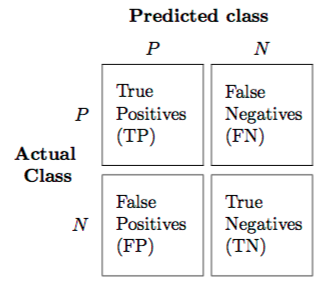
**4. Methodology**

To predict customers’ responses to bank direct marketing term deposit subscription, four algorithms are used, which are the Logistic Regression, Classification and Regression Tree, Random Forest, and Artificial Neural Network. In data processing section, I used Synthetic Minority Oversampling Technique (SMOTE) to balance the data. The balanced data is then randomly divided into training data (50% of the balanced data), and the remaining data serves as the test dataset. I used 70% training data and 30% test data before, but it showed an error message indicating that all the arguments must have the same length, so I change to 50% to 50%. Since the dataset is already balanced, the accuracy comparison is now a direct approach to validate each model and find the model that is the most accurate. The accuracy measures how accurate the certain test performs the classification, and it can be computed from the confusion matrix. The confusion matrix is a table that defines the classification performance by illustrating the actual and predicted number of classification. There are Type I and Type II errors involved in the confusion matrix, which are True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). TP represents the actual case is true and the predicted result is also true, while TN are cases which are actually false, and the test also predicted them as false. FP represents cases which are actually false but are predicted as true, while FN are those cases that are actually true, but predicted as false. Among those four elements, FP is referred as the Type-I error, and FN is known as the Type-II error. The Table ----- below shows the classical layout of the confusion matrix, and it is followed by the Accuracy formula. The reason that I am not using RMSE as a performance measurement is that RMSE is commonly used in regression, where the predictor variable is a real number. However, in classification, I have class labels or categories, so it is not corresponded to numbers. Besides, it is hard for RMSE to find difference between “a” and “b”, so I use accuracy that can be computed from the confusion matrix instead.

Table 1. Sample confusion matrix layout



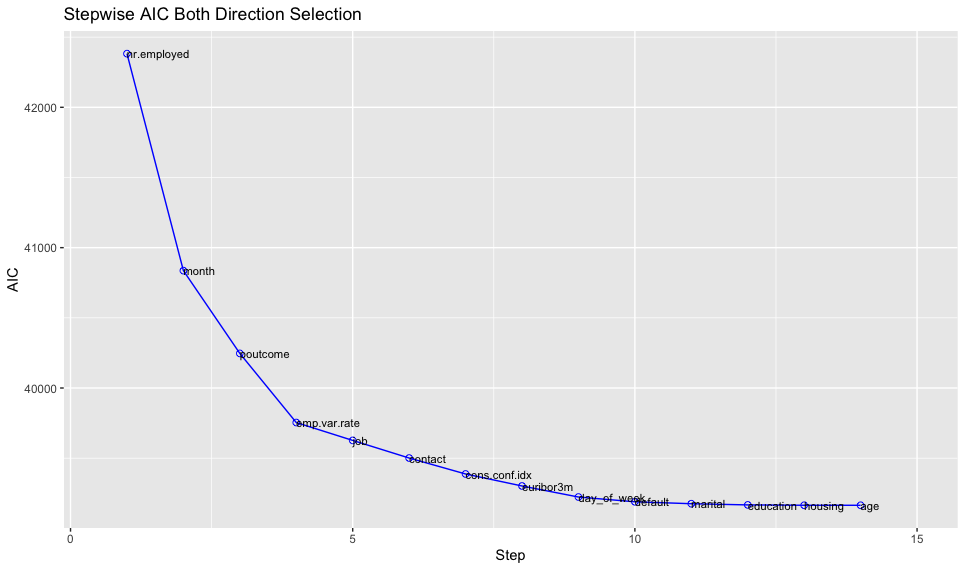
Alwis, Roshan (2016). Introduction to Confusion Matrix [Classification Modeling]. URL= <https://medium.com/tech-vision/introduction-to-confusion-matrix-classification-modeling-54d867169906>

Moreover, I used feature selection only for logistic regression and artificial neural networks, since the classification and regression tree and random forest can automatically select variables by themselves. I applied the mix selection method for logistic regression and neural networks, and the selected variables are listed in Table 2. It started with no variables in the model, and as with forward selection, I added the variable that provides the best fit, then continued to add variable one-by-one. If any point the p-value for one of the variables in the model rose above a certain threshold, we removed that variable, then continued until all variables in the model had a sufficiently low p-value. After applying this method to the balanced data, I got 14 out of 20 variables remained, which are listed below, and also the Figure 1. introduces the variable added in each step.

Table 2. Feature Selection

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ﻿nr.employed | ﻿month | ﻿poutcome | ﻿emp.var.rate | ﻿job |
| ﻿ contact | ﻿cons.conf.idx | ﻿euribor3m | ﻿day\_of\_week | ﻿default |
| ﻿marital | ﻿education | ﻿housing | ﻿age | ﻿ |

Figure 1. Feature Selection by Steps



4.1. Logistic Regression

I am interested in predicting customers’ responses to bank direct marketing campaign of term deposit subscription, and y=1 if customer subscribes the term deposit, while y=0 if customer does not subscribe. Logistic Regression is a kind of probabilistic statistical classification model, and it works well for testing hypothesis between a categorical dependent variable and one or more categorical or numerical independent variables (Elsalamony 2014). Since my dependent variable is binary, I would like to start with fitting and testing the logistic model. The Logistic Regression produces linear decision boundary, and the classification in Logistic Regression is allowed to be uncertain, which is reflected by the intermediate values between 0 and 1.

Logistic Regression Accuracy Table 3 shows that Logistic Regression did a not-so- good job in prediction accuracy, and it only obtained 71.23% prediction accuracy in validation set. Table 4. shows the analysis result of deviance, and it also provides information about the significance of variables. It indicates that except for “age” and “marital”, all remaining selected variables are significant. Figure 2. shows the variable importance with detailed categories, and it backs up the story about the significance of the majority selected variables. “poutcomesucess” has the highest importance, while “nr.employed” ranks the second. “poutcomesucess” stands for successful previous marketing campaign for certain customer, and “nr.employed” represents number of employees of direct marketing.







|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 3. Logistic Regression Accuracy | | |  | |
| LR | Training data | Test data | |
| Accuracy | 71.24% | 71.23% | |

Table 4. Logistic Regression ANOVA result

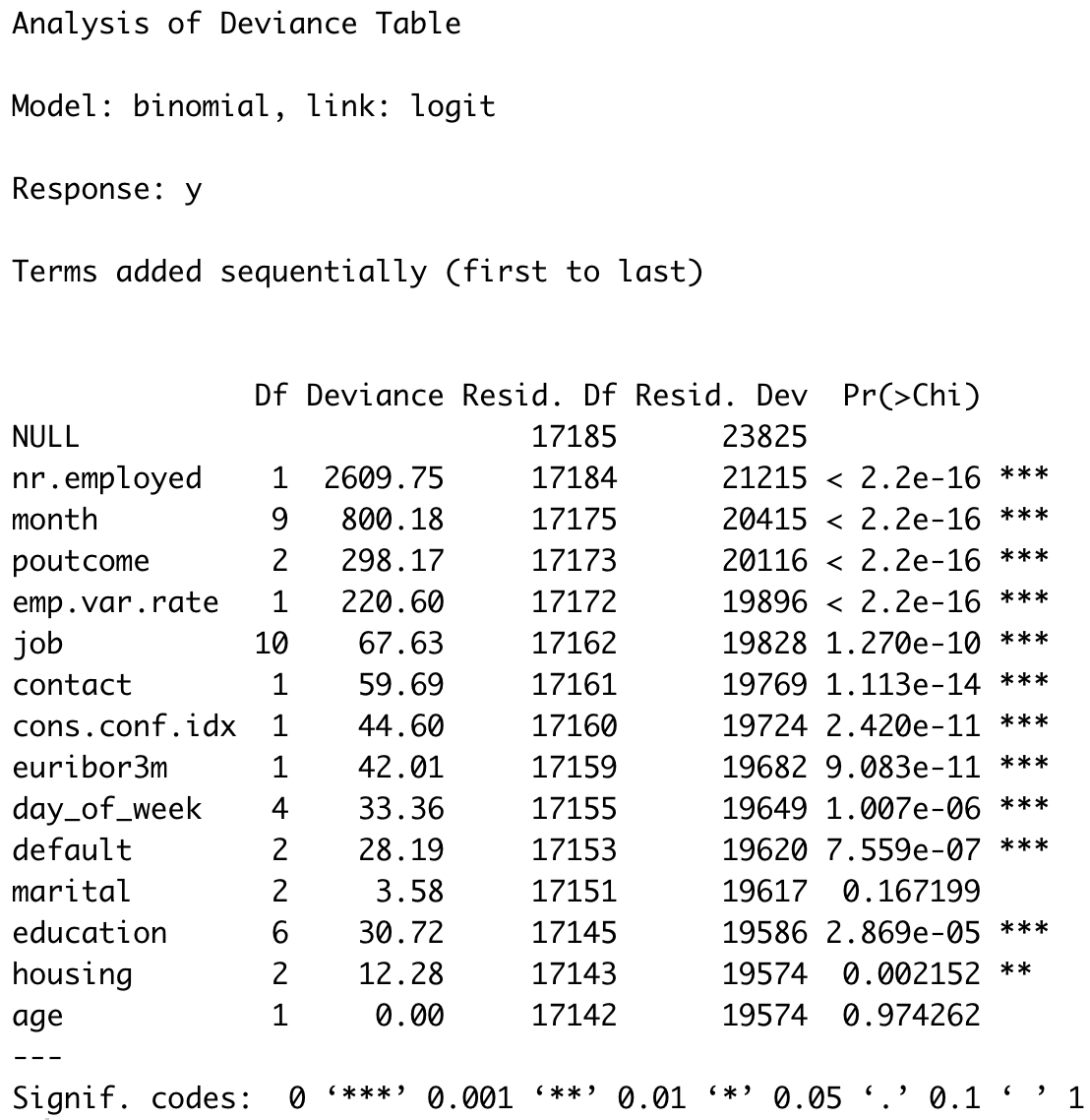


Figure 2. Logistic Regression Variable Importance



4.2. Classification and Regression Tree (CART)

Unlike logistic and linear regression, CART does not have a specific equation. Data are partitioned along the predictor variables into subsets with homogeneous values of the dependent variable and to reduce class mixing at each split. It is similar to growing a large tree and then prune it. Pruning can be done by randomly selecting a test sample and computing the error by running it down the large tree and subtrees. The tree with the smallest error will be the final tree. The tree before pruning and after are exact the same, with the result of 82.99% training accuracy and 82.51% test accuracy. The variables that are actually used in tree construction are “nr.employed” and “pdays”. The final tree with lowest error has 4 terminal nodes.

The CART automatically selected 11 out of 20 variables, which are “pdays”, “poutcome”, “nr.employed”, “month”, “cons.conf.idx”, “emp.var.rate”, “euribor3m”, “previous”, “cons.price.idx”, “campaign”, “age”. Among those 11 variables, there are 3 variables of great importance, which are “pdays”, “poutcome”, and “nr.employed”. Figure 5. illustrates the variable importance for CART. Compared with “poutcome” and “nr.employed”, “pdays” has the greatest importance, and that corresponded to the Figure 4. tree structure graph. The tree structure graph shows that the “pdays” is the initial and essential split node, then it is followed by “nr. employed”. The optimal size of tree is with 4 terminal nodes and 3 splits, which is also shown in Figure 3. It is worth noticing that “poutcome” shows equally importance as “nr.employed”in Figure 5, but does not appear in the Figure 4 final tree structure. The reason might be in terms of importance, they all have potential to win. But the final tree structure is determined by purity, meaning that “poutcome” is less purer in classification compare with “nr.employed”.

|  |  |  |
| --- | --- | --- |
| Table 5. CART Accuracy | |  |
| CART | Training data | Test data |
| Accuracy | 82.99% | 82.51% |

Figure 3. Size of tree



Figure 4. Tree Structure





Figure 5. CART Variable Importance



4.3. Random Forest

The Random Forest works no matter the dependent variable is categorical or continuous. Unlike the decision tree performs as individual trivial tree, Random Forest forms a forest with many decision trees. When building decision tress on bootstrapped training samples, each time a split in a tree is considered, and a random sample of the predictors is chosen as split candidates from the full set of p predictors (---textbook reference later---). On average, each bagged tree makes use of around 2/3 of the observations, and the remaining 1/3 that is not used to fit a given bagged tree are the OOB (out-of-bag) observations. The OOB error is the classification and regression error for the test data, and it is the same as the error rate of the test data. The OOB error for individual tree is accumulated and averaged as a measure of prediction accuracy of test data. The Random Forest works well for prediction accuracy, and it improves by decorrelating individual decision trees.

The Figure 6. shows the error rate of Random Forest. The black line represents the overall averaged OOB error (1- test accuracy), and the red line is the prediction error rate for the y=0 (1-specificity), and the green line shows the prediction error rate for y=1 (1-sensitivity). This figure also indicates that the errors do not continue to decrease sharply after the number of trees in the forest reaches 40. Table 6. Random Forest Accuracy backed up my initial assumption that Random Forest always works well in terms of prediction accuracy. The accuracy for training data almost reaches 100%, with the accuracy for test data (87.39%) that is higher than other models.

The Random Forest selects its own variables, and it kept 19 out of 20 input variables, which were “pdays”, "cons.conf.idx", “nr.employed", "month", “previous”, "euribor3m", “cons.price.idx”, "emp.var.rate", “job”, “age”, “day\_of\_week”, “education”, “contact”, “default”, “poutcome”, “campaign”, “marital”, “housing”, “loan”. The right part of Figure 7 shows that if a variable is assigned values by random permutation, and how much the accuracy will decrease. On the other part of the graph, the node purity is measured by Gini index, which is the difference between RSS before and after the split on that variable. Since there is no fixed criterion of the best measurement of variable importance, and I identified the accuracy as the uniform performance measurement, thus I choose to refer to Mean Decrease Accuracy. The variable importance declines from the top “pdays” all the way down to “loan”. The “pdays” is such the most important feature that the model cannot bear losing it, because it will lose almost 100% accuracy. The “pdays” is the number of days that passed by after the client was last contacted from a previous campaign. The relationship makes sense because whether the customer are willing to subscribe the term deposit or not can be largely influenced by the frequency of bank’s direct marketing.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 6. RF Accuracy | | |  | |
| RF | Training data | | Test data | |
| Accuracy | 100.00% | | 87.39% | |

Figure 6. The Error Rate of Random Forest



Figure 7. Random Forest Variable Importance



4.4. ANN

ANN are composed of many nodes, and these nodes are connected to each other and function together, by passing information. They consist of a number of layers. Here, each layer performs a different function on the received data. Figure 8 shows the final structure of my Neural Networks with one input layer, one output layer, and one hidden layer which has 5 neurons. From the Table 7 and Table 8 below, we know that the best accuracy is obtained at 5 neurons with decay=0.1. As a result, the accuracy with training data is 74.86%, and with test data is 73.76%. The results shows the Neural Network does a bad job in predicting accuracy, compared with Random Forest and Classification and Regression Tree. Figure 9 suggests the most important variable selected by Neural Network is “nr.employed”, and its importance is way higher than others. It represents the number of employees of direct marketing. It makes sense that the more employees they used to make direct the phone calls, the more customers will be contacted, and the probability of subscription will increase.

|  |  |  |
| --- | --- | --- |
| Table 7. ANN Accuracy | |  |
| ANN | Training data | Test data |
| Accuracy | 74.86% | 73.76% |

Figure 8. ANN Structure

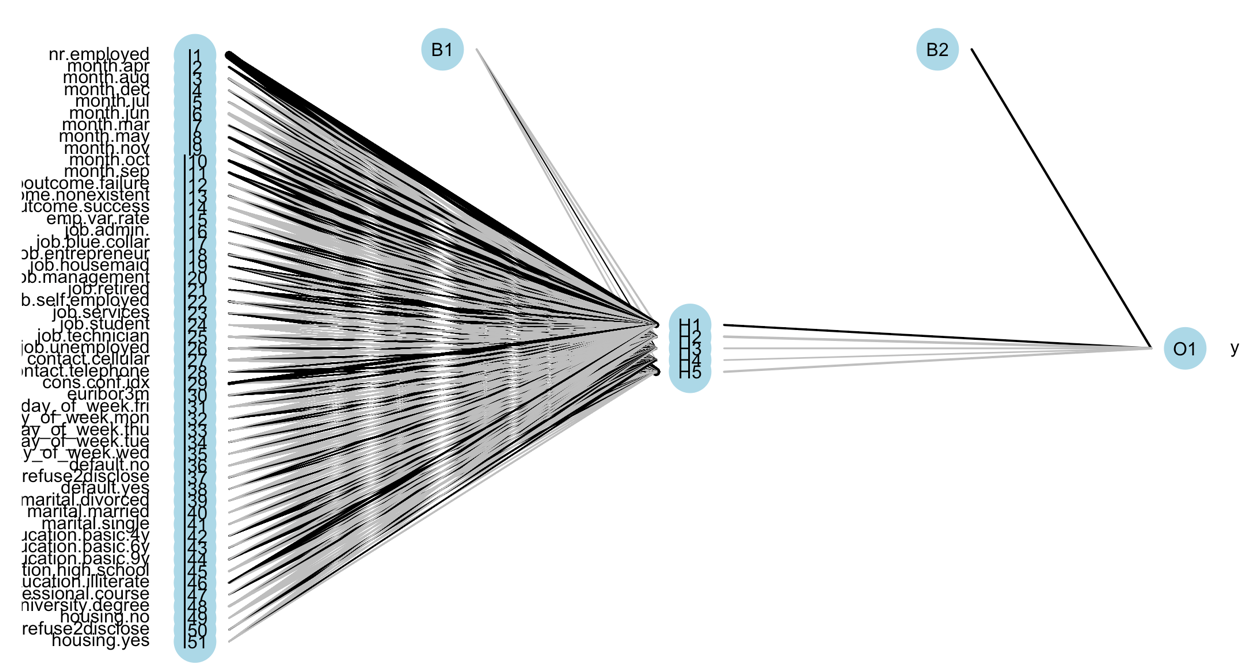


Table 8. Neural Network Result

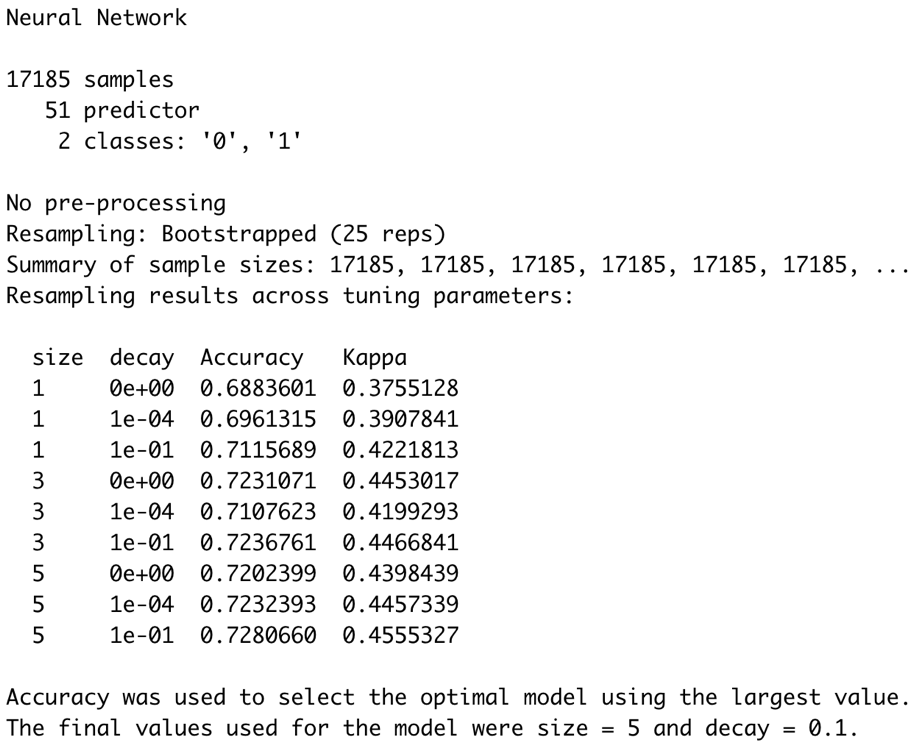




Figure 9. Variable Importance for ANN



**5. Result Analysis**

Figure 9. Accuracy Comparisons

Figure 9. illustrates the model accuracy comparisons, and it indicates that Random Forest stands out and performs the best no matter in training data or test data, it achieves 100% accuracy with training data and 87.39% accuracy with test dataset. Therefore, Random Forest has the best prediction accuracy compared with Logistic Regression, Classification and Regression Tree, and Artificial Neural Networks. Random Forest won the game because it is very stable. Even if a new data point is introduced in the dataset, the overall forest is not affected much since the new data may impact one tree, but it is very hard for it to impact all the trees in the forest. Besides, it has the ability to reduce overfitting and variance so as to improve the accuracy, and overfitting is a very common problem in other algorithms. It works well with both categorical and continuous variables, and automatically handle missing values, outliers, non-linear parameters, and no feature scaling is required. As for important variable, the “pdays” is such the most important feature that the model cannot bear losing it, because it will lose almost 100% accuracy. The “pdays” stands for the number of days that passed by after the client was last contacted from a previous campaign. It makes sense because whether the customer are willing to subscribe the term deposit can be largely influenced by the frequency of bank’s direct marketing. Besides, the second important variable “nr.employed” represents the number of employees of direct marketing. After all, the more employees they used to make direct the phone calls, the more customers will be contacted, and the probability of subscription will increase.

The Classification and Regression Tree ranks the second, and the third is Neural Networks. The logistic regression works the worst in prediction accuracy. The Logistic Regression lost the game maybe due to the parameter estimates are unstable. Besides, from the bar plot we know there are some categorical variables that have unbalanced distributions, thus it may perform worse than others. The reason of the Neural Network’s failure is unknown, and it is the most important problem. Because when it produces a probing solution, it does not give a clue as to why and how it works, thus it remains a blackspot. The Neural Networks requires data scaling and hence need to transform all the categorical variables into numerical variables, which increases its complexity of usage. As for Classification and Regression Tree, its failure might due to overfitting and unstableness, and a small variation in data may result in a completely different tree. Moreover, it is worth noticing that although it is not in our case, because I used SMOTE to balance the data, the tree become biased if some classes dominate in the unbalanced dataset.

It is interesting to see that for Logistic Regression, the first two important variables are “poutcomesucess” and “nr.employed”; for Classification and Regression Tree, the most important two variables are “pdays” and “nr.employed”; for Random Forest, they are “pdays” and “nr.employed”; for Neural Networks, “nr.employed” is much higher than others. We can see the commonly important variable for all 4 methods is “nr.employed”. The “pdays” ranks the first both in Random Forest and Classification and Regression Tree, and not in the rest maybe because the feature selection for Logistic Regression and Neural Networks does not include “pdays”, But the final result indeed indicates that “nr.employed” plays an essential part in affecting customers’ responses to bank direct marketing.

General Comments:

Please refer to my slides talking about how to present your research in paper and in presentation. Clarity is the top priority! Define variable before using them (or concurrently); define terms immediately when you introduce them; explain table/figure first before show them;…

What is the purpose of your project? I thought you have a well defined research question and you just need to show that the answers you deliver to the questions are solid and well grounded. I don’t think you need to show case all possible models that you can think of. You need to choose one to be “THE ONE” model that you will use and you can say that you tried others and here is why this one stands out.

You need to show the flow of thoughts in constructing your regression model. Starting with your research questions, then data, then how that leads to your model selection and the final econometrics model that you will estimate. You may identify the econometric issues in the model construction (e.g. variable endogeneity) and how you will deal with it. You will (very likely) encounter other econometric issues in executing the regression (e.g. homoskedasticity test, weak instrument test…) and you need to figure out how to deal with those issues, which will be part of your next assignment.