



AutoML Modeling Report

Christopher O'Hara

Binary Classifier with Clean/Balanced Data

Train/Test Split

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

Labels	Images	Train	Validation	Test
Normal		300	229	44
Pneumonia		300	233	23

Initially, 200 images were used (balanced). However, it was decided to increase the number of images to 600 to evaluate the Confusion Matrix based on adding more data.

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?

	Predicted Positives	Predicted Negatives
Actual Positives	TP	FN
Actual Negatives	FP	TN

True Positive (TP) – Correctly classified input as positive

True Negative (TN) – Correctly classified input as negative

False Positive (FP) – Misclassified input as positive

False Negative (FN) – Misclassified input as negative

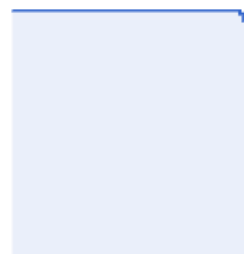
Essentially, the true positive rate for pneumonia would be the ratio of images that have the presence of pneumonia correctly classified. The false positive rate for the normal class would be cases of pneumonia that were not properly detected. The Confusion Matrix (next page) shows that the model was better at correctly identifying cases of pneumonia than cases of nominal lung condition.

True Label	Predicted Label	
	Normal	Pneumonia
Normal	93%	7%
Pneumonia	2%	98%

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)?

AutoMLDataset_v20190604000353



Average precision ?

0.998

Precision* ?

95.77%

Recall* ?

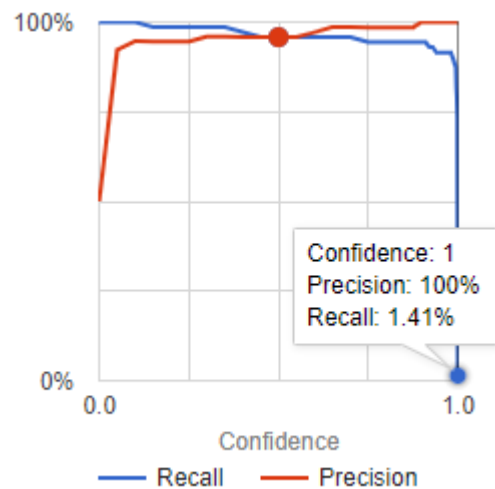
95.77%

* Using a score threshold of 0.5

Precision is the ratio of **true positives** to **predicted positives**.
Recall is the ratio of **true positives** to **actual positives**.

Score Threshold

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





When the threshold is increase, the precision increase to the maximum (approaches 1) while the recall typically plummets (approaches zero). Essentially, less false positives are returned as the acceptance criteria has been adjusted by the confidence level for a single label.

Binary Classifier with Clean/Unbalanced Data

Train/Test Split

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

Labels	Images		Train	Validation	Test
normal		100	80	10	10
pneumonia		300	240	30	30

Confusion Matrix

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

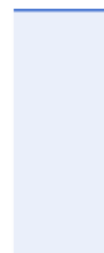
True Label	Predicted Label	
	pneumonia	normal
pneumonia	100%	-
normal	-	100%

The imbalanced data has actually improved the performance values within the Confusion Matrix. This might have results from training the 400 images on 16 node hours (an increase). It was initially expected that the results of having an unbalanced set would increase the number of false positives and false negatives.

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)?

CleanUnbalanced_1_20200112031918



Average precision ?

1

Precision* ?

100%

Recall* ?

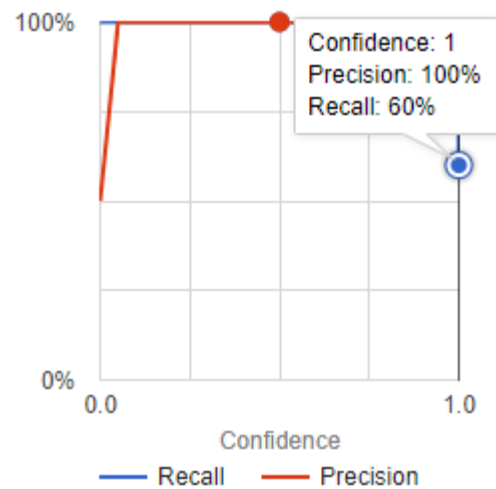
100%

* Using a score threshold of 0.5

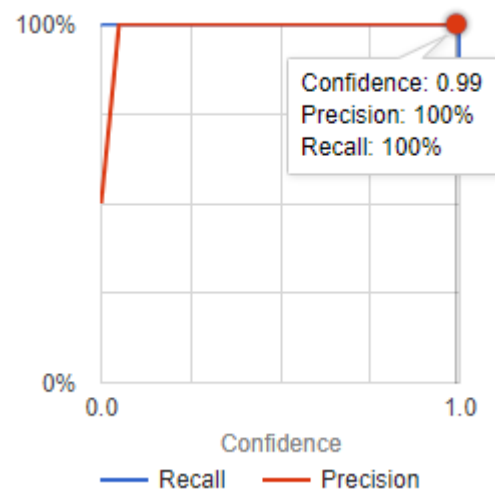
As a result of the True values above, the precision and recall are now at 100% as there are no errors present.

Unbalanced Classes

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



In general, unbalanced data introduces biases that impact the accuracy of the model. However, with such a high accuracy, the recall at a confidence value of one is unusually high (60%). For comparison at a confidence value of 0.99, both precision and recall remain at 100% (see below).



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.

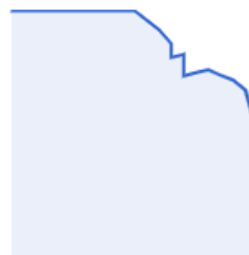
True Label	Predicted Label	
	normal	pneumonia
normal	60%	40%
pneumonia	10%	90%

Observing the Confusion Matrix, the “dirty data” as vastly decreased the true positives for the normal image set. As a result, it becomes difficult for the model to properly evaluate images that are fed into it. With a ratio of 70:30 (clean vs dirty), the model becomes ineffective at classifying images while reaffirms the importance of ensuring improper data is never mixed.

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?

DirtyBalanced_157_20200112031033



Average precision ?

0.906

Precision* ?

75%

Recall* ?

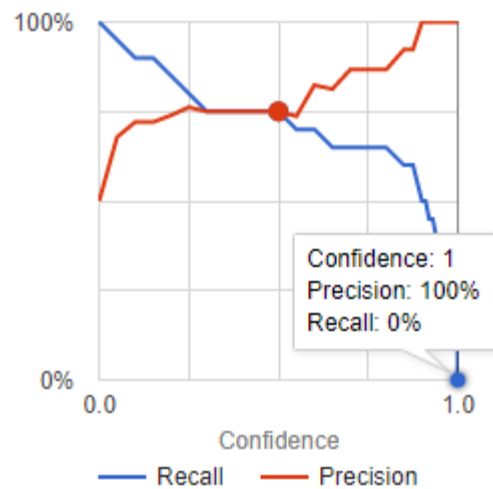
75%

* Using a score threshold of 0.5

As a result, the precision and recall have decrease but are *artificially high*. That is, using such a model without knowing that the data has issues will fail to meet quality standards and would be ineffective in practical usage. Thus far, the Clean-Unbalanced image set has had the best precision and recall though this could be due to overtraining (node hours) and “luck” in image sorting taking during preprocessing (as opposed to selecting on the first 100 images, for example).

Dirty Data

From what you have observed, how does dirty data affect a machine learning model?



Analyzing the cross-over point, the impact of the “dirty data” can be seen to impact the number of correct predications based on the confidence level. As such, cleaning data and proper allocation is a necessity for any model.

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.

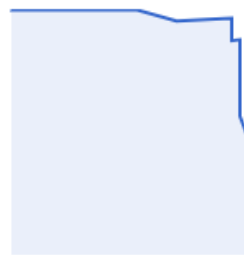
True Label	Predicted Label		
	bacteria	viral	normal
bacteria	100%	-	-
viral	10%	90%	-
normal	-	9%	91%

Intuitively, the bacterial and viral pneumonia cases are the most similar which leads to inaccuracies in proper prediction. Even for human observers, the differences can be difficult to observe (which is why binary classification leads to better results when predicting the existence of an object or ailment as compared to proper identification and classification). To remedy this, I would train with much more data as well as change the labeling scheme since the keyword "pneumonia" is in the metadata for both types of pneumonia (which is a logical choice that inexperienced technicians could easily make).

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)?

ThreeClass_157883_20200112035929



Average precision ?

0.95

Precision* ? 96.55%

Recall* ? 90.32%

* Using a score threshold of 0.5

$$P_{model} = \frac{\sum_{i=1}^n P_i}{n}$$

$$R_{model} = \frac{\sum_{i=1}^n R_i}{n}$$

F1 Score

What is this model's F1 score?

$$F1 = \frac{2 * Precision * Recall}{(Precision + Recall)} = \frac{2 * \Pi\left(\frac{TP}{(TP+FP)}\right) * \Pi\left(\frac{TP}{(TP+FN)}\right)}{\Pi\left(\frac{TP}{(TP+FP)}\right) + \Pi\left(\frac{TP}{(TP+FN)}\right)}$$
$$F1 = 93.33\%$$