README

Capstone Project:

Sentiment Analysis On Tripadvisor Reviews using ML + DL model

By:

Wong Kar Mun

Content Page

EDA + Machine Learning Models

EDA

Dataset: 79.6k rows

Notebook: ML Models 76.8k dataset_EDA.ipynb

ML Models Results

Dataset: 79.6k rows

Notebook: ML Models 76.8k dataset-Final.ipynb

KERAS (parameter comparison)

Final Model

Dataset: 79.6k rows
Epoch: 10
Hidden Layers: 2
Nodes: 64

Notebook: KERAS 79.6k-64 nodes (final)-V5.ipynb

Hyperparameter Tuning for CNN

Dataset: 10k rows

Notebook: CNN 10k-64 nodes - Hyperparameter Tuning (sampled).ipynb

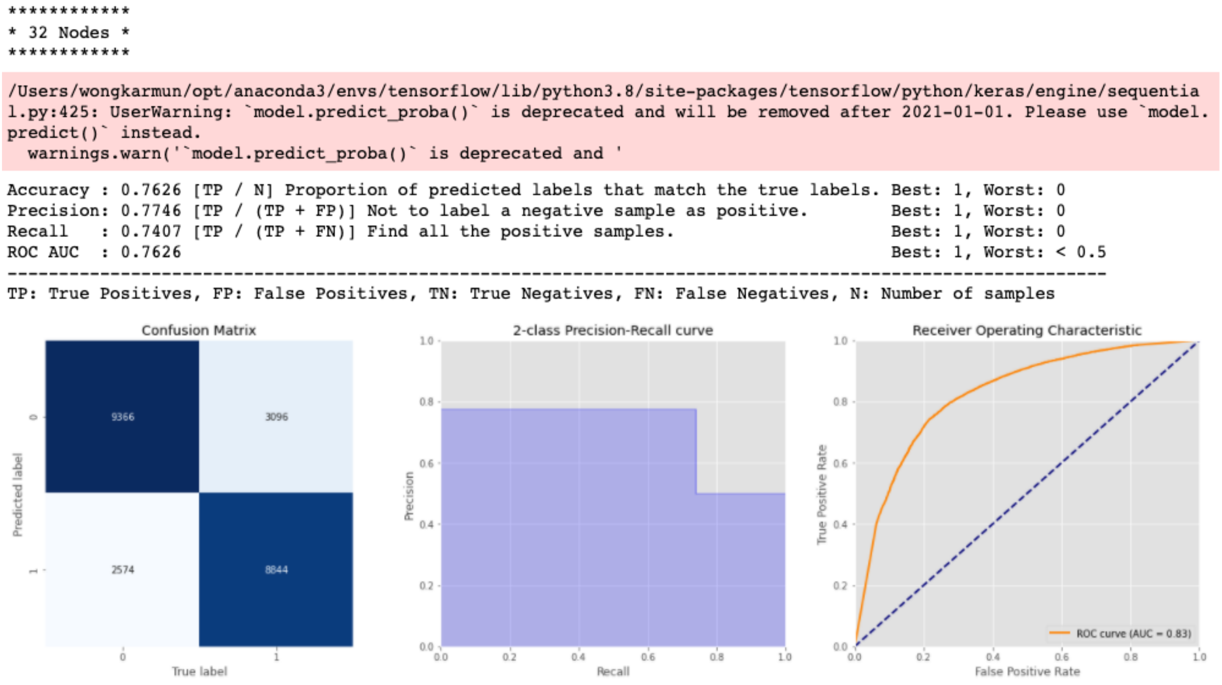
32 nodes VS 64 nodes

Training 70%, Test 30%,
Epoch = 20, 2 Hidden Layers

Notebook: KERAS 79.6k-2 layer - Nodes Comparison.ipynb

32 Nodes

Test:



```
************
* 32 Nodes *
************
```

```
/Users/wongkarmun/opt/anaconda3/envs/tensorflow/lib/python3.8/site-packages/tensorflow/python/keras/engine/sequentia
l.py:425: UserWarning: `model.predict_proba()` is deprecated and will be removed after 2021-01-01. Please use `model.
predict()` instead.
  warnings.warn('`model.predict_proba()` is deprecated and '
```

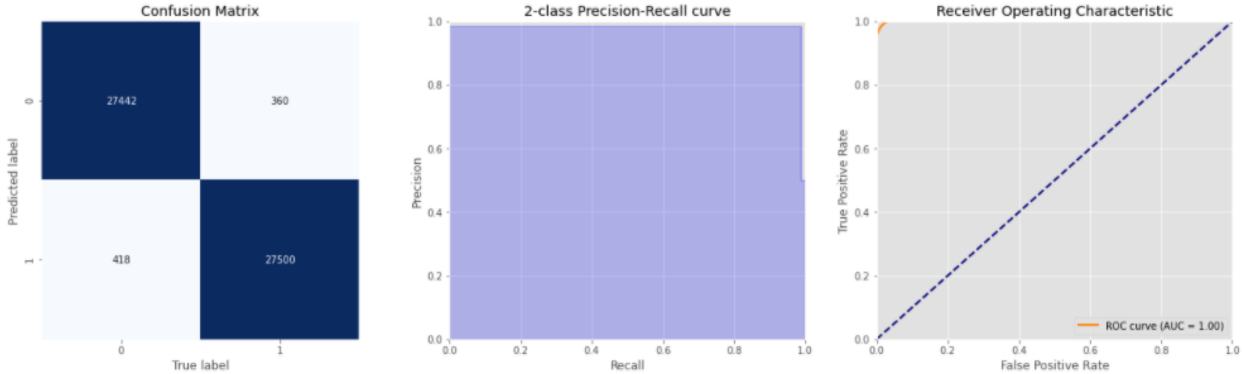
```
Accuracy : 0.7626 [TP / N] Proportion of predicted labels that match the true labels. Best: 1, Worst: 0
Precision: 0.7746 [TP / (TP + FP)] Not to label a negative sample as positive.       Best: 1, Worst: 0
Recall   : 0.7407 [TP / (TP + FN)] Find all the positive samples.                    Best: 1, Worst: 0
ROC AUC  : 0.7626                                                                    Best: 1, Worst: < 0.5
-----------------------------------------------------------------------------------------------------------
TP: True Positives, FP: False Positives, TN: True Negatives, FN: False Negatives, N: Number of samples
```

Trained:

```
************
* 32 Nodes *
************
```

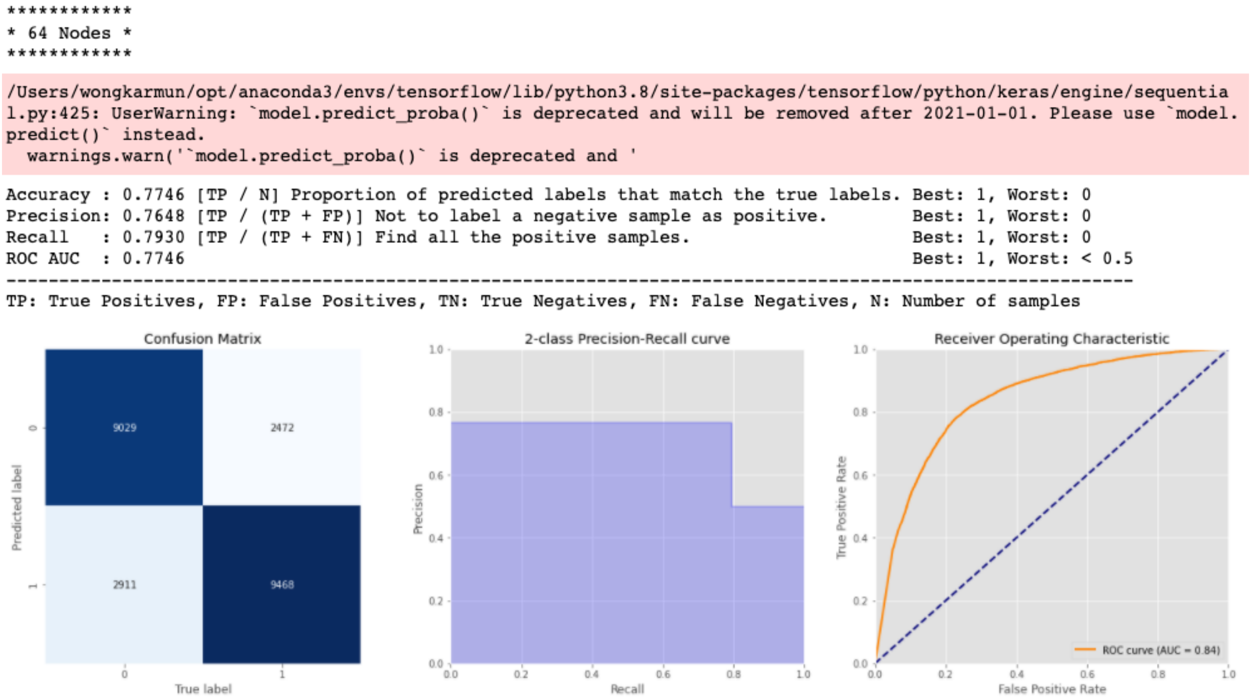
```
/Users/wongkarmun/opt/anaconda3/envs/tensorflow/lib/python3.8/site-packages/tensorflow/python/keras/engine/sequentia
l.py:425: UserWarning: `model.predict_proba()` is deprecated and will be removed after 2021-01-01. Please use `model.
predict()` instead.
  warnings.warn('`model.predict_proba()` is deprecated and '
```

```
Accuracy : 0.9860 [TP / N] Proportion of predicted labels that match the true labels. Best: 1, Worst: 0
Precision: 0.9850 [TP / (TP + FP)] Not to label a negative sample as positive.        Best: 1, Worst: 0
Recall   : 0.9871 [TP / (TP + FN)] Find all the positive samples.                     Best: 1, Worst: 0
ROC AUC  : 0.9860                                                                     Best: 1, Worst: < 0.5
-----------------------------------------------------------------------------------------------------------
TP: True Positives, FP: False Positives, TN: True Negatives, FN: False Negatives, N: Number of samples
```



64 Nodes

Test:

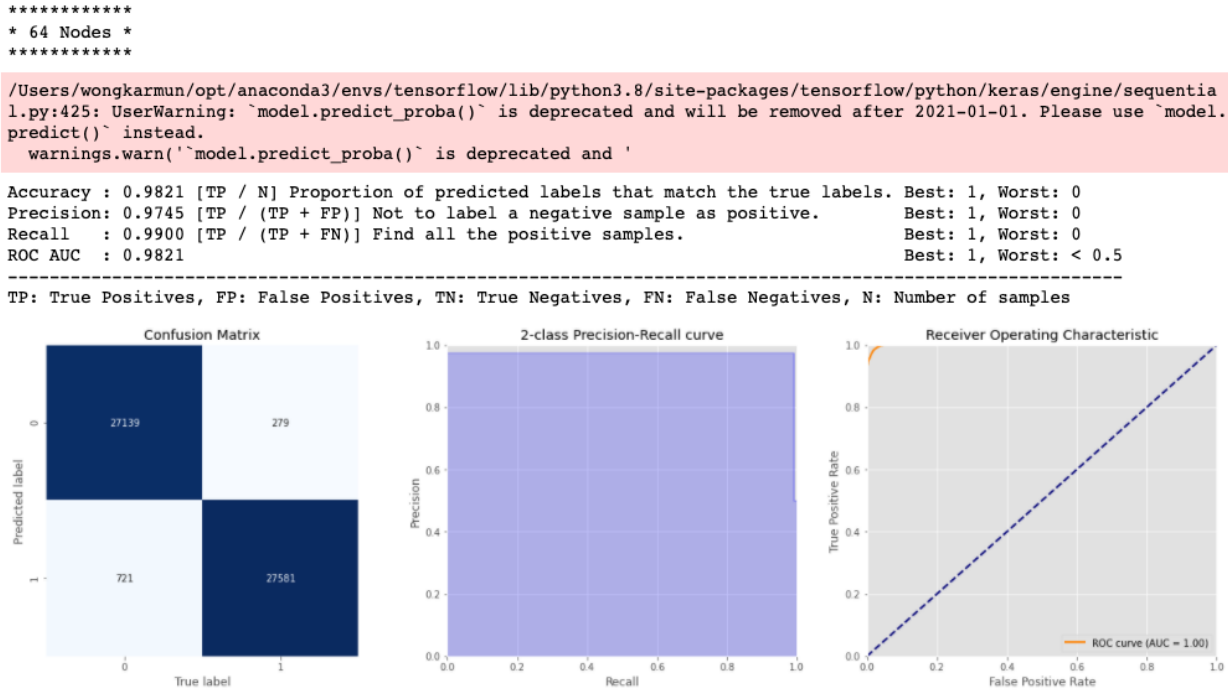


```
************
* 64 Nodes *
************
```

```
/Users/wongkarmun/opt/anaconda3/envs/tensorflow/lib/python3.8/site-packages/tensorflow/python/keras/engine/sequentia
l.py:425: UserWarning: `model.predict_proba()` is deprecated and will be removed after 2021-01-01. Please use `model.
predict()` instead.
  warnings.warn('`model.predict_proba()` is deprecated and '
```

```
Accuracy : 0.7746 [TP / N] Proportion of predicted labels that match the true labels. Best: 1, Worst: 0
Precision: 0.7648 [TP / (TP + FP)] Not to label a negative sample as positive.        Best: 1, Worst: 0
Recall   : 0.7930 [TP / (TP + FN)] Find all the positive samples.                     Best: 1, Worst: 0
ROC AUC  : 0.7746                                                                     Best: 1, Worst: < 0.5
-----------------------------------------------------------------------------------------------------------
TP: True Positives, FP: False Positives, TN: True Negatives, FN: False Negatives, N: Number of samples
```

Trained:



```
************
* 64 Nodes *
************
```

```
/Users/wongkarmun/opt/anaconda3/envs/tensorflow/lib/python3.8/site-packages/tensorflow/python/keras/engine/sequentia
l.py:425: UserWarning: `model.predict_proba()` is deprecated and will be removed after 2021-01-01. Please use `model.
predict()` instead.
  warnings.warn('`model.predict_proba()` is deprecated and '
```

```
Accuracy : 0.9821 [TP / N] Proportion of predicted labels that match the true labels. Best: 1, Worst: 0
Precision: 0.9745 [TP / (TP + FP)] Not to label a negative sample as positive.        Best: 1, Worst: 0
Recall   : 0.9900 [TP / (TP + FN)] Find all the positive samples.                     Best: 1, Worst: 0
ROC AUC  : 0.9821                                                                     Best: 1, Worst: < 0.5
-----------------------------------------------------------------------------------------------------------
TP: True Positives, FP: False Positives, TN: True Negatives, FN: False Negatives, N: Number of samples
```

	32 Nodes		64 Nodes	
	Train	Test	Train	Test
Accuracy	0.9860	0.7626	0.9821	0.7746
Precision	0.9850	0.7746	0.9745	0.7648
Recall	0.9871	0.7407	0.9900	0.7930
ROC AUC	0.9860	0.7626	0.9821	0.7746

Observations: 64 Nodes have a higher accuracy for Test data, hence chosen for further modeling.

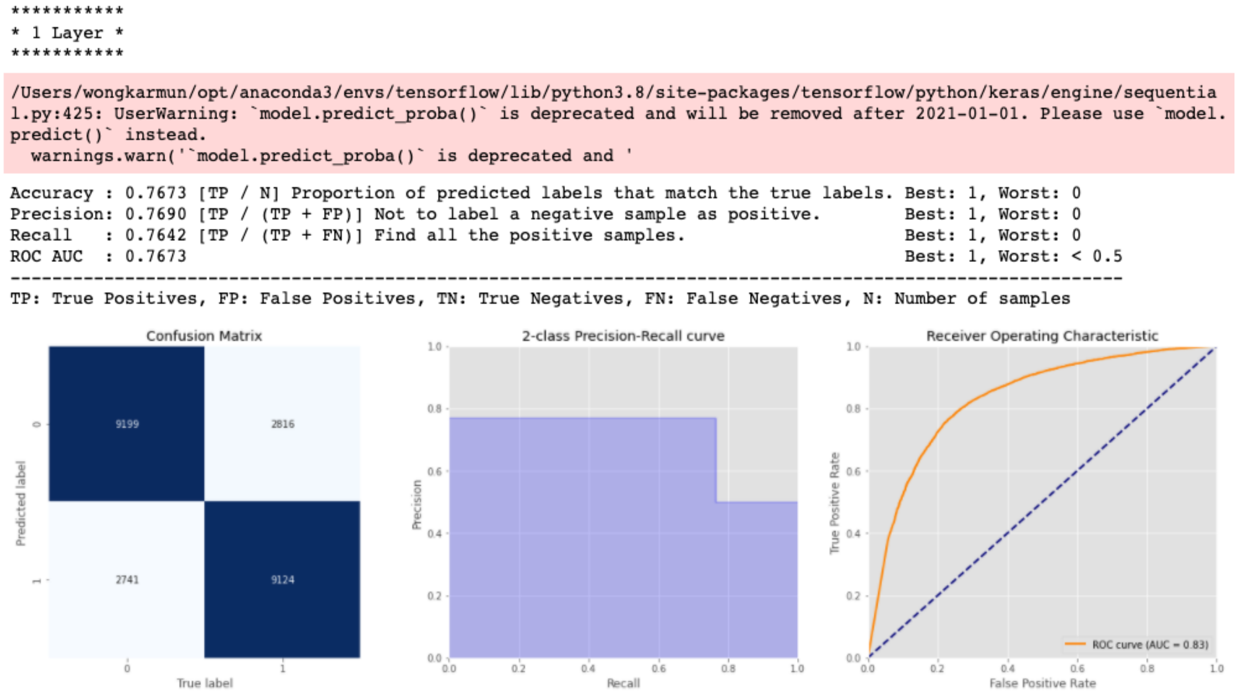
1 Layer VS 2 Layers

Training 70%, Test 30%,
Epoch = 20, Nodes = 64

Notebook: KERAS 79.6k- 64 Nodes - Layers Comparison.ipynb

1 Layer

Test:

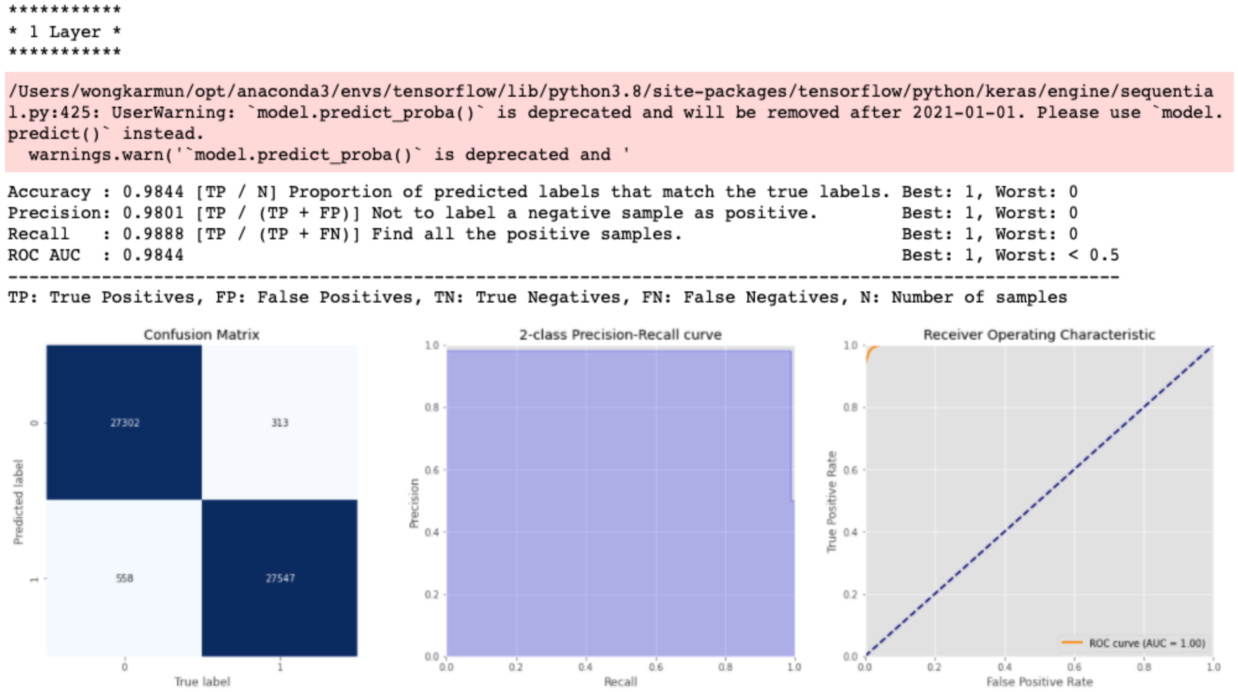


```
***********
* 1 Layer *
***********
```

```
/Users/wongkarmun/opt/anaconda3/envs/tensorflow/lib/python3.8/site-packages/tensorflow/python/keras/engine/sequentia
l.py:425: UserWarning: `model.predict_proba()` is deprecated and will be removed after 2021-01-01. Please use `model.
predict()` instead.
  warnings.warn('`model.predict_proba()` is deprecated and '
```

```
Accuracy : 0.7673 [TP / N] Proportion of predicted labels that match the true labels. Best: 1, Worst: 0
Precision: 0.7690 [TP / (TP + FP)] Not to label a negative sample as positive.        Best: 1, Worst: 0
Recall   : 0.7642 [TP / (TP + FN)] Find all the positive samples.                     Best: 1, Worst: 0
ROC AUC  : 0.7673                                                                     Best: 1, Worst: < 0.5
-----------------------------------------------------------------------------------------------------------
TP: True Positives, FP: False Positives, TN: True Negatives, FN: False Negatives, N: Number of samples
```

Trained:



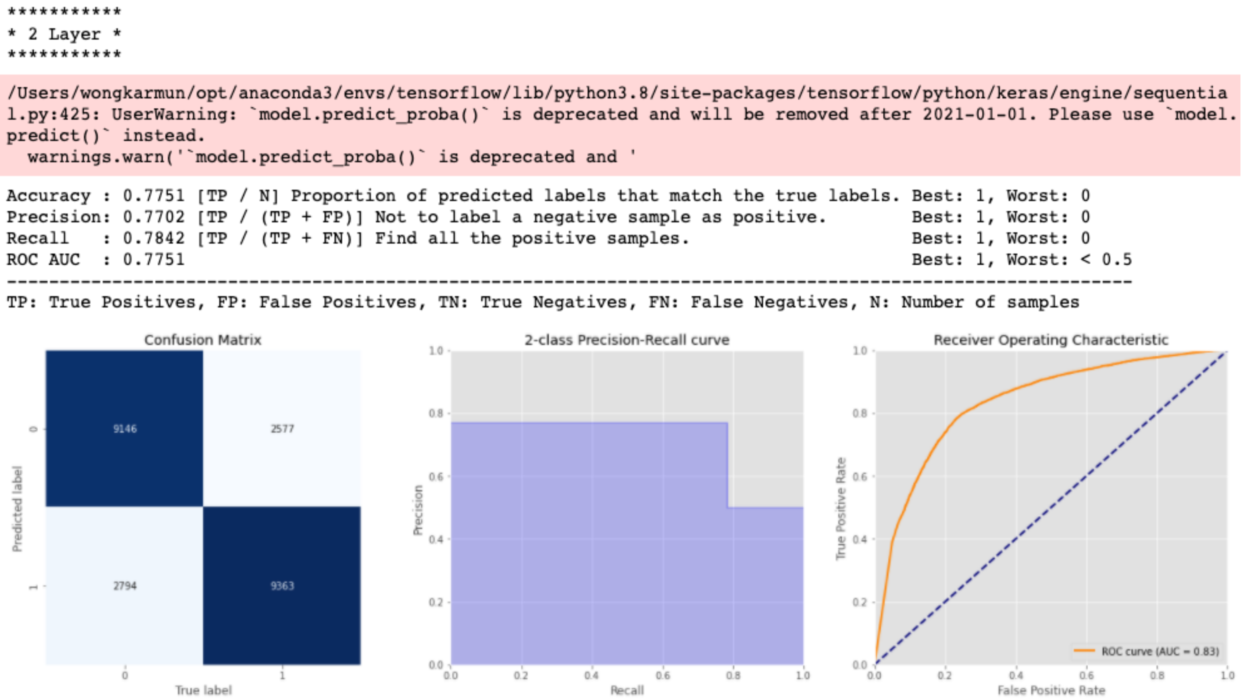
```
***********
* 1 Layer *
***********
```

```
/Users/wongkarmun/opt/anaconda3/envs/tensorflow/lib/python3.8/site-packages/tensorflow/python/keras/engine/sequentia
l.py:425: UserWarning: `model.predict_proba()` is deprecated and will be removed after 2021-01-01. Please use `model.
predict()` instead.
  warnings.warn('`model.predict_proba()` is deprecated and '
```

```
Accuracy : 0.9844 [TP / N] Proportion of predicted labels that match the true labels. Best: 1, Worst: 0
Precision: 0.9801 [TP / (TP + FP)] Not to label a negative sample as positive.        Best: 1, Worst: 0
Recall   : 0.9888 [TP / (TP + FN)] Find all the positive samples.                     Best: 1, Worst: 0
ROC AUC  : 0.9844                                                                     Best: 1, Worst: < 0.5
-----------------------------------------------------------------------------------------------------------
TP: True Positives, FP: False Positives, TN: True Negatives, FN: False Negatives, N: Number of samples
```

2 Layer

Test:

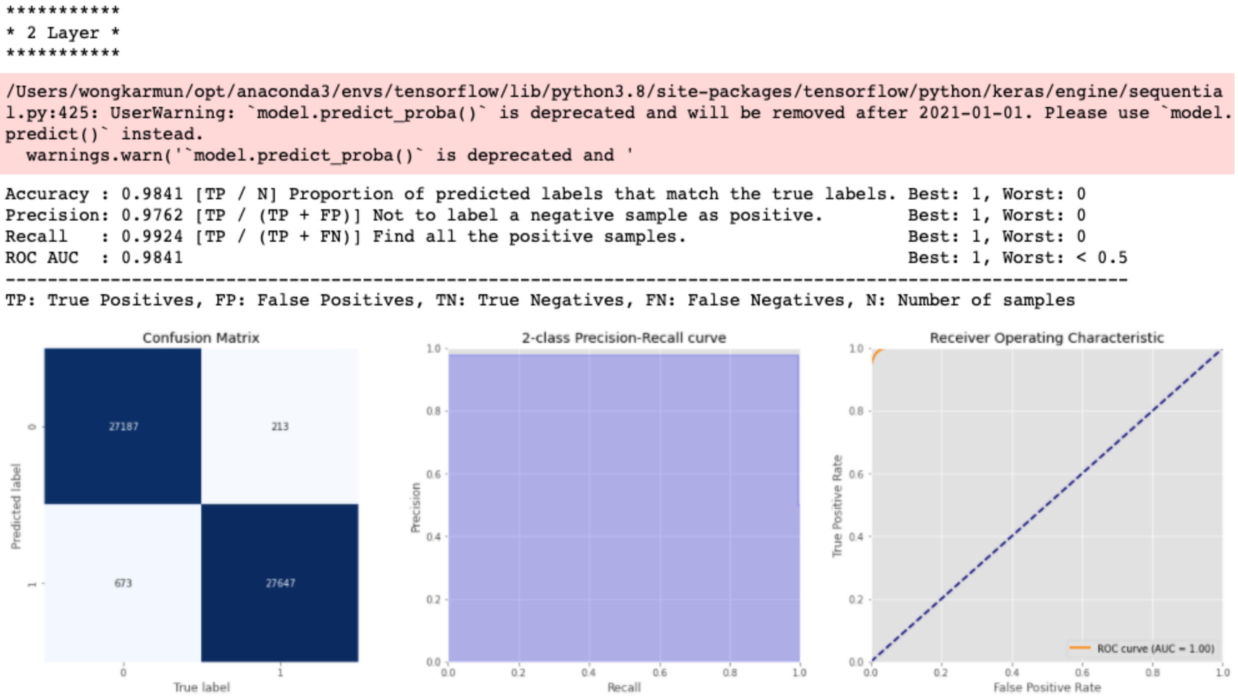


```
***********
* 2 Layer *
***********
```

```
/Users/wongkarmun/opt/anaconda3/envs/tensorflow/lib/python3.8/site-packages/tensorflow/python/keras/engine/sequentia
l.py:425: UserWarning: `model.predict_proba()` is deprecated and will be removed after 2021-01-01. Please use `model.
predict()` instead.
  warnings.warn('`model.predict_proba()` is deprecated and '
```

```
Accuracy : 0.7751 [TP / N] Proportion of predicted labels that match the true labels. Best: 1, Worst: 0
Precision: 0.7702 [TP / (TP + FP)] Not to label a negative sample as positive.        Best: 1, Worst: 0
Recall   : 0.7842 [TP / (TP + FN)] Find all the positive samples.                     Best: 1, Worst: 0
ROC AUC  : 0.7751                                                                     Best: 1, Worst: < 0.5
-----------------------------------------------------------------------------------------------------
TP: True Positives, FP: False Positives, TN: True Negatives, FN: False Negatives, N: Number of samples
```

Trained:



```
***********
* 2 Layer *
***********
```

```
/Users/wongkarmun/opt/anaconda3/envs/tensorflow/lib/python3.8/site-packages/tensorflow/python/keras/engine/sequentia
l.py:425: UserWarning: `model.predict_proba()` is deprecated and will be removed after 2021-01-01. Please use `model.
predict()` instead.
  warnings.warn('`model.predict_proba()` is deprecated and '
```

```
Accuracy : 0.9841 [TP / N] Proportion of predicted labels that match the true labels. Best: 1, Worst: 0
Precision: 0.9762 [TP / (TP + FP)] Not to label a negative sample as positive.        Best: 1, Worst: 0
Recall   : 0.9924 [TP / (TP + FN)] Find all the positive samples.                     Best: 1, Worst: 0
ROC AUC  : 0.9841                                                                     Best: 1, Worst: < 0.5
-----------------------------------------------------------------------------------------------------
TP: True Positives, FP: False Positives, TN: True Negatives, FN: False Negatives, N: Number of samples
```

	1 Hidden Layer		2 Hidden Layer	
	Train	Test	Train	Test
Accuracy	0.9844	0.7673	0.9841	0.7751
Precision	0.9801	0.7690	0.9762	0.7702
Recall	0.9888	0.7642	0.9924	0.7842
ROC AUC	0.9844	0.7673	0.9841	0.7751

Observations: 2 Hidden layers have a higher accuracy for Test data, hence chosen for further modeling.

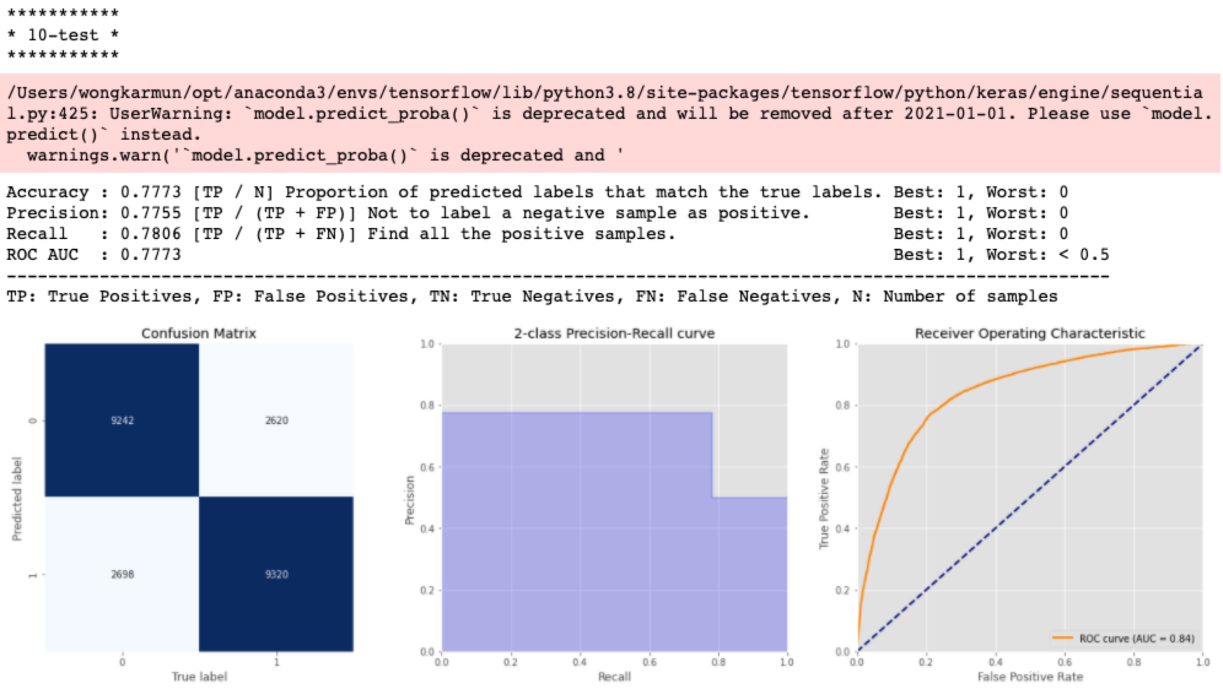
10 Epoch VS 20 Epoch

Training 70%, Test 30%,
Nodes = 64, 2 Hidden Layers

Notebook: KERAS 79.6k-2 layer - Epoch Comparison.ipynb

10 Epoch

Test:

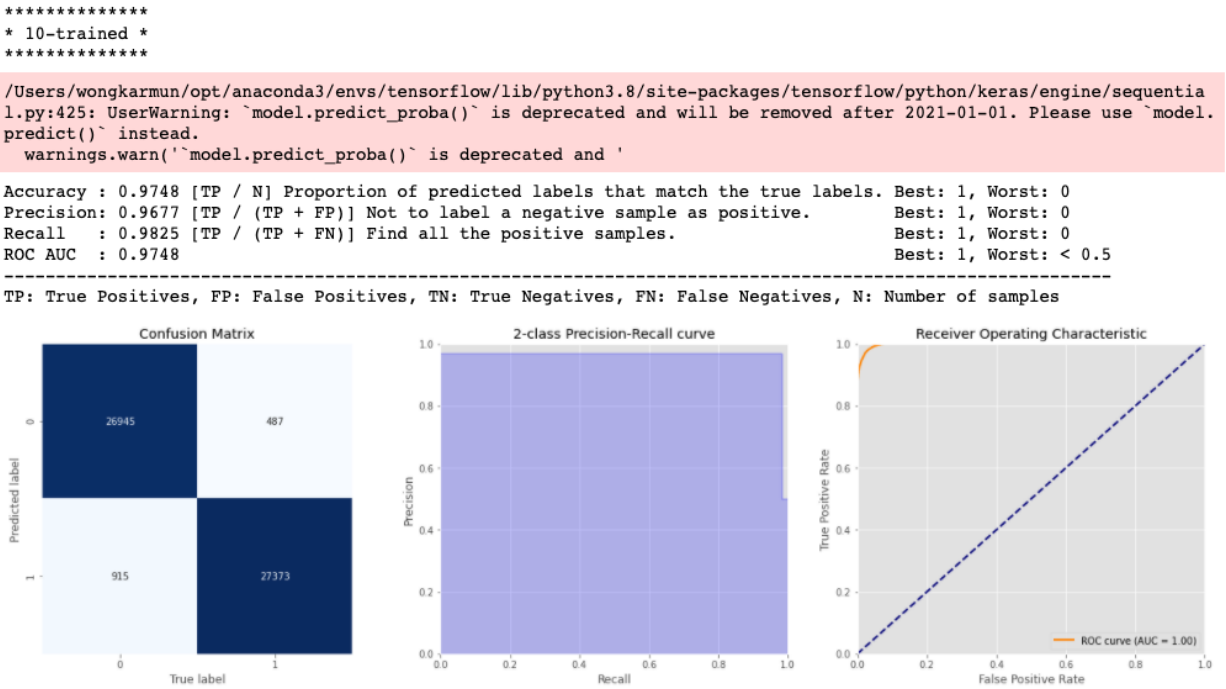


```
***********
* 10-test *
***********
```

```
/Users/wongkarmun/opt/anaconda3/envs/tensorflow/lib/python3.8/site-packages/tensorflow/python/keras/engine/sequentia
l.py:425: UserWarning: `model.predict_proba()` is deprecated and will be removed after 2021-01-01. Please use `model.
predict()` instead.
  warnings.warn('`model.predict_proba()` is deprecated and '
```

```
Accuracy : 0.7773 [TP / N] Proportion of predicted labels that match the true labels. Best: 1, Worst: 0
Precision: 0.7755 [TP / (TP + FP)] Not to label a negative sample as positive.       Best: 1, Worst: 0
Recall    : 0.7806 [TP / (TP + FN)] Find all the positive samples.                    Best: 1, Worst: 0
ROC AUC   : 0.7773                                                                    Best: 1, Worst: < 0.5
-------------------------------------------------------------------------------------------------------------
TP: True Positives, FP: False Positives, TN: True Negatives, FN: False Negatives, N: Number of samples
```

Trained:



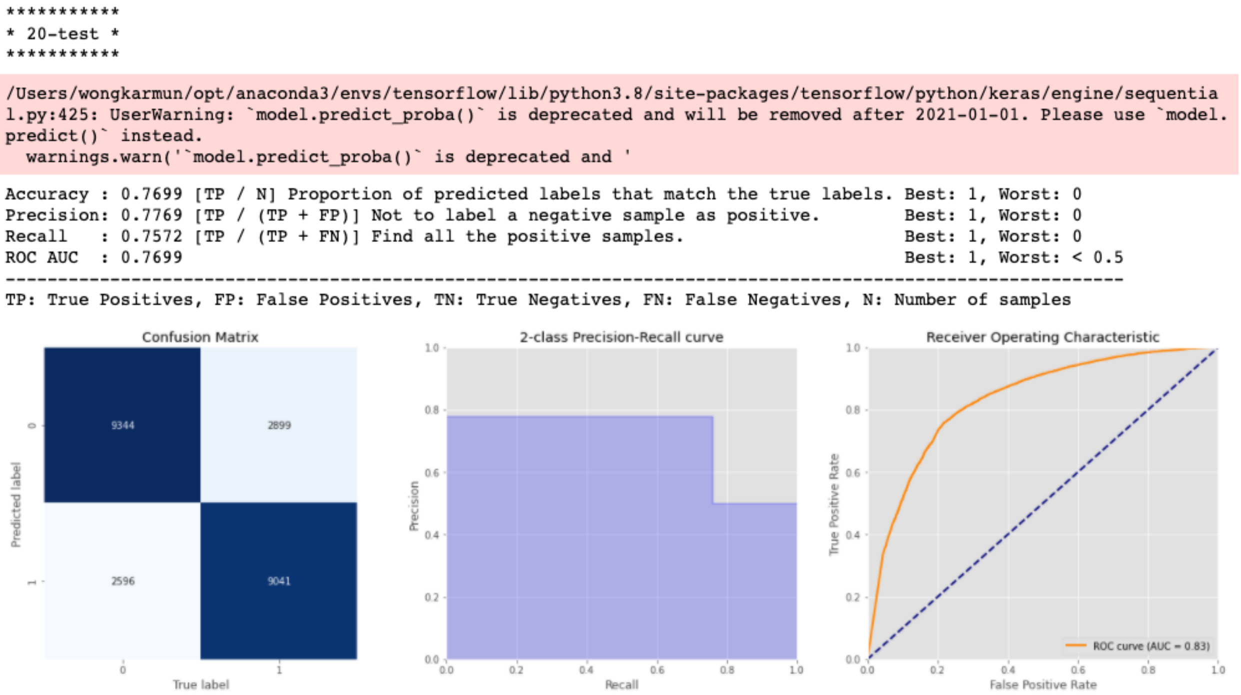
```
**************
* 10-trained *
**************
```

```
/Users/wongkarmun/opt/anaconda3/envs/tensorflow/lib/python3.8/site-packages/tensorflow/python/keras/engine/sequentia
l.py:425: UserWarning: `model.predict_proba()` is deprecated and will be removed after 2021-01-01. Please use `model.
predict()` instead.
  warnings.warn('`model.predict_proba()` is deprecated and '
```

```
Accuracy : 0.9748 [TP / N] Proportion of predicted labels that match the true labels. Best: 1, Worst: 0
Precision: 0.9677 [TP / (TP + FP)] Not to label a negative sample as positive.       Best: 1, Worst: 0
Recall    : 0.9825 [TP / (TP + FN)] Find all the positive samples.                    Best: 1, Worst: 0
ROC AUC   : 0.9748                                                                    Best: 1, Worst: < 0.5
-------------------------------------------------------------------------------------------------------------
TP: True Positives, FP: False Positives, TN: True Negatives, FN: False Negatives, N: Number of samples
```

20 Epoch

Test:

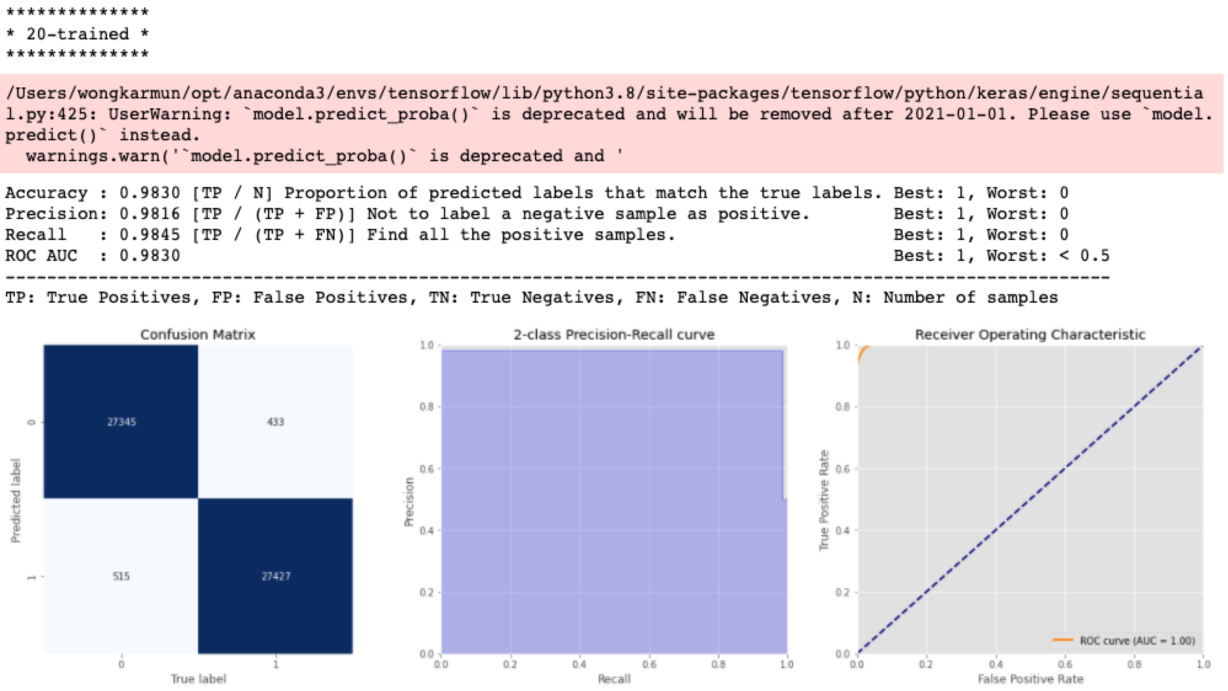


```
***********
* 20-test *
***********
```

```
/Users/wongkarmun/opt/anaconda3/envs/tensorflow/lib/python3.8/site-packages/tensorflow/python/keras/engine/sequentia
l.py:425: UserWarning: `model.predict_proba()` is deprecated and will be removed after 2021-01-01. Please use `model.
predict()` instead.
  warnings.warn('`model.predict_proba()` is deprecated and '
```

```
Accuracy : 0.7699 [TP / N] Proportion of predicted labels that match the true labels. Best: 1, Worst: 0
Precision: 0.7769 [TP / (TP + FP)] Not to label a negative sample as positive.        Best: 1, Worst: 0
Recall   : 0.7572 [TP / (TP + FN)] Find all the positive samples.                     Best: 1, Worst: 0
ROC AUC  : 0.7699                                                                     Best: 1, Worst: < 0.5
------------------------------------------------------------------------------------------------
TP: True Positives, FP: False Positives, TN: True Negatives, FN: False Negatives, N: Number of samples
```

Trained:



```
**************
* 20-trained *
**************
```

```
/Users/wongkarmun/opt/anaconda3/envs/tensorflow/lib/python3.8/site-packages/tensorflow/python/keras/engine/sequentia
l.py:425: UserWarning: `model.predict_proba()` is deprecated and will be removed after 2021-01-01. Please use `model.
predict()` instead.
  warnings.warn('`model.predict_proba()` is deprecated and '
```

```
Accuracy : 0.9830 [TP / N] Proportion of predicted labels that match the true labels. Best: 1, Worst: 0
Precision: 0.9816 [TP / (TP + FP)] Not to label a negative sample as positive.        Best: 1, Worst: 0
Recall   : 0.9845 [TP / (TP + FN)] Find all the positive samples.                     Best: 1, Worst: 0
ROC AUC  : 0.9830                                                                     Best: 1, Worst: < 0.5
------------------------------------------------------------------------------------------------
TP: True Positives, FP: False Positives, TN: True Negatives, FN: False Negatives, N: Number of samples
```

	10		20	
	Train	Test	Train	Test
Accuracy	0.9748	0.7773	0.9830	0.7699
Precision	0.9677	0.7755	0.9816	0.7769
Recall	0.9825	0.7806	0.9845	0.7572
ROC AUC	0.9748	0.7773	0.9830	0.7699

Observations: 10 Epoch have a higher accuracy for Test data, hence chosen for further modeling.