AutoML Modeling Report

Christina Loiacono

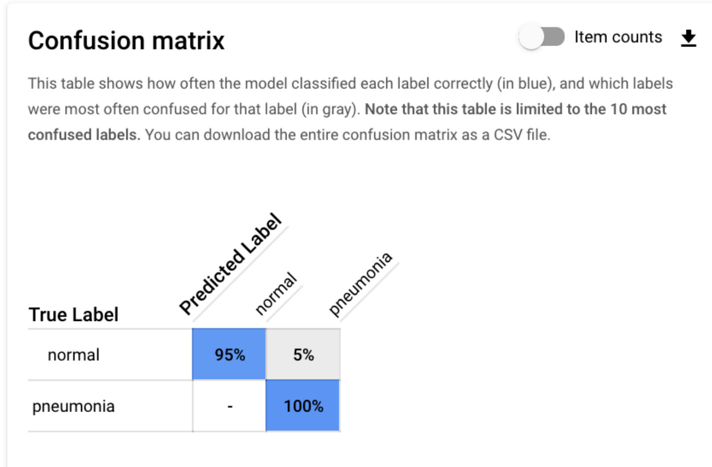
Binary Classifier with Clean/Balanced Data

<table>
<tr>
<td>Train/Test Split
How much data was used for training? How much data was used for testing?</td>
<td>Data used:
200 Normal Images
200 Pneumonia Images

Total number of images used for this model: 400

I used a high number of images for this model (400) because the directions say to use sets of 100-300 images, and 200 is in the middle. What makes this clean/balanced data, is that the number of “normal” class images, and the number of “pneumonia” class images are equal. By using more images than required on the rubric (originally 100 “normal” class images, and 100 “pneumonia” class images) this gives the model more data to work with, providing more correct examples, which should increase the accuracy/True Positives.

How the images were distributed:
360 for training and 40 for testing</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>What each of the cells in the confusion matrix describe:
<table>
<tr>
<td>True Positive for “normal” class.
How often the model classified each label correctly</td>
<td>False Positive for “normal” class.
Labels were most often confused for that label</td>
</tr>
<tr>
<td>False Negative</td>
<td>True Positive for “pneumonia” class.
How often the model classified each label correctly</td>
</tr>
</table>
</td>
</tr>
</table>

<table>
<tr>
<td></td>
<td>True positive rate for the “pneumonia” class: 100%
False positive rate for the “normal” class: 5%

</td>
</tr>
<tr>
<td>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)?</td>
<td>What precision measures:
Precision tells us how many images (that were assigned a label) were supposed to be categorized with that label.
A high precision model produces fewer false positives.

What recall measures:
Recall tells us that out of all the images that should have had the label assigned, how many were actually assigned the correct label.
A high recall model produces fewer false negatives.

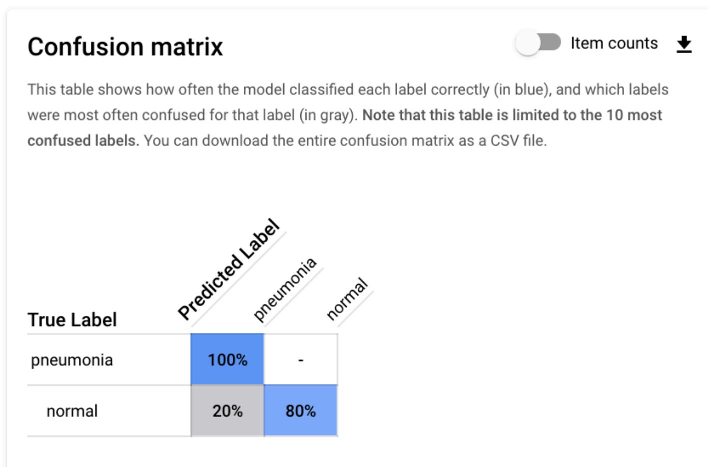
Model Achieved (values for a score threshold of 0.5):
Precision: 97.5%
Recall: 97.5%</td>
</tr>
<tr>
<td>Score Threshold
When you increase the threshold what happens to precision? What happens to recall? Why?</td>
<td>Increasing the threshold – what happens to Precision:
The precision goes up: 100% (if threshold = 1)
Because: The model produces fewer false-positives.

Increasing the threshold – what happens to Recall:
The recall goes down: 0% (if threshold = 1)
Because: The model produces fewer false-negatives.</td>
</tr>
</table>

Binary Classifier with Clean/Unbalanced Data

<table>
<tr>
<td>Train/Test Split
How much data was used for training? How much data was used for testing?</td>
<td>Data used:
100 Normal Images
300 Pneumonia Images
Total number of images used for this model: 400

How the images were distributed:
360 for training and 40 for testing</td>
</tr>
<tr>
<td>Confusion Matrix
How has the confusion matrix been affected by the unbalanced data? Include a screenshot of the new confusion matrix.</td>
<td>How the confusion matrix has been affected by the unbalanced data:
The labels are now switched – the “pneumonia” class is on the top/left, while the “normal” class is on the bottom and right sides of the confusion matrix.

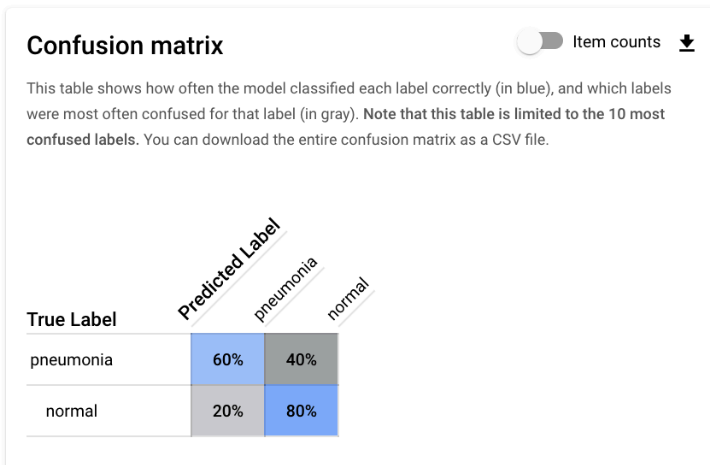
<table>
<tr><td>True Positive for “pneumonia” class</td><td>False Negative</td></tr>
<tr><td>False Positive for “normal” class</td><td>True Positive for “normal” class</td></tr>
</table>


</td>
</tr>
<tr>
<td>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)?</td>
<td>Model Achieved (values for a score threshold of 0.5):
Precision: 95%
Recall: 95%

The unbalanced data is now lower in both precision and recall, compared to the model with clean/balanced data.</td>
</tr>
<tr>
<td>Unbalanced Classes
From what you have observed, how do unbalanced classes affect a machine learning model?</td>
<td>Unbalanced classes have a higher percentage of false positives, which means the labels are getting confused more often. This model is most likely to be biased towards predicting “pneumonia” since there are a significantly higher number of “pneumonia” class images in the data.</td>
</tr>
</table>

	Since there were 300 “pneumonia” images, the True Positive/accuracy was 100%, while for the 100 “normal” images there was less data, so the True Positive/accuracy was only 80%.

Binary Classifier with Dirty/Balanced Data

<table>
<tr>
<td>Confusion Matrix
How has the confusion matrix been affected by the dirty data? Include a screenshot of the new confusion matrix.</td>
<td>How the confusion matrix has been affected by the dirty data:
There were 30 “pneumonia” images added to the “normal” class
There were 30 “normal” images added to the “pneumonia” class.
<table>
<tr><td>True Positive for “pneumonia” class</td><td>False Positive for “pneumonia” class</td></tr>
<tr><td>False Positive for “normal” class</td><td>True Positive for “normal” class</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>Model Achieved (values for a score threshold of 0.5):
Precision: 70%
Recall: 70%

The precision and recall have gone down significantly in this model.
It labeled the “pneumonia” class correctly: 60% (as True Positive)
It labeled the “normal” class correctly: 80% (as True Positive)</td>
</tr>
</table>

	The binary classifier with the highest precision is: Clean/Balanced data The binary classifier with the highest recall is: Clean/Balanced data
Dirty Data From what you have observed, how does dirty data affect a machine learning model?	Even though there were 100 images in the “pneumonia” class, and 100 images in the “normal” class, the “normal” class had higher accuracy/True Positives of 80%, than the “pneumonia” class accuracy/True Positives of 60%. Dirty data also creates more False Positive results, which is why the precision and recall are much lower in this model.

3-Class Model

<table>
<tr>
<td>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.</td>
<td>Summary of the 3-class confusion matrix.
Classes the model is most likely to confuse:
“viral_pneumonia” and “bacterial_pneumonia” classes, because the confusion matrix shows false positives for the “viral_pneumonia” class.

Class(es) the model is most likely to get right:
“bacterial_pneumonia” and “normal” classes, because the True Positive/accuracy was 100% for both.

Why you might try to remedy the model’s “confusion”:
The confusion is occurring in the two classes labeled as “pneumonia”, but having to choose whether it is “bacterial_pneumonia” or “viral_pneumonia”. In this model, some of the “viral_pneumonia” images were labeled as “bacterial_pneumonia” instead.
The “normal” class is being labeled accurately.

<table>
<tr><td>True Positive for “bacterial_pneumonia” class</td><td>False Negative</td><td>False Negative</td></tr>
<tr><td>False Positive for “viral_pneumonia” class</td><td>True Positive for “viral_pneumonia” class</td><td>False Negative</td></tr>
<tr><td>False Negative</td><td>False Negative</td><td>True Positive for “normal” class</td></tr>
</table>
</td>
</tr>
</table>

<table>
<tr>
<td></td>
<td>

Confusion matrix

Item counts

This table shows how often the model classified each label correctly (in blue), and which labels were most often confused for that label (in gray). **Note that this table is limited to the 10 most confused labels.** You can download the entire confusion matrix as a CSV file.

True Label \ Predicted Label	bacterial_pneumonia	viral_pneumonia	normal
bacterial_pneumonia	100%	-	-
viral_pneumonia	30%	70%	-
normal	-	-	100%

To remedy the situation:
You could look at the incorrectly classified images, and include more “bacterial_pneumonia” class images that look similar to these.
The additional images will go through more training cycles, which will remedy this model’s confusion, where the “viral_pneumonia” class is currently being mislabeled.
</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>

Model Achieved (values for a score threshold of 0.5):
Precision: 90%
Recall 90%

These values are calculated by:
Taking the True Positives and the False Positives, and seeing which ones were mislabeled.
Precision is calculated by: seeing how many labeled images were supposed to be categorized with that label. To do this, you divide the true positives by the sum of the true positives and the false positives.

Recall is calculated by looking at: from all the labeled images that should have had the label assigned, how many were actually assigned that label. To do this, you divide the true positives by the sum of the true positives and the false negatives.
</td>
</tr>
<tr>
<td>

F1 Score
What is this model’s F1 score?
</td>
<td>

This model’s F1 score is calculated by first calculating each of the classes individually.

Steps:
You calculate the recall for each of the classes.
Next, you calculate the precision for each of the classes.
</td>
</tr>
</table>

Then you find the overall model recall and precision.

Once you have these, you plug them into the formula for the F1 score and get the overall model performance.

Variables used:
bp = bacterial_pneumonia class
vp = viral_pneumonia class
n = normal class

Precision:
Precision bp = 100% correct/130 predictions
= 0.77
Precision vp = 70% correct/70 predictions
= 1.0
Precision n = 100% correct/100 predictions
= 1.0
Precision for model = 0.77 + 1.0 + 1.0/3 classes
= 0.92 or 92%

Recall:
Recall bp = 100%/100 total images
= 1.00
Recall vp = 70%/100 total images
= .70
Recall n = 100%/100 total images
= 1.00
Recall for model = 1.0 + 0.7 + 1.0/3 classes
= 0.9 or 90%

So now I have this:
Precision for model = 0.92, or 92%
Recall for model = 0.9, or 90%

F1 score is calculated by using this formula:

$$\textbf{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Math:
F1 = 2 * 0.92*0.9 / 0.92+0.9
F1 = 2 * 0.828 / 1.82
F1 = 1.66 / 1.82
F1 = 0.91, or 91%

The F1 score for this model is 91%.