# CHARACTER RECOGNITION WITH TREE-BASED ALGORITHM

**By: HOANG UYEN LE**

---

## 1. Build random forest algorithm from "scratch" with MNIST.

Following is the approach I chased to write the script for building random forest algorithm from "scratch"

1. Create a training dataset for each trees created:

- Draw a random number between 392 and 784 (called N), which will be the size of training dataset.

- For each row in the original dataset (excluding label column), draw N cells and add them to the training dataframe. Every row can share the same feature. After that, adding label column to this dataframe.

- After all the rows are added to the traning dataframe, we have a completed traning set with

(N variables + 1 label) x 60,000 observations.

2. Use training dataset created above to build a classification model. I used randomForest in caret library to grow this tree, with default parameter with number of tree is 1 and do.trace =1

3. Apply this random forest tree to predict label for testing dataframe (without label column)

4. Loop step 1 – 2 - 3 for 1 times, I have 1 random forest tree and 1 predicted label set for testing dataframe.

5. Loop step 1 – 2 - 3 for 10 times, I have 10 random forest trees and 10 predicted label sets for testing dataframe. Selecting the appropriate predicted label for each image by majority voting rule.

6. Loop step 1 – 2 - 3 for 500 times, I have 500 random forest trees and 500 predicted label sets for testing dataframe. Selecting the appropriate predicted label for each image by majority voting rule.

7. Produce the confusionmatrix to compare the matching rate between actual lablesl and predicted labels for each case: 1 tree, 10 trees, 500 trees. Combining the confusion matrix and accuracy obtained with logistic regression to have a comprehensive comparison.

| Confusion matrix of Logistic regression – Accuracy rate: 81.9% | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Prediction | | | | | | | | | |
| | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| **Actuals** | **0** | **964** | 0 | 0 | 2 | 0 | 0 | 6 | 1 | 7 | 0 |
| | **1** | 0 | **921** | 1 | 5 | 0 | 0 | 6 | 1 | 200 | 1 |
| | **2** | 12 | 3 | **813** | 25 | 2 | 0 | 15 | 2 | 155 | 5 |
| | **3** | 6 | 0 | 6 | **902** | 1 | 0 | 6 | 3 | 82 | 4 |
| | **4** | 6 | 0 | 3 | 5 | **793** | 0 | 12 | 0 | 103 | 59 |
| | **5** | 34 | 1 | 1 | 90 | 5 | **288** | 19 | 4 | 434 | 16 |
| | **6** | 16 | 1 | 3 | 1 | 3 | 2 | **894** | 0 | 38 | 0 |
| | **7** | 11 | 2 | 16 | 15 | 5 | 0 | 1 | **805** | 75 | 97 |
| | **8** | 7 | 1 | 0 | 10 | 4 | 0 | 3 | 1 | **948** | 0 |
| | **9** | 11 | 2 | 0 | 10 | 5 | 0 | 0 | 1 | 119 | **861** |

| Confusion matrix of 1 random forest tree – Accuracy rate: 82.85% | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Prediction | | | | | | | | | |
| | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| **Actuals** | **0** | **884** | 0 | 2 | 10 | 9 | 20 | 14 | 11 | 17 | 13 |
| | **1** | 1 | **1083** | 9 | 5 | 3 | 1 | 6 | 11 | 13 | 3 |
| | **2** | 16 | 9 | 826 | 45 | 17 | 16 | 21 | 39 | 29 | 14 |
| | **3** | 15 | 4 | 32 | **785** | 7 | 68 | 9 | 17 | 39 | 34 |
| | **4** | 10 | 5 | 8 | 7 | **815** | 11 | 27 | 16 | 19 | 64 |
| | **5** | 23 | 6 | 9 | 55 | 14 | **676** | 27 | 12 | 38 | 32 |
| | **6** | 20 | 5 | 12 | 9 | 28 | 34 | **815** | 5 | 23 | 7 |
| | **7** | 3 | 14 | 38 | 18 | 13 | 12 | 2 | **884** | 14 | 29 |
| | **8** | 22 | 8 | 37 | 57 | 29 | 34 | 19 | 15 | **714** | 39 |
| | **9** | 14 | 5 | 16 | 21 | 49 | 19 | 6 | 41 | 36 | **802** |

| Confusion matrix of 10 random forest trees – Accuracy rate: 94.91% | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Prediction | | | | | | | | | |
| | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| Actuals | **0** | **971** | 1 | 0 | 1 | 0 | 3 | 2 | 1 | 1 | 0 |
| | **1** | 0 | **1128** | 2 | 2 | 0 | 1 | 2 | 0 | 0 | 0 |
| | **2** | 13 | 5 | **990** | 3 | 2 | 0 | 3 | 9 | 6 | 2 |
| | **3** | 5 | 2 | 16 | **951** | 1 | 15 | 0 | 8 | 8 | 4 |
| | **4** | 1 | 3 | 6 | 1 | **932** | 1 | 2 | 4 | 4 | 28 |
| | **5** | 7 | 5 | 4 | 29 | 7 | **816** | 10 | 1 | 8 | 5 |
| | **6** | 10 | 4 | 1 | 4 | 7 | 7 | **921** | 0 | 4 | 0 |
| | **7** | 2 | 11 | 27 | 1 | 6 | 2 | 0 | **963** | 5 | 10 |
| | **8** | 8 | 6 | 14 | 23 | 5 | 9 | 9 | 3 | **887** | 10 |
| | **9** | 9 | 6 | 5 | 16 | 21 | 3 | 3 | 7 | 8 | **931** |

| Confusion matrix of 500 random forest trees – Accuracy rate: 97.22% | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Prediction | | | | | | | | | |
| | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| Actuals | **0** | **972** | 1 | 0 | 0 | 0 | 3 | 1 | 1 | 1 | 1 |
| | **1** | 0 | **1125** | 2 | 2 | 0 | 2 | 2 | 0 | 1 | 1 |
| | **2** | 6 | 0 | **1002** | 4 | 3 | 0 | 4 | 8 | 5 | 0 |
| | **3** | 0 | 0 | 9 | **976** | 0 | 7 | 0 | 9 | 7 | 2 |
| | **4** | 1 | 0 | 2 | 0 | **957** | 0 | 6 | 0 | 2 | 14 |
| | **5** | 3 | 0 | 1 | 10 | 3 | **861** | 6 | 2 | 5 | 1 |
| | **6** | 6 | 3 | 0 | 0 | 4 | 3 | **938** | 0 | 4 | 0 |
| | **7** | 1 | 4 | 18 | 2 | 0 | 0 | 0 | **988** | 4 | 10 |
| | **8** | 3 | 0 | 2 | 8 | 3 | 5 | 1 | 4 | **938** | 10 |
| | **9** | 5 | 4 | 1 | 9 | 11 | 2 | 1 | 4 | 8 | **964** |

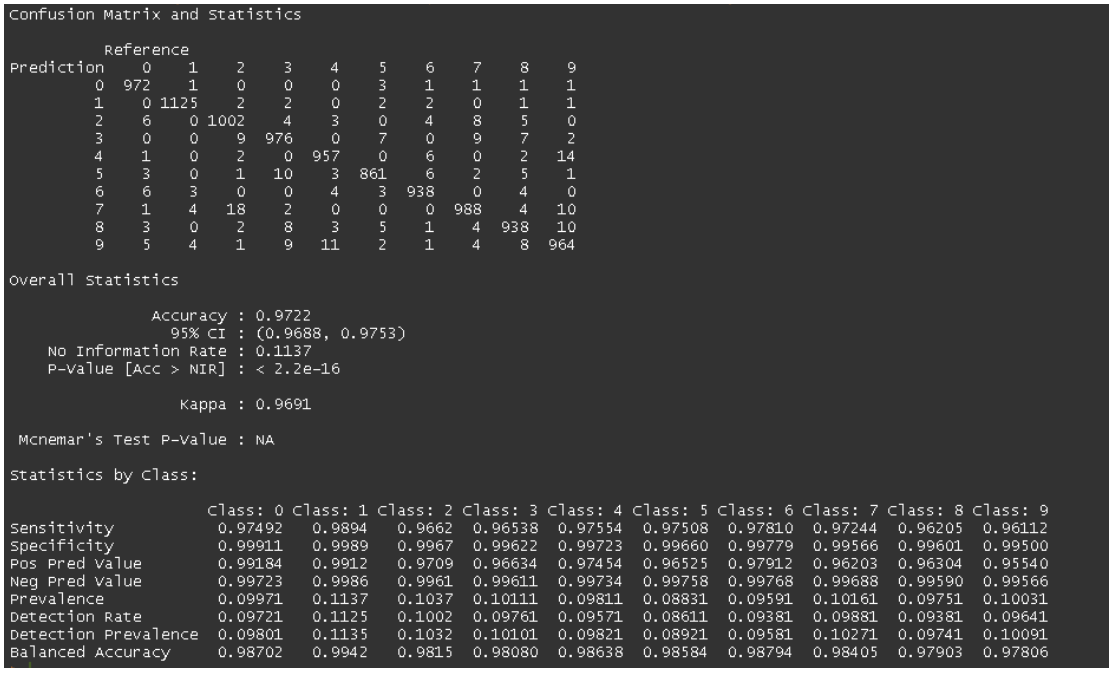
Overall, the more tree models I built, the higher accuracy I obtained. While in logistics regression algorithm, all the digits were highly wrong-predicted to digit 8, with some significant failure such as 454 times of digit 5, or 200 times of digit 1, this situation was improved when using ranfom forest algorithm. After looping random forest for 500 times, the accuracy is approximately 97%, with the sensitivity and balanced accuracy of each digit keep increasing.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1    2    3    4    5    6    7    8    9
         0  884    0    2   10    9   20   14   11   17   13
         1    1 1083    9    5    3    1    6   11   13    3
         2   16    9  826   45   17   16   21   39   29   14
         3   15    4   32  785    7   68    9   17   39   34
         4   10    5    8    7  815   11   27   16   19   64
         5   23    6    9   55   14  676   27   12   38   32
         6   20    5   12    9   28   34  815    5   23    7
         7    3   14   38   18   13   12    2  884   14   29
         8   22    8   37   57   29   34   19   15  714   39
         9   14    5   16   21   49   19    6   41   36  802

Overall Statistics

               Accuracy : 0.8285
                 95% CI : (0.8209, 0.8358)
    No Information Rate : 0.1139
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.8093

 Mcnemar's Test P-Value : 0.09765

Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6 Class: 7 Class: 8 Class: 9
Sensitivity           0.87698   0.9508  0.83519  0.77569  0.82825  0.75870  0.86152  0.84110  0.75796  0.77338
Specificity           0.98932   0.9941  0.97714  0.97496  0.98148  0.97628  0.98420  0.98402  0.97129  0.97690
Pos Pred Value        0.90204   0.9542  0.80039  0.77723  0.82994  0.75785  0.85073  0.86076  0.73306  0.79485
Neg Pred Value        0.98625   0.9937  0.98182  0.97475  0.98126  0.97639  0.98551  0.98139  0.97474  0.97386
Prevalence            0.10081   0.1139  0.09891  0.10121  0.09841  0.08911  0.09461  0.10511  0.09421  0.10371
Detection Rate        0.08841   0.1083  0.08261  0.07851  0.08151  0.06761  0.08151  0.08841  0.07141  0.08021
Detection Prevalence  0.09801   0.1135  0.10321  0.10101  0.09821  0.08921  0.09581  0.10271  0.09741  0.10091
Balanced Accuracy     0.93315   0.9725  0.90616  0.87533  0.90486  0.86749  0.92286  0.91256  0.86463  0.87514
```

*Figure 1: Confusion matrix of 1 random forest tree*

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1    2    3    4    5    6    7    8    9
         0  971    1    0    1    0    3    2    1    1    0
         1    0 1128    2    2    0    1    2    0    0    0
         2   13    5  990    3    2    0    2    9    6    2
         3    5    2   16  951    1   15    0    8    8    4
         4    1    3    6    1  932    1    2    4    4   28
         5    7    5    4   29    7  816   10    1    8    5
         6   10    4    1    4    7    7  921    0    4    0
         7    2   11   27    1    6    2    0  963    5   10
         8    8    6   14   23    5    9    9    3  887   10
         9    9    6    5   16   21    3    3    7    8  931

Overall Statistics

               Accuracy : 0.9491
                 95% CI : (0.9446, 0.9533)
    No Information Rate : 0.1171
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9434

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6 Class: 7 Class: 8 Class: 9
Sensitivity           0.94639   0.9633  0.92958  0.92241  0.95005  0.95216  0.96845  0.96687  0.95274  0.94040
Specificity           0.99900   0.9992  0.99530  0.99342  0.99446  0.99169  0.99591  0.99289  0.99041  0.99134
Pos Pred Value        0.99082   0.9938  0.95930  0.94158  0.94908  0.91480  0.96138  0.93768  0.91068  0.92270
Neg Pred Value        0.99390   0.9951  0.99164  0.99110  0.99457  0.99550  0.99668  0.99632  0.99512  0.99344
Prevalence            0.10261   0.1171  0.10651  0.10311  0.09811  0.08571  0.09511  0.09961  0.09311  0.09901
Detection Rate        0.09711   0.1128  0.09901  0.09511  0.09321  0.08161  0.09211  0.09631  0.08871  0.09311
Detection Prevalence  0.09801   0.1135  0.10321  0.10101  0.09821  0.08921  0.09581  0.10271  0.09741  0.10091
Balanced Accuracy     0.97270   0.9812  0.96244  0.95791  0.97225  0.97192  0.98218  0.97988  0.97157  0.96587
```

*Figure 2: Confusion matrix of 10 random forest tree*



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1    2    3    4    5    6    7    8    9
         0  972    1    0    0    0    3    1    1    1    1
         1    0 1125    2    2    0    2    2    0    1    1
         2    6    0 1002    4    3    0    4    8    5    0
         3    0    0    9  976    0    7    0    9    7    2
         4    1    0    2    0  957    0    6    0    2   14
         5    3    0    1   10    3  861    6    2    5    1
         6    6    3    0    0    4    3  938    0    4    0
         7    1    4   18    2    0    0    0  988    4   10
         8    3    0    2    8    3    5    1    4  938   10
         9    5    4    1    9   11    2    1    4    8  964

Overall Statistics

               Accuracy : 0.9722
                 95% CI : (0.9688, 0.9753)
    No Information Rate : 0.1137
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9691

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6 Class: 7 Class: 8 Class: 9
Sensitivity           0.97492   0.9894   0.9662  0.96538  0.97554  0.97508  0.97810  0.97244  0.96205  0.96112
Specificity           0.99911   0.9989   0.9967  0.99622  0.99723  0.99660  0.99779  0.99566  0.99601  0.99500
Pos Pred Value        0.99184   0.9912   0.9709  0.96634  0.97454  0.96525  0.97912  0.96203  0.96304  0.95540
Neg Pred Value        0.99723   0.9986   0.9961  0.99611  0.99734  0.99758  0.99768  0.99688  0.99590  0.99566
Prevalence            0.09971   0.1137   0.1037  0.10111  0.09811  0.08831  0.09591  0.10161  0.09751  0.10031
Detection Rate        0.09721   0.1125   0.1002  0.09761  0.09571  0.08611  0.09381  0.09881  0.09381  0.09641
Detection Prevalence  0.09801   0.1135   0.1032  0.10101  0.09821  0.08921  0.09581  0.10271  0.09741  0.10091
Balanced Accuracy     0.98702   0.9942   0.9815  0.98080  0.98638  0.98584  0.98794  0.98405  0.97903  0.97806
```

*Figure 3: Confusion matrix of 500 random forest tree*