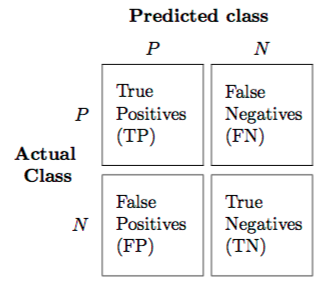
**4. Methodology**

To predict customers’ responses to bank direct marketing term deposit subscription, four algorithms are used, which are the Logistic model (LR), Classification and Regression Tree (CART), Random Forest (RF), and Artificial Neural Network (ANN). The balanced data is randomly divided into training data (50% of the balanced data), and the remaining data serves as the test dataset. 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 four elements 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.

Table-----. 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>

To identify casual effect, I plan to use GLM, Conditional Inference Tree (CIT), and Instrumental Variable (IV). I will only touch a little bit about the IV, since even though we know the criteria for choosing a good IV, I do not have the IV data to test on, thus I will just keep it hypothetical.

**4.1. LR**

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. LR 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 confusion matrix results of LR for training and test data are listed below, and the Table----. LR Accuracy shows that LR did a not-so- good job in prediction accuracy, and it only obtained 71.38% prediction accuracy in validation set. The reason why the LR made a not-so-good prediction in our case might due to the classes of the response variable y or not were well-separated, and the parameter estimates for the LR are unstable. Besides, the LR cannot handle non-linear or interactive effects of feature variables. Whereas the LR also has its advantage, since it produces straight forward probabilistic classification formula.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table--Confusion matrix of LR with training data** | | | | |
|  |  | Predicted subscription status | | |
|  |  | No | Yes | Total |
| True subscription status | No | 6974 | 3265 | 10239 |
| Yes | 1649 | 5290 | 6939 |
|  | Total | 8623 | 8555 | 34356 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table--Confusion matrix of LR with test data** | | | | |
|  |  | Predicted subscription status | | |
|  |  | No | Yes | Total |
| True subscription status | No |  |  |  |
| Yes |  |  |  |
|  | Total |  |  |  |

Note: there are some problem with R code for logistic model test data confusion matrix.----Fix next time!!

|  |  |  |
| --- | --- | --- |
| **Table---. LR Accuracy** | |  |
| **LR** | Training data | Test data |
| Accuracy | 71.39% | 71.38% |

**4.2. CART**

(Do we have to introduce the algorithm? -) (+Advantages &disadvantages of CART)------------------------------The tree before pruning and after are exact the same, with the result of 83.76% training accuracy and 83.46% test accuracy. It resulted tree has 6 terminal nodes.

****

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table--Confusion matrix of CART with training data** | | | | |
|  |  | Predicted subscription status | | |
|  |  | No | Yes | Total |
| True subscription status | No | 7683 | 940 | 8623 |
| Yes | 1850 | 6705 | 8555 |
|  | Total | 9533 | 7645 | 17178 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Table--Confusion matrix of CART with test data** | | | | | | | |
|  | |  | | Predicted subscription status | | | |
|  | |  | | No | | Yes | Total |
| True subscription status | | No | | 7642 | | 980 | 8622 |
| Yes | | 1861 | | 6693 | 8554 |
|  | | Total | | 9503 | | 7673 | 17176 |
| **Table---. CART Accuracy** | | |  | |
| **CART** | Training data | | Test data | |
| Accuracy | 83.76% | | 83.46% | |

**4.3. RF**

The RF works no matter the dependent variable is categorical or continuous. Unlike the decision tree performs as individual trivial tree, RF forms a forest with may 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---). The out-of-bag (OOB) error for individual tree is accumulated and averaged as a measure of prediction accuracy of test data. The RF works well for prediction accuracy, and it improves by decorrelating individual decision trees. There is no need to prune the tree, and even when the size of data is large and has so many independent variables, RF performs stably without overfitting.

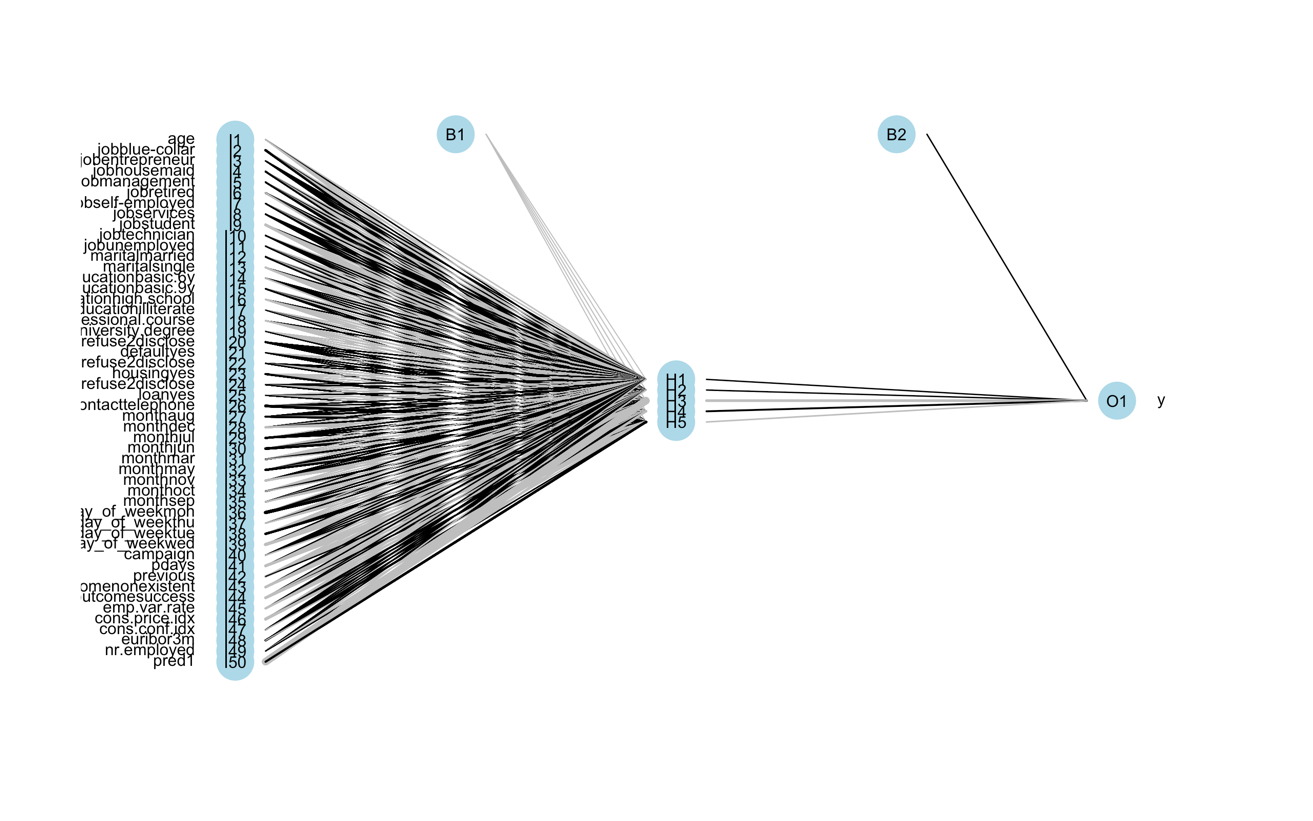
The Figure---. shows the error rate of RF. The black line represents the overall averaged OOB error (1- test accuracy), and the red line is the prediction error rate for the y=0 (1-specificity), and the green line shows the prediction error rate for y=1 (1-sensitivity). This figure also indicates that the errors do not continue to decrease sharply after the number of trees in the forest reaches 40. The confusion matrix(Table---&Table--) and Table--. RF Accuracy backed up my initial assumption that RF always works well in terms of prediction accuracy. The accuracy for training data almost reaches 100%, with the accuracy for test data (86.41%) that is higher than other models.

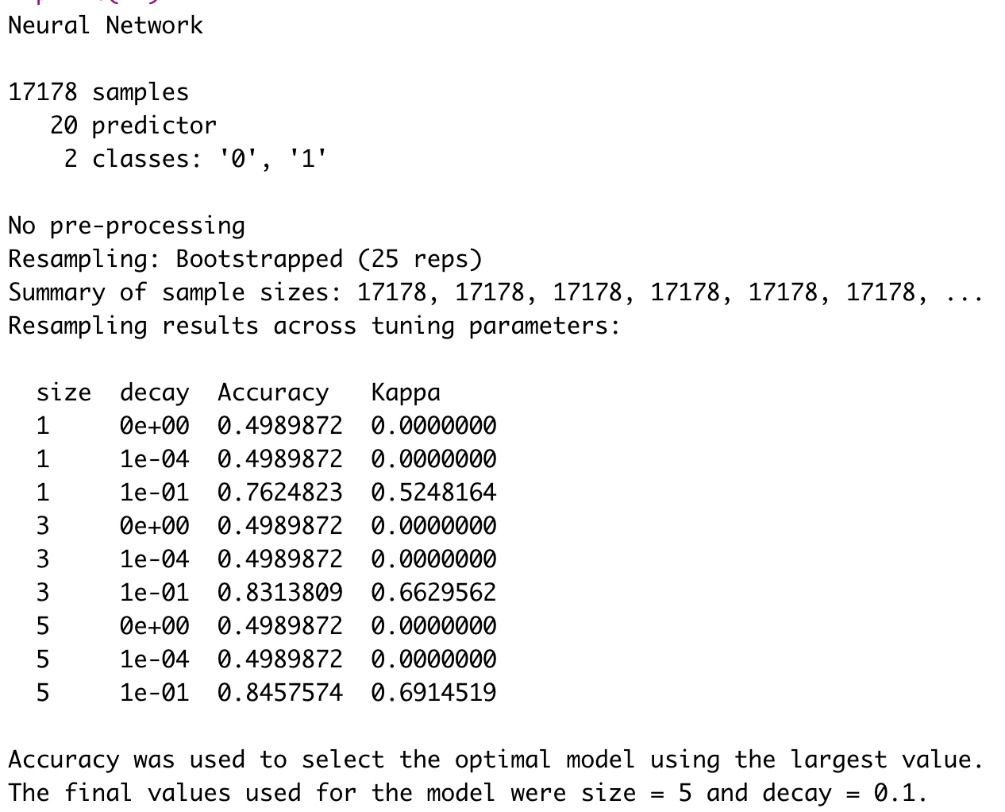
****

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Table--Confusion matrix of RF with training data** | | | | | | | |
|  | |  | | Predicted subscription status | | | |
|  | |  | | No | | Yes | Total |
| True subscription status | | No | | 8623 | | 0 | 8623 |
| Yes | | 2 | | 8553 | 8555 |
|  | | Total | | 8625 | | 8553 | 17178 |
| **Table--Confusion matrix of RF with training data** | | | | | | | |
|  | |  | | Predicted subscription status | | | |
|  | |  | | No | | Yes | Total |
| True subscription status | | No | | 7591 | | 1031 | 8622 |
| Yes | | 1302 | | 7252 | 8554 |
|  | | Total | | 8893 | | 8283 | 17176 |
| **Table---. RF Accuracy** | | |  | |
| **RF** | Training data | | Test data | |
| Accuracy | 99.98% | | 86.41% | |

**4.4. ANN**

The ANN has large capacity to deal with large size data, and it is suitable to estimate non-linear relationship. Besides, it prevents the model from overfitting and outliers. From the Table and Figure------ 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 84.57%, and with test data is 84.09%. The results shows the ANN does a good job in predicting accuracy, even though it is still lower than RF.

****

****

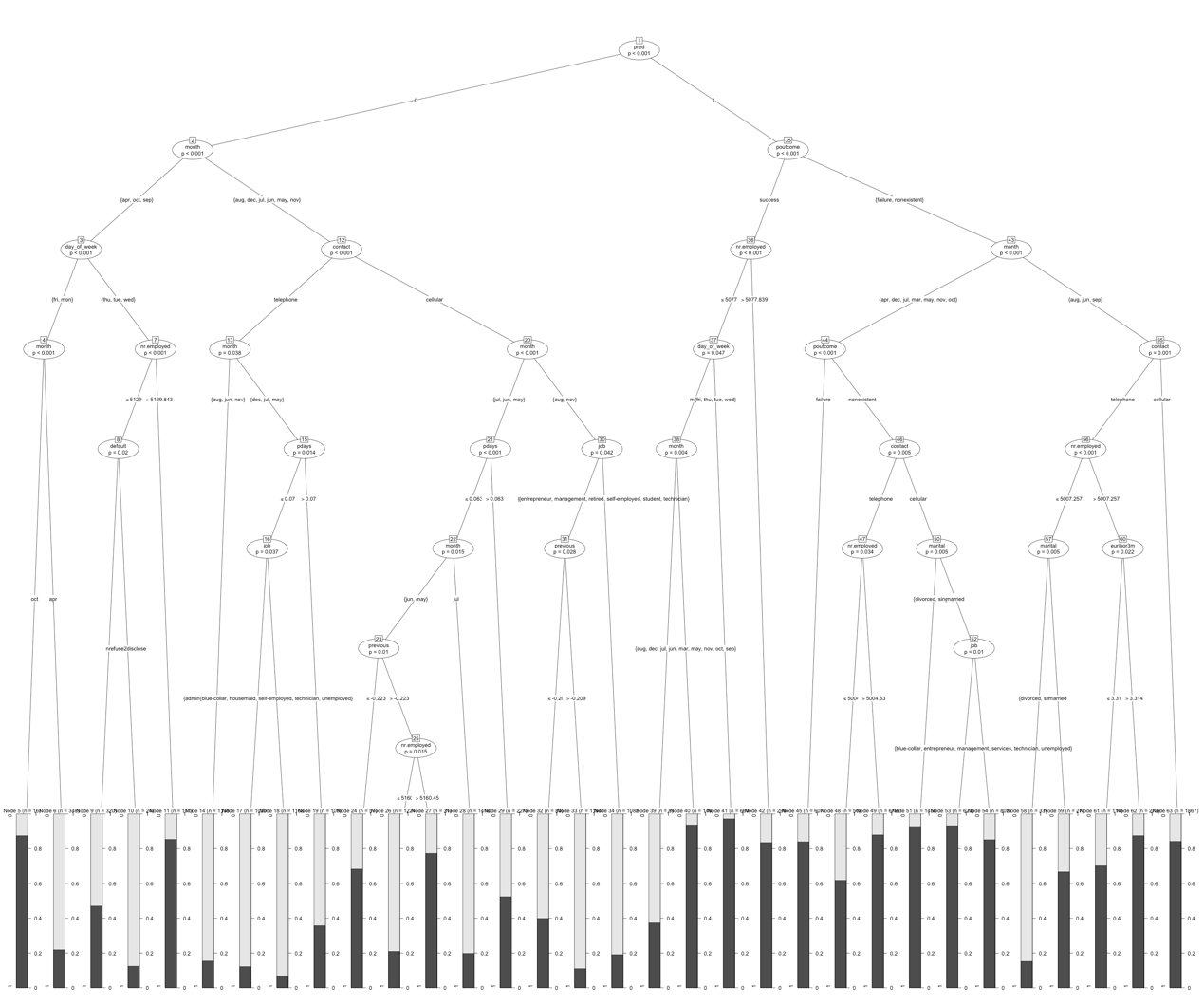
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Table--Confusion matrix of ANN with training data** | | | | | | | |
|  | |  | | Predicted subscription status | | | |
|  | |  | | No | | Yes | Total |
| True subscription status | | No | | 7586 | | 1614 | 9200 |
| Yes | | 1031 | | 6941 | 7972 |
|  | | Total | | 8617 | | 8555 | 17172 |
| **Table--Confusion matrix of ANN with training data** | | | | | | | |
|  | |  | | Predicted subscription status | | | |
|  | |  | | No | | Yes | Total |
| True subscription status | | No | | 7513 | | 1624 | 9137 |
| Yes | | 1109 | | 6930 | 8039 |
|  | | Total | | 8622 | | 8554 | 17176 |
| **Table---. ANN Accuracy** | | |  | |
| **ANN** | Training data | | Test data | |
| Accuracy | 84.57% | | 84.09% | |

The Figure---. illustrates the model accuracy comparisons, and it indicates that Random Forest performs best no matter in training data or test data. Therefore, Random Forest has the best prediction accuracy compared with GLM, CART, and ANN. The ANN ranks the second, and the third is CART. The logistic regression works the worst in prediction accuracy. (Forget to include the casual inference tree prediction results-----)

**4.5. Casual inference**

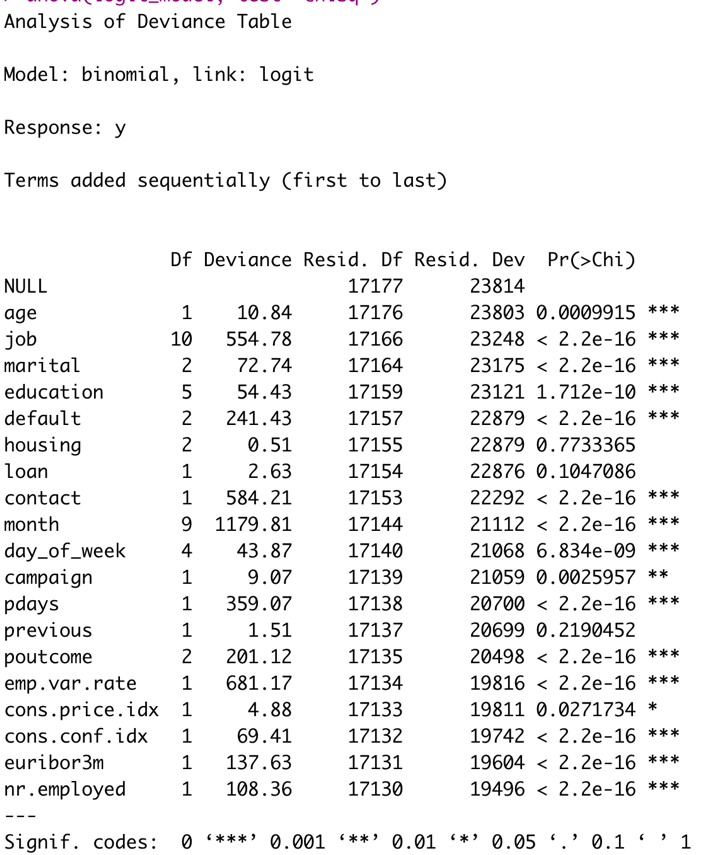
**4.5.1. CIT**

CIT solves the problems of CART, including overfitting, selection bias towards the covariates, interpretation difficulties, no concept of statistical significance in terms of variable selection. CIT does unbiased variable selection, and it structured differently from trees partitioned with exhaustive search procedures (Needs further knowledge and investigation------Maybe ask Shledon).

****

**4.5.2 Logistic regression(Needs further investigation for significance test Wald?)**

In order to analyze casual inference, the test of significance of each covariate is in real need, and thus I conducted a likelihood ratio test. The test uses chi-squared distribution to calculate p-values, and adds one predictor at a time and compares nested models with increasing complexity against the full model sequentially. The Table---. below shows that age, job, marital, education, default, contact, month, day\_of\_week, pdays, poutcome, emp.var.rate, cons.conf.idx, euribor3m, nr.employed are significant. We should consider

****

**4.5.3. Deep IV(Not familiar with this concept, prepare to ask Sheldon)**

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.