

# AutoML Modeling Report



<Ashish Anand>

## Binary Classifier with Clean/Balanced Data

### Train/Test Split

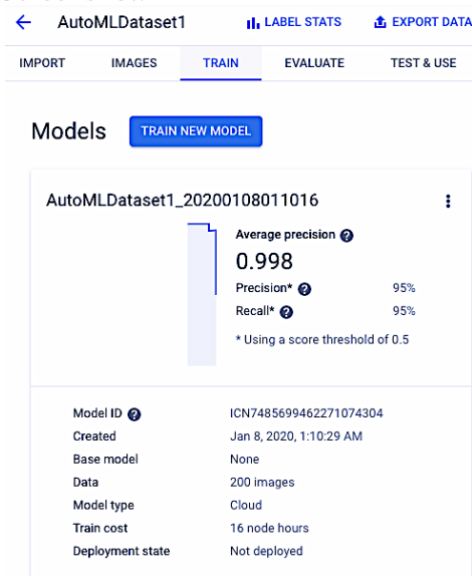
How much data was used for training? How much data was used for testing?

Model Name: [AutoMLDataset1\\_20200108011016](#)

Total Training Image = 200 images (Pneumonia = 100, Normal = 100)

Total Test items = 20

Screenshot:



### Confusion Matrix

What do each of the cells in the confusion matrix describe? What values did you observe (include a screenshot)? What is the true positive rate for the “pneumonia” class? What is the false positive rate for the “normal” class?

With Clean and balanced data, AutoML model has achieved high precision and recall.

Other factor, which helped to achieve highly sophisticated results, is quality of the data. Downloaded from Kaggle, X-ray images were pretty clean and well labeled.

True Positive (TP) of Pneumonia = 0.9 or 90%

False Positive of Normal = 0.1 or 10%

Screenshot of the Confusion Matrix:

True Label	Predicted Label	
	NORMAL	PNEUMONIA
NORMAL	100%	-
PNEUMONIA	10%	90%

### Precision and Recall

What does precision measure?  
What does recall measure? What precision and recall did the model achieve (report the values for a score threshold of 0.5)?

Precision measures the correctness of evaluating of the pneumonia or normal image in the given dataset. It indicates how often is the classifier correct to determine success criteria of the trained model.

Formula:  $TP/(TP+FP)$

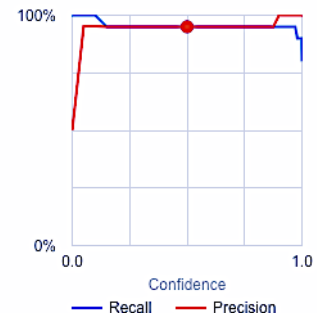
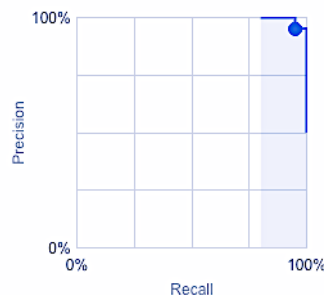
Recall measure the validation of the true positive image vs ground image. It indicates how often the model predict correctly, when it is actually correct.

Formula:  $TP/(TP+FN)$

For the score threshold of 0.5, both precision and recall for the model were 95%.

Models

[TRAIN NEW MODEL](#)



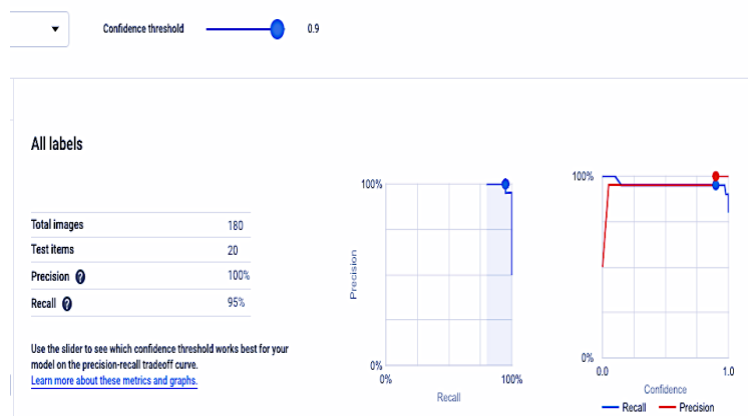
## Score Threshold

When you increase the threshold what happens to precision? What happens to recall? Why?

Changing the score threshold from 0.5 to 0.9 has changed the precision from 95% to 100%, but it has not changed the recall.

Score threshold indicates the confidence level of the model that it has to assign a category to the test run. Higher threshold lowers the risk of misclassifying images, which in case here, increased the precision rate but didn't change recall rate overall.

Screenshot of the change in precision level with confidence threshold of 0.9:



## Binary Classifier with Clean/Unbalanced Data

### Train/Test Split

How much data was used for training? How much data was used for testing?

Model Name: [AutoMLDataset2\\_20200108031200](#)  
Total Training Image = 399 images (Pneumonia = 299, Normal = 100)  
Total Test items = 40

AutoMLDataset2\_20200108031200



Average precision ?

0.991

Precision\* ?

97.5%

Recall\* ?

97.5%

\* Using a score threshold of 0.5

Model ID ?

ICN4441969801610395648

Created

Jan 8, 2020, 3:13:33 AM

Base model

None

Data

399 images

Model type

Cloud

Train cost

8 node hours

Deployment state

Not deployed

### Confusion Matrix

How has the confusion matrix been affected by the unbalanced data? Include a screenshot of the new confusion matrix.

Despite the increase of the training data, model form a biased towards the prediction of pneumonia (False positive) even if the x-ray images were normal. This is because there were 3 times more training data of pneumonia case compared to normal case.

We can see, 10% of the normal image was falsely predicted as pneumonia. Overall average precision has also declined to 0.991 as compared to previous case when it was 0.998.

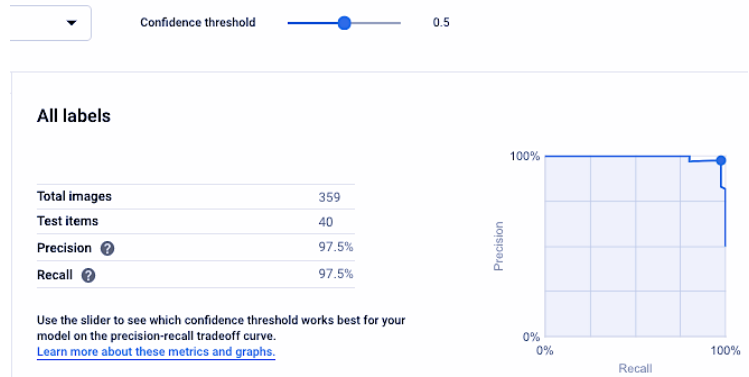
Screenshot of the Confusion Matrix:

True Label	Predicted Label	
	PNEUMONIA	NORMAL
PNEUMONIA	100%	-
NORMAL	10%	90%

### Precision and Recall

How have the model's precision and recall been affected by the unbalanced data (report the values for a score threshold of 0.5)?

Despite having much larger data (total 399), precision and recall has not changed significantly. On the other side, model has formed biased toward pneumonia case, which was not the case when we had balanced data in previous case. Please refer the screenshot below:



### Unbalanced Classes

From what you have observed, how do unbalanced classes affect a machine learning model?

We can notice overall that the model algorithm has a visible bias towards pneumonia case, because the number of images uploaded for pneumonia cases were three times higher than the normal case.

So, it's important to train a model with correctly labeled and large data BUT it's also important to feed the data which are balanced among all classes the model has.

## Binary Classifier with Dirty/Balanced Data

### Confusion Matrix

How has the confusion matrix been affected by the dirty data? Include a screenshot of the new confusion matrix.

Model Name: [AutoMLDataset3\\_20200108023527](#)  
Total Training Image = 200 images (Pneumonia = 100, Normal = 100), Total test items = 20

The confusion matrix has been severely affected by the dirty data. Precision and Recall of the model have dropped to 70%. The trained model is mislabeling 30% of total images as pneumonia when it is normal and vice versa.

Screenshot of confusion matrix:

True Label	Predicted Label	
	NORMAL	PNEUMONIA
NORMAL	70%	30%
PNEUMONIA	30%	70%

### Precision and Recall

How have the model's precision and recall been affected by the dirty data (report the values for a score threshold of 0.5)? Of the binary classifiers, which has the highest precision? Which has the highest recall?

For the value of Score threshold of 0.5, both precision and recall have dropped down to 70% with 30% FP and FN cases.

Since we added 30% of dirty data to each input class, the model is trained to mislabel both pneumonia and normal cases accordingly.

#### All labels

Total images	180
Test items	20
Precision ?	70%
Recall ?	70%

Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve.  
[Learn more about these metrics and graphs.](#)



Both binary classifiers, normal and pneumonia, surprisingly have same precision and recall (70%).

Screenshots:

#### NORMAL

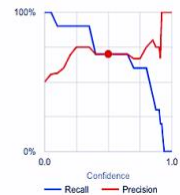
Total images	180
Test items	0
Precision ?	70%
Recall ?	70%


Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve.  
[Learn more about these metrics and graphs.](#)

All test images are evaluated at the time of training. If you modify your dataset after training, these results will not be accurate.

#### True positives

Your model correctly predicted NORMAL on these images



	<p><b>PNEUMONIA</b></p> <table border="1"> <tr> <td>Total images</td><td>180</td></tr> <tr> <td>Test items</td><td>0</td></tr> <tr> <td>Precision ?</td><td>70%</td></tr> <tr> <td>Recall ?</td><td>70%</td></tr> </table> <p>Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve.  <a href="#">Learn more about these metrics and graphs.</a></p> <p>All test images are evaluated at the time of training. If you modify your dataset after training, these results will not be accurate.</p> <p><b>True positives</b>  Your model correctly predicted PNEUMONIA on these images</p> 	Total images	180	Test items	0	Precision ?	70%	Recall ?	70%
Total images	180								
Test items	0								
Precision ?	70%								
Recall ?	70%								
<p><b>Dirty Data</b>  From what you have observed, how does dirty data affect a machine learning model?</p>	<p><b>Garbage-in = Garbage-out.</b> This holds true when we train a ML model. We altered the dataset and mislabeled pneumonia and normal images, which has impacted the model accuracy severally. So, correct and consistent data are equally important as large amount of data to create a accurate ML model.</p>								

### 3-Class Model

#### Confusion Matrix

Summarize the 3-class confusion matrix. Which classes is the model most likely to confuse? Which class(es) is the model most likely to get right? Why might you do to try to remedy the model's "confusion"? Include a screenshot of the new confusion matrix.

Model Name: [AutoMLDataset4\\_20200108043226](#)

Total Training Image = 294 images (Viral Pneumonia = 97, Bacteria Pneumonia = 98, Normal = 99)

Total test items = 30

Confusion Matrix Screenshot:

True Label	Predicted Label	VIRAL PNEUMONIA	BACTERIAL PNEUMONIA	NORMAL
VIRAL PNEUMONIA	80%	20%	-	
BACTERIAL PNEUMONIA	10%	90%	-	
NORMAL	-	10%	90%	

Normal images are most likely to predict correctly by the model.

Next to normal images, bacterial pneumonia is more likely to be predicted correctly.

Viral pneumonia is the class which is most likely to get confused.

Overall, we can see that model is struggling to identify between Viral and bacterial pneumonia in many cases.

Best remedy here is to increase the number of labeled images of each class and maintain the balance of each class.

Further, I decided to run the model with total training Image = 1501 images (Viral Pneumonia = 500, Bacteria Pneumonia = 500, Normal = 501).

Model with above larger and balance trained data has improved the accuracy of the prediction.

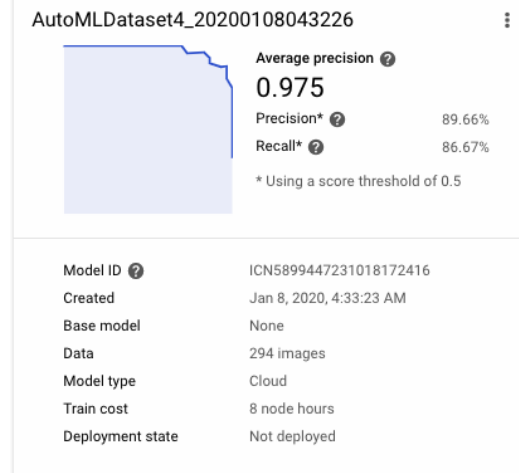


True Label \ Predicted Label	VIRAL PNEUMONIA	BACTERIAL PNEUMONIA	NORMAL
VIRAL PNEUMONIA	90%	-	10%
BACTERIAL PNEUMONIA	10%	90%	-
NORMAL	-	10%	90%

### Precision and Recall

What are the model's precision and recall? How are these values calculated (report the values for a score threshold of 0.5)?

Taking the consideration of below model:



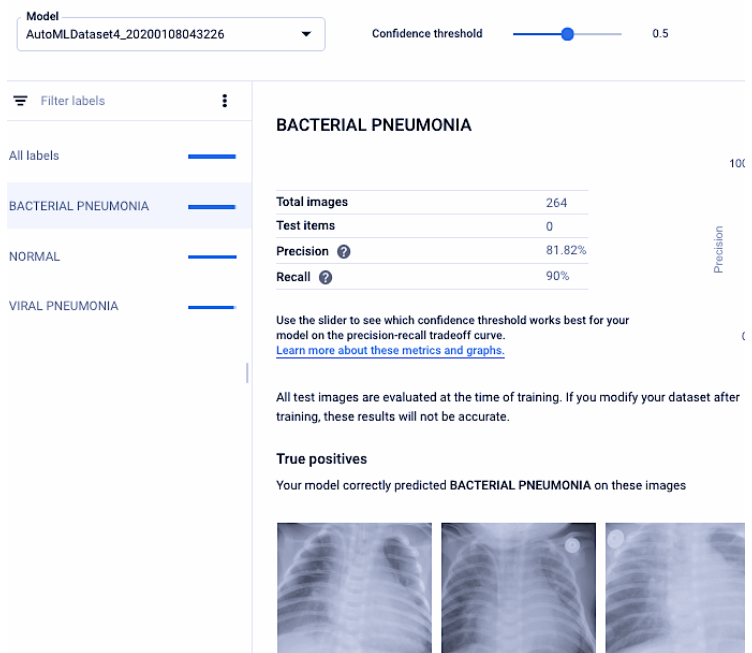
Overall all precision of this model is 89.66% and recall is 86.67%.

Precision and recall of individual class are separately calculated based on True Positive (TP), False Positive (FP), and False negative of each case by below formula:

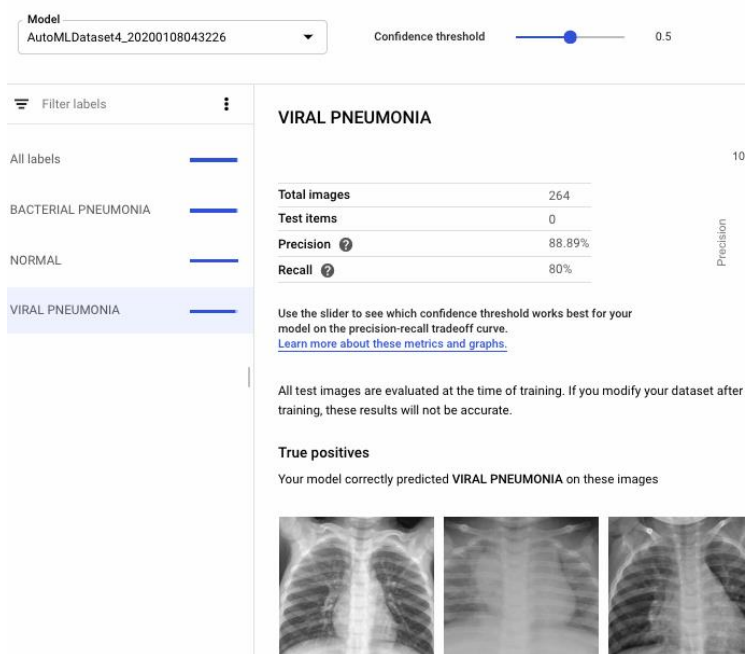
$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$


In Bacterial Pneumonia, precision is 81.82 % and recall is 90%.



In Viral Pneumonia, precision is 88.89 % and recall is 80%



In Normal case, precision is 100 % and recall is 90%

	<div> <div> Model  AutoMLDataset4_20200108043226 </div> <div> Confidence threshold <div> <div></div> 0.5 </div> </div> </div> <div> <div> Filter labels </div> <div> <div> All labels </div> <div> BACTERIAL PNEUMONIA </div> <div> <b>NORMAL</b> </div> <div> VIRAL PNEUMONIA </div> </div> <div> <div> <b>NORMAL</b> </div> <div> Total images264 </div> <div> Test items0 </div> <div> Precision100% </div> <div> Recall90% </div> <div> Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve.  <a href="#">Learn more about these metrics and graphs.</a> </div> <div> All test images are evaluated at the time of training. If you modify your dataset after training, these results will not be accurate. </div> <div> <b>True positives</b>  Your model correctly predicted <b>NORMAL</b> on these images </div> <div>  </div> </div> </div>
<b>F1 Score</b> What is this model's F1 score?	<p> F1 Score = <math>2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})</math> </p> <p> Precision of the model is 89.66% = 0.8966  and recall of the model is 86.67% = 0.8667 </p> <p> Substituting above value in the formula, </p> <p> F1 Score = <math>2 * 0.8966 * 0.8667 / (0.8966 + 0.8667)</math>  = 1.5541/1.7633  = 0.8814 </p> <p> This means the 3 classes trained model is 88.14% accurate overall. </p>