trees, Model Evaluation. Retrieved from <http://www.users.cs.umn.edu/~kumar/dmbook/ch4.pdf>.