AutoML Modeling Report

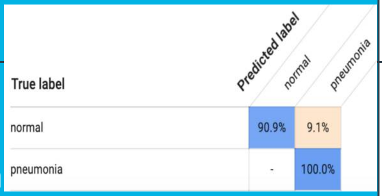
< Eric Armstrong >

Binary Classifier with Clean/Balanced Data

<table>
<tr>
<td><strong>Train/Test Split</strong>

How much data was used for training? How much data was used for testing?</td>
<td>This first variation used 100 "normal" images, balanced with 100 "pneumonia" images - either bacterial or viral.

Google AutoML divided these up into training images (147), validation images (32), and test images (21).</td>
</tr>
<tr>
<td><strong>Confusion</strong>


<table>
<tr><th>True label</th><th>Predicted label
normal</th><th>pneumonia</th></tr>
<tr><td>normal</td><td>90.9%</td><td>9.1%</td></tr>
<tr><td>pneumonia</td><td>-</td><td>100.0%</td></tr>
</table>
What do ea
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?</td>
<td>True Positive (top left, 90.9%) - Nearly ninety percent of the normal images were correctly identified as being normal.

False Negative (top right, 9.1%) - About 9 percent of the normal images were mistakenly predicted as having pneumonia.

False Positive (bottom left, null) - None of the pneumonia images were predicted to be normal.

True Negative (bottom right, 100%) - All of the actual cases of pneumonia were correctly predicted as pneumonia by the model.

The true positive rate for the "pneumonia" class is 100%
The false positive rate for the "normal" class is 9.1%</td>
</tr>
<tr>
<td><strong>Precision & Recall</strong>

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)?</td>
<td>Precision relates to false positives. The closer to 100%, the lower the fewer false positives, and vice versa. This model achieved a precision of 95.238%.

Recall relates to false negatives. The closer to 100%, the lower the false negatives, and vice versa. This model achieved a recall of 95.238%. (yes, the same percent as precision!)</td>
</tr>
<tr>
<td><strong>Score Threshold</strong>

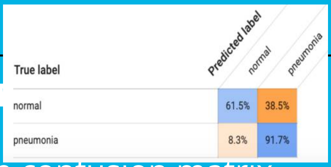
When you increase the score threshold, what happens to precision? What happens to recall? Why?</td>
<td>When the score threshold is increased from .5 to .75, precision increases to 100.0%, while recall decreases to 81.0%

With precision at 100%, the model should now produces fewer false positives. The model should still not predict actual pneumonia cases to be normal.

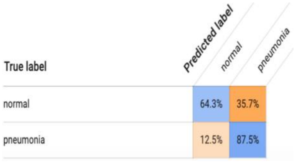
With recall at 81%, the model now produces more false negatives.

In other words, when you ask the model to be more confident in identifying whether an image is normal, it goes a bit overboard in calling out what it thinks to be pneumonia. The model here is becoming a hypochondriac.</td>
</tr>
</table>

Binary Classifier with Clean/Unbalanced Data

Train/Test Split How much data was used for training? How much data was used for testing?	In total, 299 images were used in the Clean/Unbalanced scenairo. There are 100 "normal" images, and 199 "pneumonia" images. AutoML dropped 1 duplicate image. Google AutoML divided these up into training images (227), validation images (35), and test images (37).
Confusion  True label Predicted label normal 61.5% 38.5% pneumonia 8.3% 91.7% How has the confusion matrix been affected by the unbalanced data? Include a screenshot of the new confusion matrix.	The confusion matrix has been somewhat adversely affected by using unbalanced data... True positive (top left, 61.5%) - decreased by 29.4%. Several fewer normal images correctly predicted to be normal. False negative (top right, 38.5%) - increased by 29.4.%. Several more normal images thought to have pneumonia. False positive (bottom left, 8.3%) - increased by 8.3%. Slightly more pneumonia images thought to be normal. True negative (bottom right, ...%) - decreased by 8.3%. Slightly fewer pneumonia images correctly predicted to have pneumonia.
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.)	Using unbalanced data has caused the model's precision to decrease from the clean/balanced baseline of 95.238% to 81.081%. A decrease of 14.157%, essentially meaning a few more false positives. Furthermore, the unbalanced data has caused the model's recall to fall from the clean/balanced baseline of 95.238% to 81.081%. Also a decrease of 14.157%, meaning a marginal increase in false negatives.
Unbalanced Classes From what you've observed, how do unbalanced classes affect a machine learning model?	From what I've observed in this instance, unbalanced classes in the data make the machine learning model marginally less "useful," overall. To be specific, among precision, recall, and the confusion matrix, every desirable measure has decreased while each undesirable measure has increased. In general, the model is worse at distinguishing normal and pneumonia images. Having trained on more pneumonia images, it seems more ready to predict pneumonia when none is present. At the same time, interestingly, some actual pneumonia cases are incorrectly predicted to be normal.

Binary Classifier with Dirty/Balanced Data



True label

Predicted label

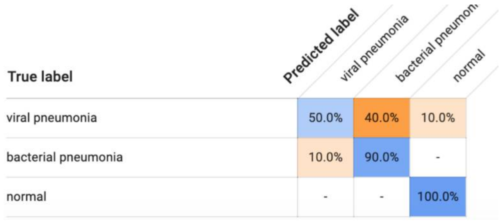
normal 64.3% 35.7%

pneumonia 12.5% 87.5%

Confusion Matrix How has the confusion matrix been affected by the dirty data? Include a screenshot of the new confusion matrix.	The confusion matrix has been adversely affected by the dirty data... True positive (top left, 64.3%) - decreased by 26.6%. The model can correctly predict just under 2/3 of the normal cases as normal. False negative (top right, 35.7%)- increased by 26.6%. Over 1/3 of normal images are incorrectly predicted as having pneumonia. False positive (bottom left, 12.5%) - increased by 12.5%. 1 in 8 actual pneumonia images are incorrectly predicted to be normal. True negative (bottom right, 87.5%) - decreased by 12.5%. 7 in 8 pneumonia images are correctly predicted to have pneumonia.

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?	Using dirty data has caused the model's precision to decrease from the clean/balanced baseline of 95.238% to 72.727%. A massive decrease of 22.511%, essentially meaning a many more false positives. Furthermore, the unbalanced data has caused the model's recall to fall from the clean/balanced baseline of 95.238% to 72.727%. Also a substantial decrease of 22.511% here, meaning a large increase in false negatives. Of the binary classifiers, the one with the clean, balanced data (first) has both the highest precision and the highest recall.
Dirty Data From what you’ve observed, how do dirty data affect a machine learning model?	Without a doubt, dirty data seem to have a significant negative impact on the machine learning model. The confusion matrix, as well as precision & recall, reflect a model that is mediocre at performing its intended function: distinguishing normal, healthy chest x-ray images from similar images that depict bacterial or viral pneumonia.

3-Class Model



True label	Predicted label: viral pneumonia	bacterial pneumon.	normal
viral pneumonia	50.0%	40.0%	10.0%
bacterial pneumonia	10.0%	90.0%	-
normal	-	-	100.0%

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 3-class confusion matrix is a 3x3 grid depicting the overlay of the true labels on the data (viral pneumonia, bacterial pneumonia, and normal) against the ML model's predicted label for each of the three classes. Viral pneumonia and bacterial pneumonia are the classes the model are most likely to confuse. Normal is the class that the model is most likely to get right. It correctly identifies 100% of the normal cases as being normal. To remedy the model's "confusion," I would increase the amount of training data for both of the pneumonia cases (more for viral than for bacterial). Specifically, the model could stand to do much better at distinguishing viral pneumonia from bacterial. In general, it seems to be pretty good at predicting bacterial pneumonia.
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.)	The model's precision is 76.0%, and the recall is also 76.0%. These values are calculated based on the number of false positive and false negatives made by the model in testing. Essentially, the more incorrect types of predictions the model makes, the lower its precision and recall. The fewer incorrect predictions (and more correct predictions), the higher the precision and recall.
F1 Score What is this model’s F1 score? $\text{F1} = 2 \times \frac{Precision*Recall}{Precision+Recall}$	.76x.76=0.5776 .76+.76=1.52 .5776/1.52=.38 2x.38=.76 Based on this formula, the model's F1 score is .76 (76%).