**Discuss the model result**

**Group members**

Nattapat Tangniyom 640631032

Sonram Sirirat 640631037

Thanakorn Chaisen 640631120

|  |  |  |
| --- | --- | --- |
| Train and Test Ratio | X | Y |
| Train Shape | 22571 | 22571 |
| Test Shape | 5643 | 5643 |

In modeling, we split the data in 2 parts which are 80% of training and 20% of testing.

**Accuracy rate**

|  |  |
| --- | --- |
| k-NN (k) | Accuracy (%) |
| 1 | 0.72 |
| 2 | 0.80 |
| 3 | 0.76 |
| 4 | 0.80 |
| 5 | 0.79 |

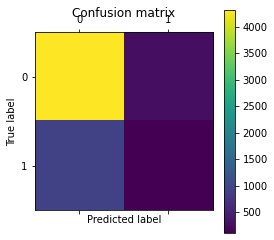
When we set k = 5, although some of them have the same accuracy, the others results (ex. Precision, recall, f1-score) may be different.

|  |  |
| --- | --- |
| Cross Validation (CV) | Accuracy (%) |
| 1 | 0.78671096 |
| 2 | 0.77868852 |
| 3 | 0.77691626 |
| 4 | 0.77536553 |
| 5 | 0.78090385 |

When we set CV = 5, Average of accuracy score = 0.7797170265412736

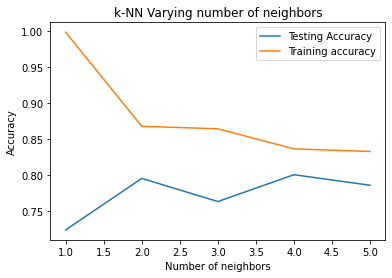
**Confusion matrix**

From test data 20%



|  |  |  |  |
| --- | --- | --- | --- |
| True Positive | 4320 | False Positive | 277 |
| True Negative | 933 | False Negative | 133 |

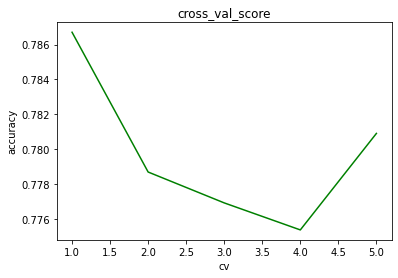
**Train vs test**



The graph shows the accuracy of each k-NN operation.

In k-NN of this dataset. We test the k-NN model by setting k = 1 to k = 5. Then find which k of this model returns the best accuracy.

**Cross validation**



The graph shows the accuracy of each cross validation operation.

In Cross validation of this dataset. We separate data from training in 5 parts and test the accuracy of each part by rotation estimation to find the performance of the mode which is the mean of accuracy.

**Optimal k**

From the k-NN model, when set k = 2 and k = 4 they return the same accuracy (0.80). But k = 4 have training and testing accuracy more close together than k = 2 from the graph accuracy of each k-NN operation. Therefore, the optimal k = 4.