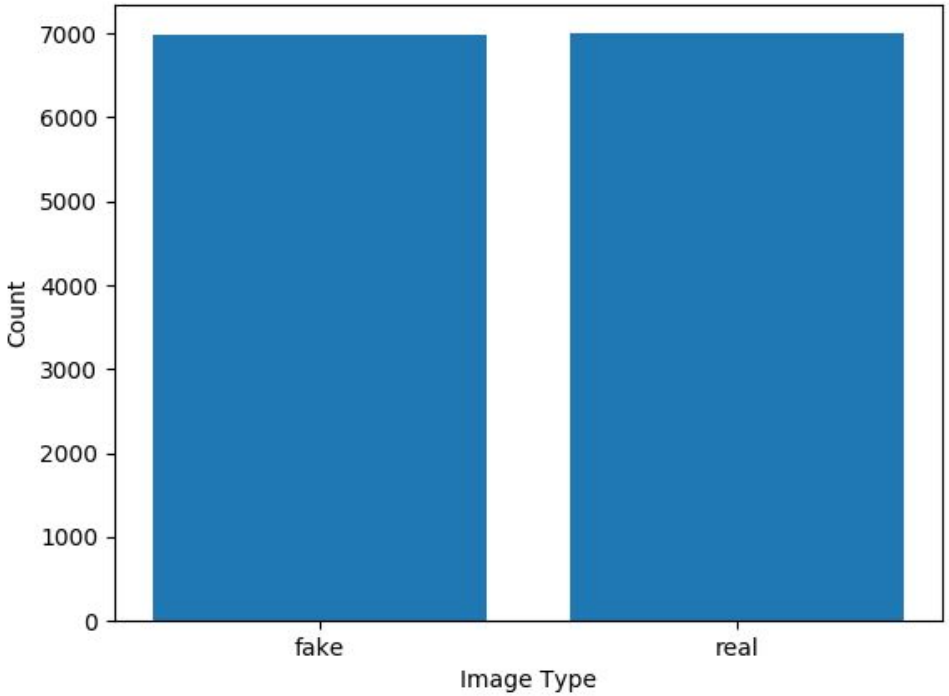
WK 10 Report

Machine Learning Tasks

Yunting Chiu

Data Visualization



- **Variables: 307200 of X + 1 of y**
- **Observations: 13984**
- **Supervised Learning**

```
[[array([134, 131, 116, ...,  68,  60,  71], dtype=uint8) 'fake']
 [array([133, 130, 115, ...,  71,  59,  71], dtype=uint8) 'fake']
 [array([117, 113, 112, ...,  43,  31,  45], dtype=uint8) 'fake']
 ...
 [array([ 33,  20,  66, ..., 188, 155, 172], dtype=uint8) 'real']
 [array([ 51,  28,  46, ..., 116,  52,  50], dtype=uint8) 'real']
 [array([174, 140, 102, ...,  23,  39,  98], dtype=uint8) 'real']]
```

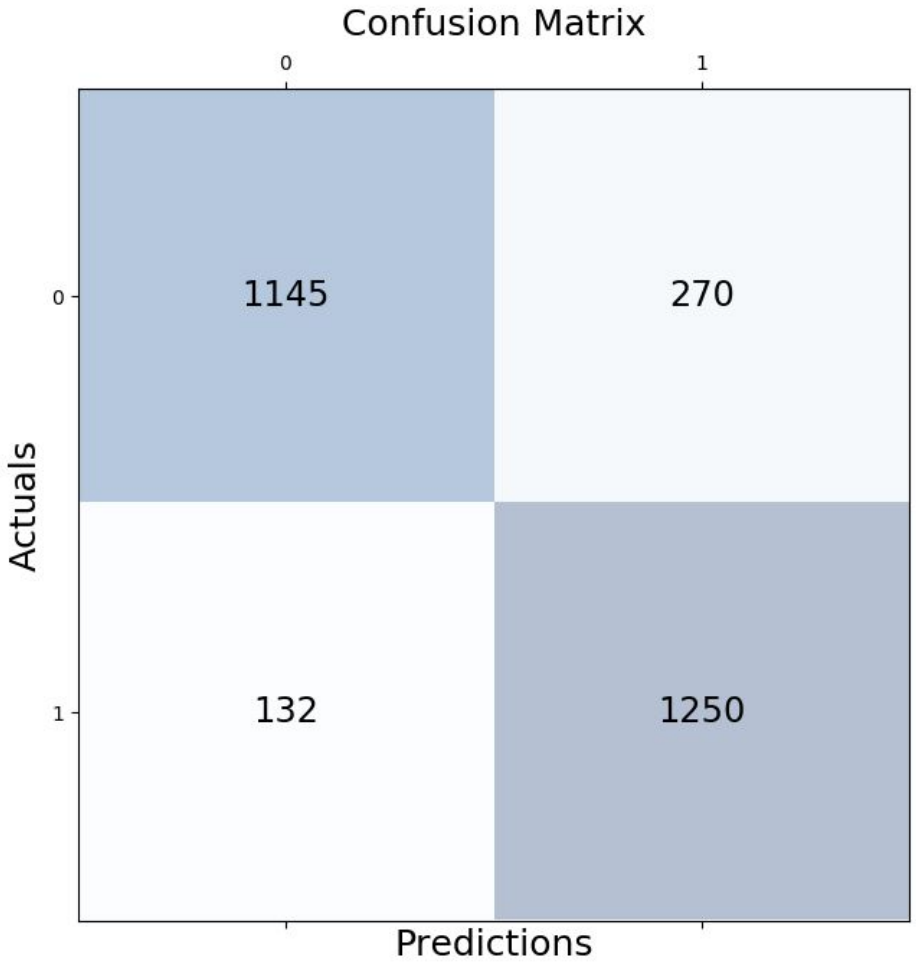
Random Forest (best)

80%train/20%test

Running time: 576 sec

Avg memory: 15481 MB

Accuracy: 85.6%



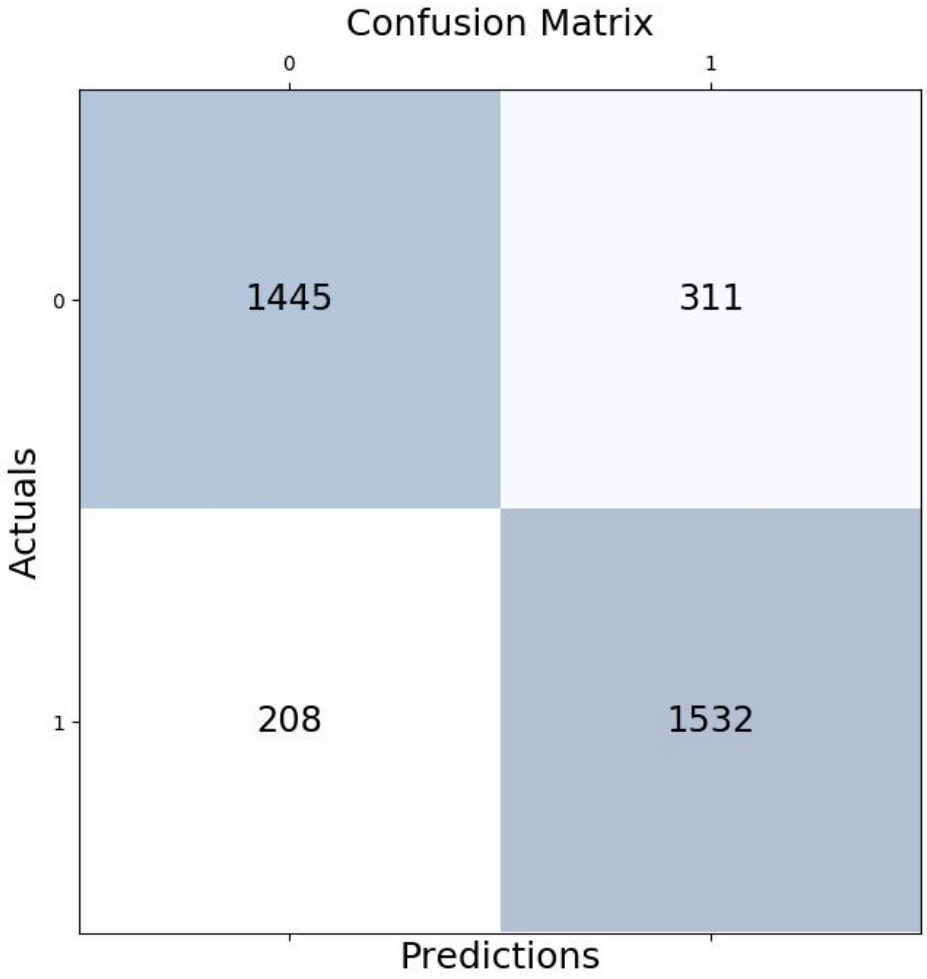
Random Forest

75%train/25%test

Running time: 570 sec

Avg memory: 14254 MB

Accuracy: 85.2%



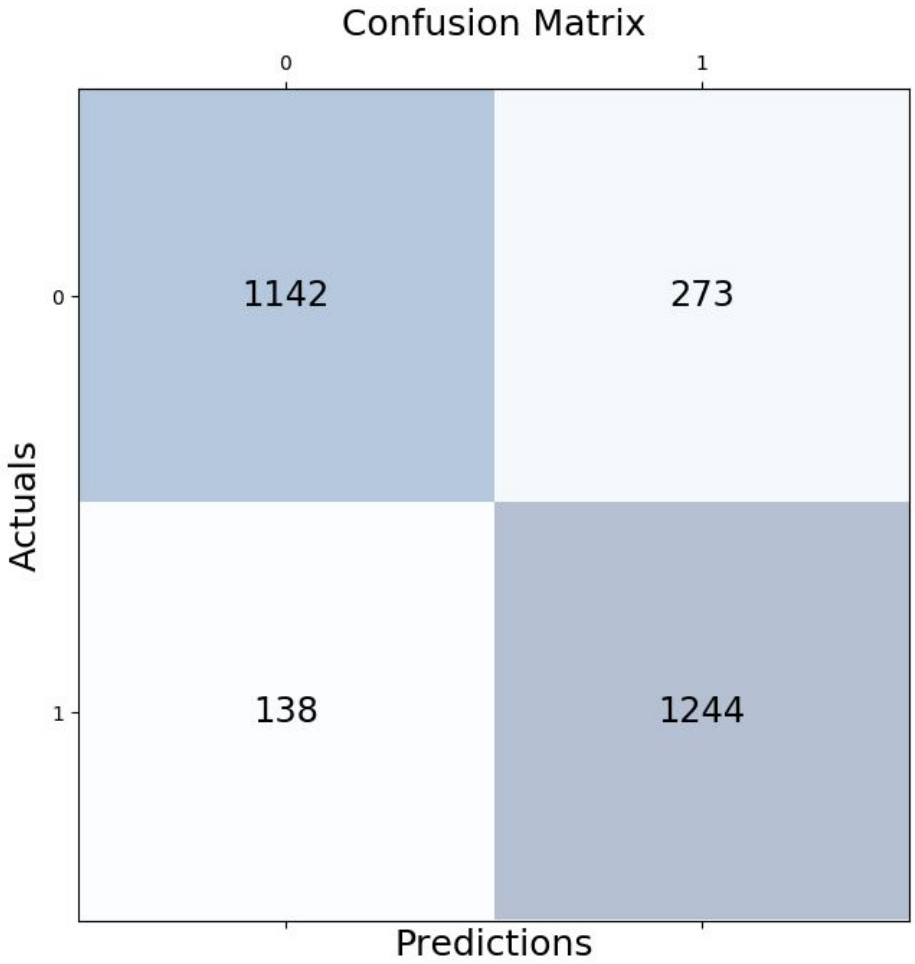
SVM

80%train/20%test, C = 1

Running time: 12.1 hrs

Avg memory: 30267 MB

Accuracy: 85.3%



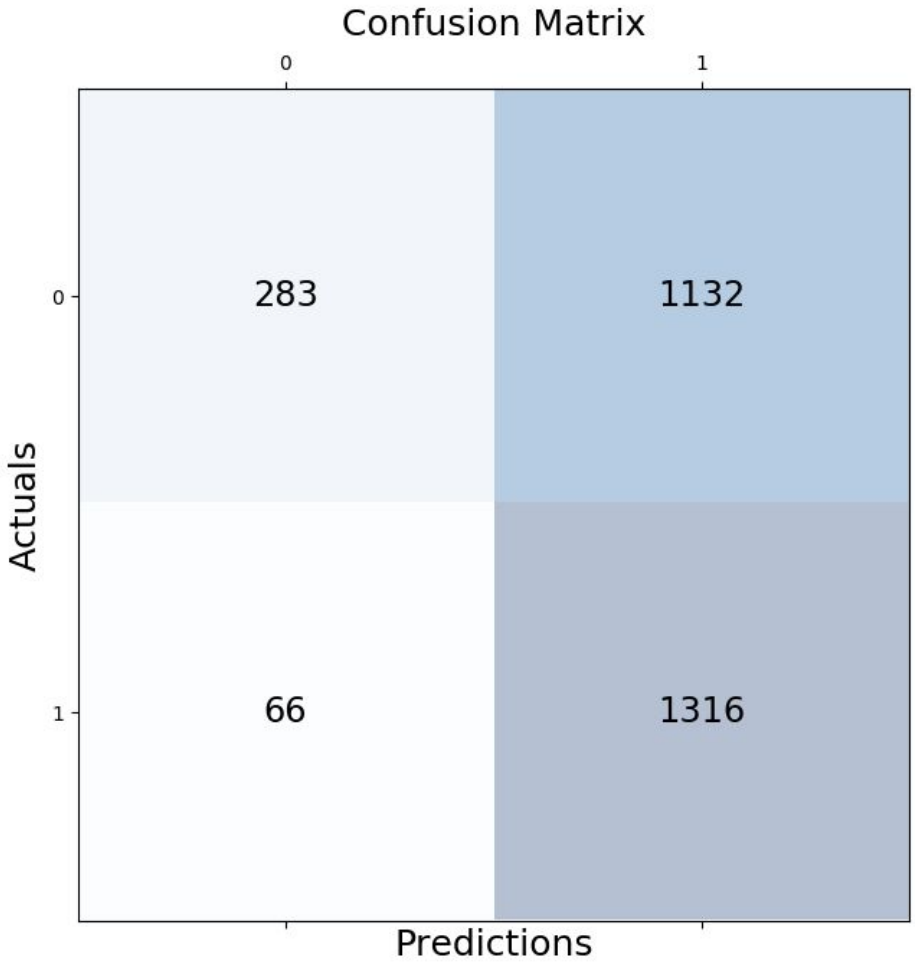
SVM

80%train/20%test, C = 0.01

Running time: 16.3 hrs

Avg memory: 31843 MB

Accuracy: 57.2%



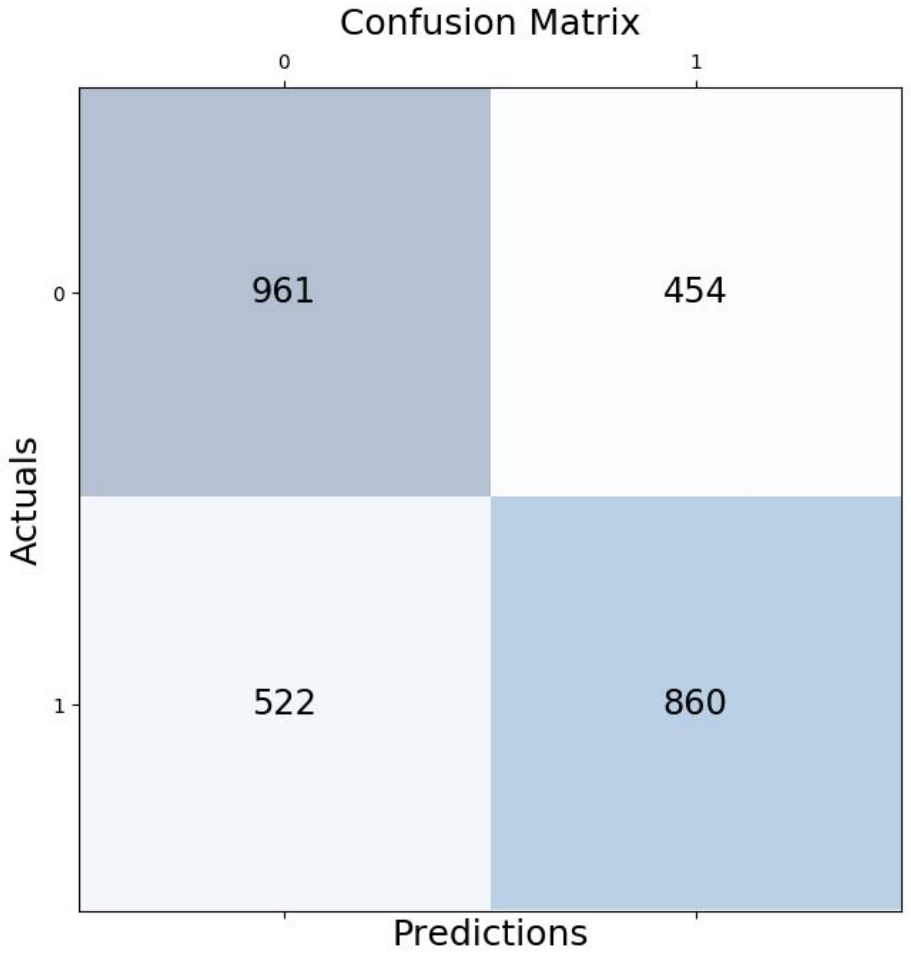
Logistic Regression

80%train/20%test

Running time: almost 1 day

Avg memory: 82565 MB

Accuracy: 65.1%



Next Steps

- improve the Random Forest model accuracy
- nested Cross-Validation
- find the optimal hyperparameter (num of trees) using Gird Search