AutoML Modeling Report

Bertrand Ndakena

Binary Classifier with Clean/Balanced Data

Train/Test Split

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

I decided to create two models for balanced data;

1. **dataset100**(dataset100_v20190626165240) : A model with 200 images(100 normal and 100 pneumonia).
2. **AutoMLDataset600**(AutoMLDataset600_v20190626165107): A model with 600 images(300 normal and 300 pneumonia).

   I aimed at seeing how the number of training samples affects the different metrics used in AutoML. See the models below and their corresponding number of training images, validation images, and Testing images.

Model	**Training images**	**Validation images**	**Test images**
dataset100	157	21	22
AutoMLDataset600	478	61	61

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?

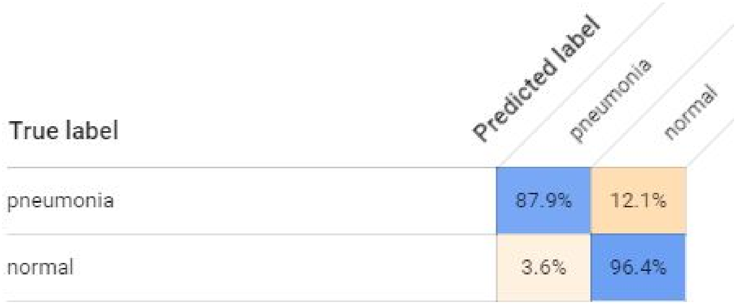
We want to know whether an image has pneumonia or not. **pneumonia** means **positive** and **normal** means **negative**

	Predicted pneumonia	Predicted normal
Actual pneumonia	True Positive(TP)	False negative(FN)
Actual normal	False positive (FP)	True negative(TN)

We look at the confusion matrix table above to explain what each cell

1. **True positive** stands for pneumonia images that were correctly predicted
2. **False negative** stands for normal images that were predicted as pneumonia.
3. **False positive** stands for pneumonia images that were predicted as normal
4. **True negative** stands for normal images that were correctly predicted

Now we will consider the description of each cell above to identify the values we obtained from our model on the balanced dataset with 600 images (300 normal and 300 pneumonia). Below is the confusion matrix.



True label	Predicted label pneumonia	normal
pneumonia	87.9%	12.1%
normal	3.6%	96.4%

TP rate(Pneumonia) =TP/(TP+FN)=87.9/(87.9+12.1)
= 87.9%

FN rate(Normal) = FN/(FN+TN)=12.1/(96.4+12.1)
= 11.2%

Precision & 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)?

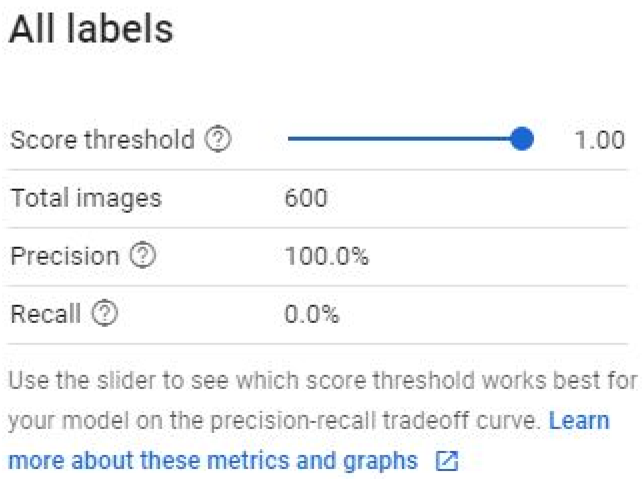
Precision measures the fraction of positive values predicted as positive with respect to total values predicted as positive irrespective of whether they are actually positive or not. In order words, when a model makes a prediction, model precision measures the likelihood for it to be correct.
Recall measures the fraction of positive values predicted as positive with respect to the total actual positive values. In other words, it measures how good a model is at identifying the actual occurrences in a dataset.
**Generally, precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned.

The model achieved,
Precision = 91.8% (for a score threshold of 0.5)
Recall = 91.8% (for a score threshold of 0.5)

Score Threshold

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

The score threshold allows us to adjust the likelihood of a decision based on inputs. In other words, if you want to put your model for use in ML applications, it is a good idea to adjust the score threshold in order to maximize the performance of the model.



All labels

Score threshold ⓘ	1.00
Total images	600
Precision ⓘ	100.0%
Recall ⓘ	0.0%

Use the slider to see which score threshold works best for your model on the precision-recall tradeoff curve. Learn more about these metrics and graphs

When we increase the threshold,

1. The precision increases because the high threshold produces fewer FP. in our context, it implies that very few normal images will be classified under the pneumonia class. This also means that there is a lower probability of saying that a child has pneumonia when he/she is healthy.
2. The recall decreases because, in addition to decreasing the FP, we also increase FN by increasing the recall. This also means that we will have more pneumonia patients considered to be healthy.

Binary Classifier with Clean/Unbalanced Data

Train/Test Split

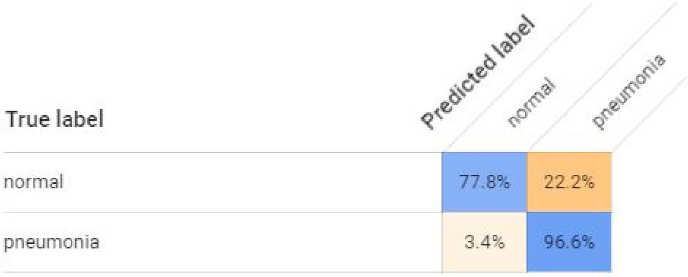
How much data was used for training? How much data was used for testing?

Model	Training images	Validation images	Test images
DatasetUnbalance_v20190626151743	326	36	38

Confusion Matrix

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

For imbalanced data (100 normal and 300 pneumonia), the confusion matrix shows that the percentage of correctly identified pneumonia cases has increased, even the percentage of normal and healthy images being identified as having pneumonia has increased.

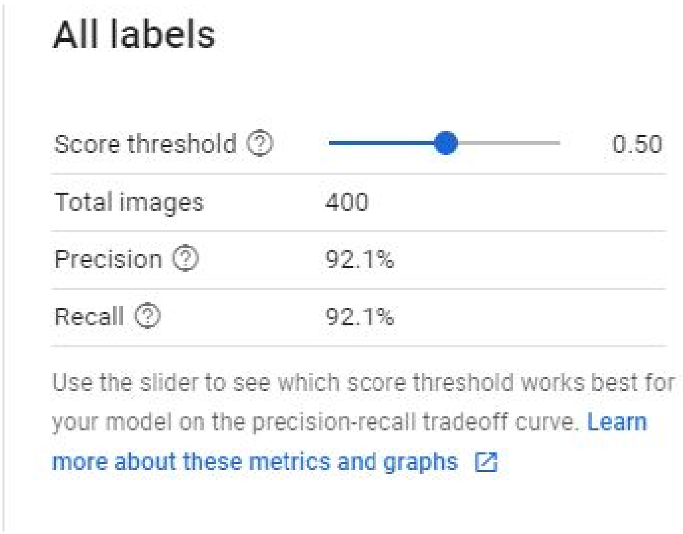


True label	normal	pneumonia
normal	77.8%	22.2%
pneumonia	3.4%	96.6%

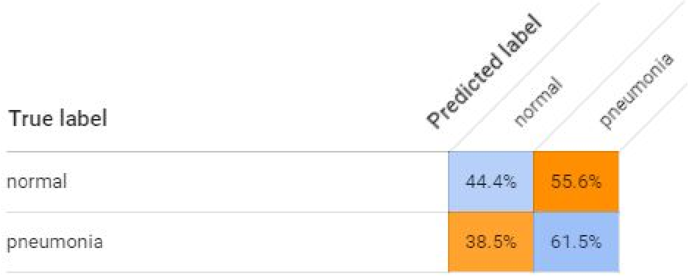
Precision & Recall

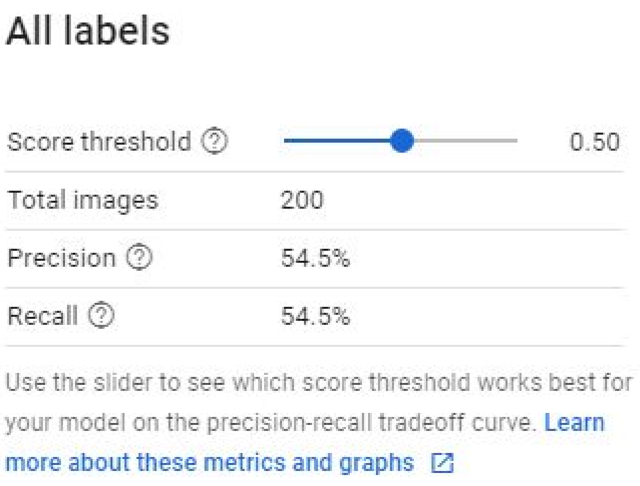
How have the model's precision and recall been affected by the unbalanced data? (Report the values for a score threshold of 0.5.)

The model's precision and recall have unexpectedly increased, implying that the model has performed better in an imbalanced dataset. This is misleading, note that it could be that the model has become more used to identifying pneumonia images and consequently has a high tendency of identifying every image as having pneumonia. See the values on the screenshot below.

	 All labels Score threshold ⓘ 0.50 Total images 400 Precision ⓘ 92.1% Recall ⓘ 92.1% Use the slider to see which score threshold works best for your model on the precision-recall tradeoff curve. Learn more about these metrics and graphs
Unbalanced Classes From what you've observed, how do unbalanced classes affect a machine learning model?	From my observation, unbalanced classes give a sort of "naive behavior", giving an impression that the accuracy of the model is high, but greatly reducing the number of TP. If the number of TP decreases, it implies that the chances of having correct identification or correct prediction have dropped.

Binary Classifier with Dirty/Balanced Data

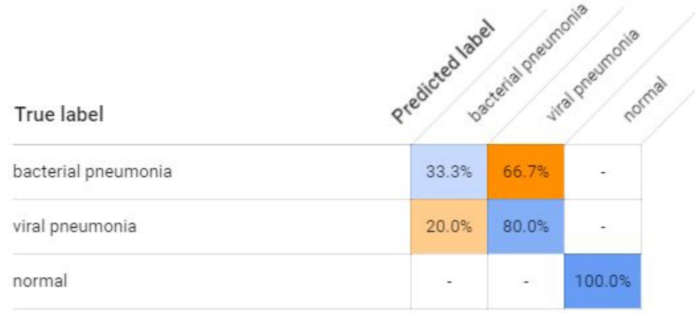
Confusion Matrix How has the confusion matrix been affected by the dirty data? Include a screenshot of the new confusion matrix.	Looking at the screenshot below, we see the confusion matrix has given a result that is basically complicated to understand, it's very difficult to determine where the model can be improved on.  True label | Predicted label: normal, pneumonia normal | 44.4% | 55.6% pneumonia | 38.5% | 61.5%

Precision & 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?.	Both the precision and the recall have dropped. Indicating that we have a lower chance of correct predictions by the model.  All labels Score threshold 0.50 Total images 200 Precision 54.5% Recall 54.5% Use the slider to see which score threshold works best for your model on the precision-recall tradeoff curve. Learn more about these metrics and graphs
Dirty Data From what you’ve observed, how do dirty data affect a machine learning model?	From observation, the nature of data is very critical for the success of Machine Learning applications. The algorithms that Google uses in AutoML can be powerful, but without the relevant or right data training, the ML system may fail to yield ideal results. Therefore, data should be clean to ensure proper training by the ML model for ML applications.

3-Class Model

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

The model is most likely to confuse the bacterial pneumonia and the viral pneumonia classes and is most likely to the the normal class right. To remedy the model's confusion, I will consider using a different naming pattern for each class to avoid confusion. The confusion is actually between bacterial pneumonia and the viral pneumonia classes as their naming patterns are similar.

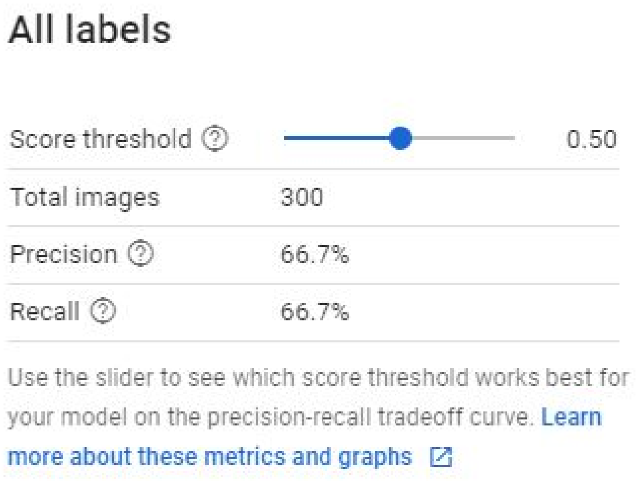


True label \ Predicted label	bacterial pneumonia	viral pneumonia	normal
bacterial pneumonia	33.3%	66.7%	-
viral pneumonia	20.0%	80.0%	-
normal	-	-	100.0%

Precision & Recall

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

Find below, the model's precision and recall. To calculate both the precision and the recall, we first calculate their individual values for each class and we take their averages.



All labels

Score threshold ⓘ	0.50
Total images	300
Precision ⓘ	66.7%
Recall ⓘ	66.7%

Use the slider to see which score threshold works best for your model on the precision-recall tradeoff curve. Learn more about these metrics and graphs

	$P_{model} = (P_{bacteria} + P_{viral} + P_{normal})/3$ $R_{model} = (R_{bacteria} + R_{viral} + R_{normal})/3$
F1 Score What is this model's F1 score?	To calculate the F1 score, we use the following formula $F_{1-score} = \frac{2 \times P_{model} \times R_{model}}{P_{model} + R_{model}}$. =2*0.667*0.667/2*0.667 = 0.667 = 66.7%