AutoML Modeling Report

Odili Charles Opute

Binary Classifier with Clean/Balanced Data

<table>
<tr>
<td>Train/Test Split
How much data was used for training? How much data was used for testing?</td>
<td>Dataset: 478 Images

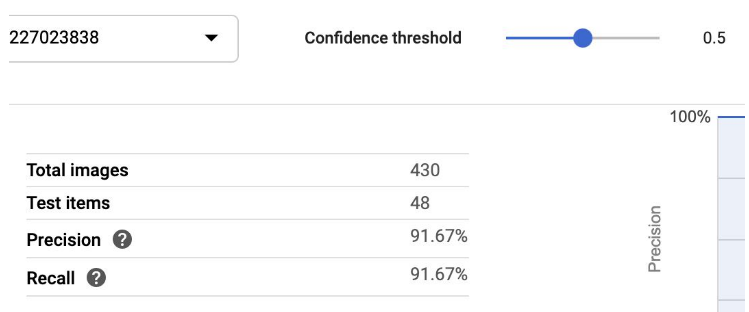
Normal Label: 239 Images
Train/Test/Validation Split: 191 / 24 / 24

Pneumonia Label: 239 Images
Train/Test/Validation Split: 191 / 24 / 24</td>
</tr>
<tr>
<td>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?</td>
<td>Below is the confusion matrix of the model :

<table>
<tr><th>True Label</th><th>Predicted Label: pneumonia</th><th>normal</th></tr>
<tr><td>pneumonia</td><td>83%</td><td>17%</td></tr>
<tr><td>normal</td><td>-</td><td>100%</td></tr>
</table>

The cells represent and visualise the correctness of the predictions that the model made for each class (label), and indicate where the model might be confused and can be improved

For instance, for the pneumonia class, the model correctly predicted 83% pneumonia cases as pneumonia (True Positive), and wrongly predicted 17%</td>
</tr>
</table>

	pneumonia cases are normal(False Negative). The model can however correctly predict the normal class (cases of no pneumonia). In summary, our model is a little confused about the pneumonia class and likely to make wrong predictions of it, than the normal class.Thus, we can spend some time improving the data and predictions for the pneumonia class. **Pneumonia TPR: 0.83** **Normal FPR: 0**
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 ability of a model to identify only the relevant data (i.e make correct predictions) **Recall** measures the ability of a model to identify and find all of the relevant data (i.e make as many correct predictions as possible) High values for Precision and Recall indicate that the model is making correct predictions (Precision) and also making the majority of such correct predictions (Recall). **Precision & Recall @ 0.5 Threshold** 
Score Threshold When you increase the threshold what happens to precision? What happens to recall? Why?	Precision appears to be directly proportional to the threshold value, while Recall appears to be inversely proportional to the threshold value. At a confidence threshold of 0, Precision is 50% while Recall is 100%. As the threshold value increases, Precision increases but Recall decreases, such that at a confidence

	threshold of 1, Precision is 100% while Recall is 0%. **Why**: Precision and Recall often have an inverse relationship that works like a trade off between them. Hence, tuning the confidence threshold (e.g to 0.5 instead of 0 or 1) is a way to optimise the model to maximise both metrics at the same time. High values for Precision and Recall indicate that the model is making correct predictions (Precision) and also making the majority of such correct predictions (Recall).

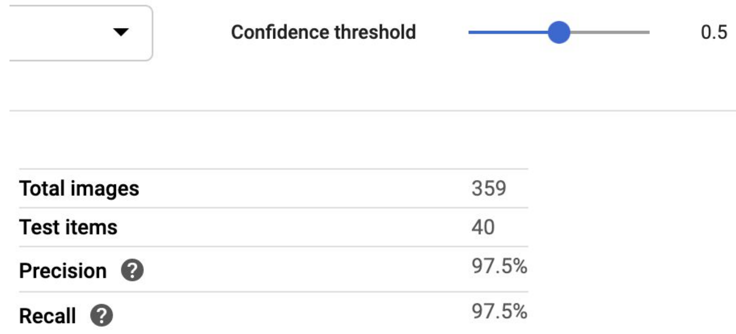
Binary Classifier with Clean/Unbalanced Data

Train/Test Split How much data was used for training? How much data was used for testing?	Dataset: 399 Images Normal Label: 100 Images Train/Test/Validation Split: 80 / 10 / 10 Pneumonia Label: 299 Images Train/Test/Validation Split: 239 / 30 / 30
Confusion Matrix How has the confusion matrix been affected by the unbalanced data? Include a screenshot of the new confusion matrix.	 The unbalanced data across the classes in our dataset has resulted in the model being 10% less likely to correctly predict cases of no pneumonia. Our model also thinks it can completely correctly predict presence of pneumonia (100%), even though we know

	from the previous model which had more data and more balanced data showed 17% False Negative for the pneumonia class

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)?

At a confidence threshold of 0.5, the model's Precision and Recall values also appear to have improved by 5.8%, going from 91.6% to 97.5%.



This improvement might however be erroneous since the dataset has unbalanced data across the classes

Unbalanced Classes From what you have observed, how do unbalanced classes affect a machine learning model?	From observation, unbalanced class data can severely affect a Machine Learning model because it can create a false representation of the model's performance (e.g accuracy, precision and recall) and make it generally difficult to rely on the confusion matrix as a guide on where to improve the model

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.	With dirty data (e.g wrongly labeled images) in our dataset, our model is now 33% further less likely able to predict presence of pneumonia. For the pneumonia label, we've gone from 17% False Negative to 50% False Negative. Though not quite visible with the confusion matrix, our normal label now has much higher False Positives since it is reporting 100% correct prediction for this class, even though some of its data are wrongly labeled.

<table>
<tr>
<td></td>
<td>
<table>
<tr><th>True Label</th><th>Predicted Label
pneumonia</th><th>normal</th></tr>
<tr><td>pneumonia</td><td>50%</td><td>50%</td></tr>
<tr><td>normal</td><td>-</td><td>100%</td></tr>
</table>
</td>
</tr>
<tr>
<td>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?</td>
<td>
Confidence threshold 0.5
<table>
<tr><td>Total images</td><td>180</td></tr>
<tr><td>Test items</td><td>20</td></tr>
<tr><td>Precision</td><td>75%</td></tr>
<tr><td>Recall</td><td>75%</td></tr>
</table>
Overall, the model's Precision and Recall took a 16.7% decline (down from 91.67%)

Of the three binary classifiers, the model with unbalanced data (100 no-pneumonia images vs 299 pneumonia images) has the highest Precision and Recall values of 97.5% each
</td>
</tr>
<tr>
<td>Dirty Data
From what you have observed, how does dirty data affect a machine learning model?</td>
<td>From observation, dirty data (e.g incorrectly annotated data) can significantly affect a model's Precision and Recall negatively, especially increasing False Positives and False Negatives</td>
</tr>
</table>

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? 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 shows that the model is confused about predicting the presence of pneumonia.

True Label \ Predicted Label	bacterial pneumonia	normal	viral pneumonia
bacterial pneumonia	69%	6%	25%
normal	-	100%	-
viral pneumonia	6%	-	94%

While the model is most likely to be confused about predicting bacterial pneumonia (31% False Negative), it will correctly predict cases of no pneumonia.

We can attempt to remedy this situation and improve the model by adding more accurately labeled data for the bacterial and viral pneumonia classes

The below confusion matrix is the result of training the model with almost 900 data points for the bacterial and viral pneumonia classes each, vs only 200 data points for the normal class.

True Label \ Predicted Label	bacterial pneumonia	viral pneumonia	normal
bacterial pneumonia	89%	10%	1%
viral pneumonia	24%	74%	2%
normal	8%	8%	84%

There still appears to be more confusion in the model, even for the normal class since it appears the model is

<table>
<tr>
<td></td>
<td>somewhat biased towards the pneumonia classes. We can improve this with more training cycles and experimenting with various data [im]balances for the 3 classes.

That said, this iteration of the model is better able to identify general pneumonia cases vs no-pneumonia cases, as highlighted by the area under the red box.</td>
</tr>
<tr>
<td>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)?</td>
<td>Using a score threshold of 0.5, the model’s Precision and Recall are both 0.875 (87.5%)

Confidence threshold 0.5

<table>
<tr><td>Total images</td><td>426</td></tr>
<tr><td>Test items</td><td>48</td></tr>
<tr><td>Precision</td><td>87.5%</td></tr>
<tr><td>Recall</td><td>87.5%</td></tr>
</table>

Precision is calculated as: True Positives / (True Positives + False Positives)

Recall is calculated as: True Positives / (True Positives + False Negatives)</td>
</tr>
<tr>
<td>F1 Score
What is this model’s F1 score?</td>
<td>Given F1 Score = (2 * Precision * Recall) / (Precision + Recall)

This model’s F1 Score = (2 * 0.875 * 0.875) / (0.875 + 0.875)
= 0.9</td>
</tr>
</table>