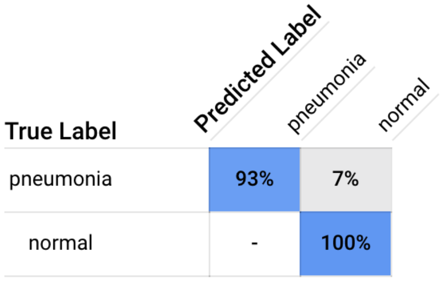
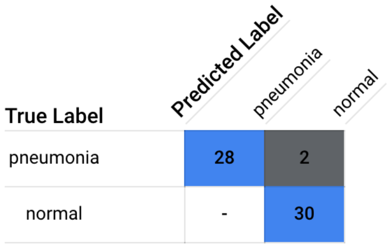
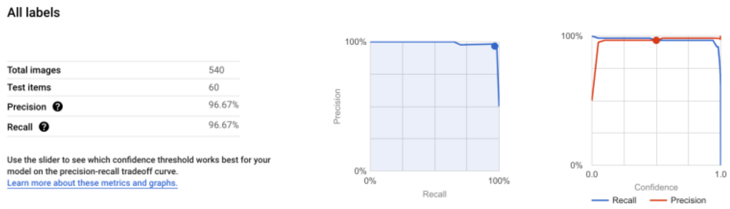
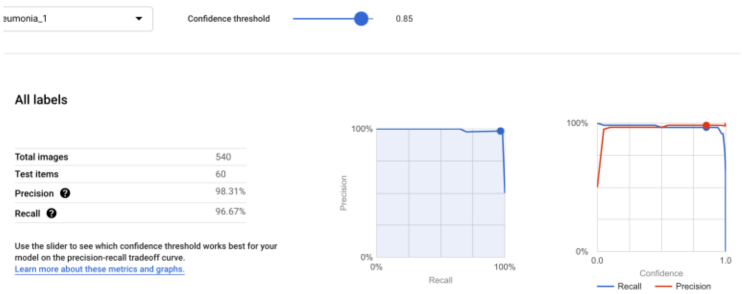
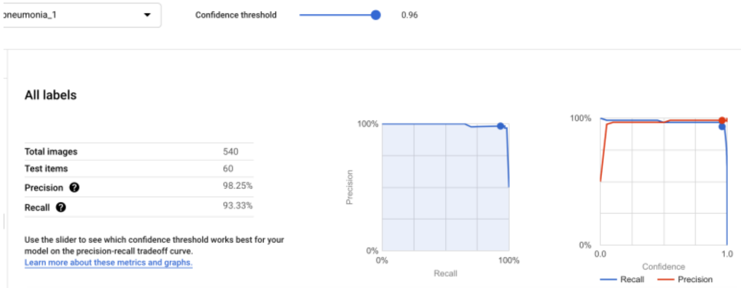
AutoML Modeling Report

Antonina Savka

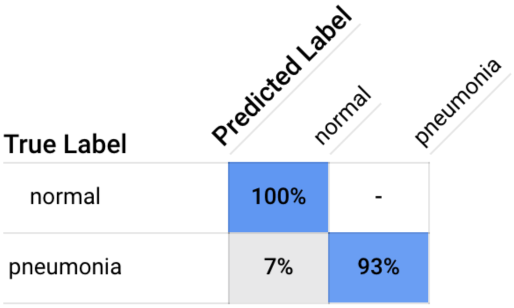
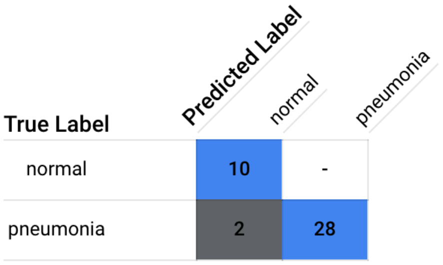
Binary Classifier with Clean/Balanced Data

<table>
<tr>
<td><b>Train/Test Split</b>
How much data was used for training? How much data was used for testing?</td>
<td>I used 300 images of normal x-ray and 300 images of pneumonia x-ray: 150 of viral and 150 of bacterial pneumonia scans.
AutoML split the dataset on 240 training images, 30 validation images and 30 test images for each label (normal and pneumonia)</td>
</tr>
<tr>
<td><b>Confusion Matrix</b>
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>
<table>
<tr><th>True Label</th><th colspan="2">Predicted Label</th></tr>
<tr><th></th><th>pneumonia</th><th>normal</th></tr>
<tr><td>pneumonia</td><td>93%</td><td>7%</td></tr>
<tr><td>normal</td><td>-</td><td>100%</td></tr>
</table>

<table>
<tr><th>True Label</th><th colspan="2">Predicted Label</th></tr>
<tr><th></th><th>pneumonia</th><th>normal</th></tr>
<tr><td>pneumonia</td><td>28</td><td>2</td></tr>
<tr><td>normal</td><td>-</td><td>30</td></tr>
</table>
Each cell in the confusion matrix tells us how often the model detected a label correctly (blue) and incorrectly (gray).
True positive rate for the "pneumonia" class is 93%
False positive rate for the "normal" class is 7%</td>
</tr>
<tr>
<td><b>Precision and Recall</b>
What does precision measure? What does recall measure? What precision and recall did the model</td>
<td>Precision measures percentage of correctly labeled data for specified class among all data labeled as this class (true positive + false positive).
Recall measures percentage of correctly labeled data for</td>
</tr>
</table>

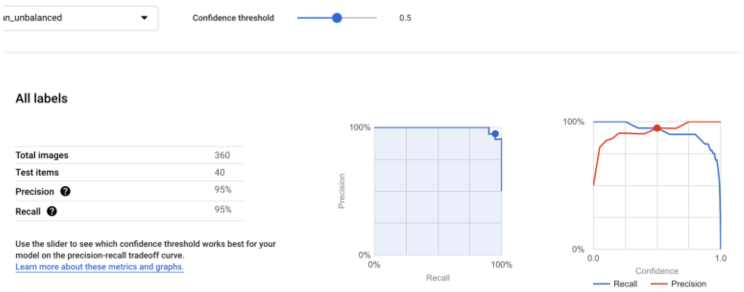
achieve (report the values for a score threshold of 0.5)?	specified class among all data in this class (true positive + false negative).  All labels Total images 540 Test items 60 Precision ❓ 96.67% Recall ❓ 96.67% Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve. Learn more about these metrics and graphs. 100% Precision 0% 0% 100% Recall 100% 0% 0.0 1.0 Confidence Recall Precision The model achieved 96.67% both for recall and precision.
Score Threshold When you increase the threshold what happens to precision? What happens to recall? Why?	 eumonia_1 ▾ Confidence threshold 0.85 All labels Total images 540 Test items 60 Precision ❓ 98.31% Recall ❓ 96.67% Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve. Learn more about these metrics and graphs. 100% Precision 0% 0% 100% Recall 100% 0% 0.0 1.0 Confidence Recall Precision By increasing threshold to e.g., 0.85 the precision number goes up to over 98% - this means the number of false positives is reduced.  pneumonia_1 ▾ Confidence threshold 0.96 All labels Total images 540 Test items 60 Precision ❓ 98.25% Recall ❓ 93.33% Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve. Learn more about these metrics and graphs. 100% Precision 0% 0% 100% Recall 100% 0% 0.0 1.0 Confidence Recall Precision Further increment of threshold slightly decreased precision (comparing to 0.85 threshold) and decreased the recall. Decreased recall means that we are getting more false negative. Threshold is a “confidence” with which the model classifies the data. From the graphs I can see that with demanding higher “confidence” from the model, we can get better precision but recalls drops dramatically when the threshold is close to 1. I suspect this means that the model will get a bias leaning to one of the labels (in our case of binary labeling). Why it happens: when data is clear (let’s say “data follows the rules”) then it is easy for the model to detect the label. The clear example here is “normal” label as all “normal” data show “dark” lungs. This gives us a high precision for “normal”. The data of the viral pneumonia (especially sever) is also “clear” – follows the rules (very high opacity). However, the bacterial pneumonia data

	falls somewhere on between normal and viral pneumonia and this is where the model lost its confidence. Because both bacterial and viral pneumonia are included under the same label, this gives us drop in recall when we increase the threshold.

Binary Classifier with Clean/Unbalanced Data

<table>
<tr>
<td><b>Train/Test Split</b>
How much data was used for training? How much data was used for testing?</td>
<td>Dataset of 100 normal and 300 pneumonia images (150 for each bacterial and viral pneumonia).
AutoML selected 80 training, 10 validation and 10 testing data for “normal” and 240 training, 30 validation and 30 testing data from “pneumonia”</td>
</tr>
<tr>
<td><b>Confusion Matrix</b>
How has the confusion matrix been affected by the unbalanced data? Include a screenshot of the new confusion matrix.</td>
<td>

<table>
<tr><th>True Label</th><th>Predicted Label
normal</th><th>pneumonia</th></tr>
<tr><td>normal</td><td>100%</td><td>-</td></tr>
<tr><td>pneumonia</td><td>7%</td><td>93%</td></tr>
</table>

<table>
<tr><th>True Label</th><th>Predicted Label
normal</th><th>pneumonia</th></tr>
<tr><td>normal</td><td>10</td><td>-</td></tr>
<tr><td>pneumonia</td><td>2</td><td>28</td></tr>
</table>
Disappointing for me – no visible influence on the confusion matrix. I expected to see the sign of a bias: more accurate prediction of pneumonia and at least a few false negative for normal. Not sure whether the result will get more visible bias if all pneumonia images are only bacterial (as some of them look quite close to normal).
Maybe increasing test data can reflect the bias. I find it dangerous because the bias may show itself quite late.
</td>
</tr>
</table>

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)?


ın_unbalanced

Confidence threshold 0.5

All labels

Total images	360
Test items	40
Precision	95%
Recall	95%

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

Comparing to the balanced data, I can see the difference around the recall and precision. Despite the number remains decent (95%, just 1% less than for balanced model), I see that the graphs are steeper: with increasing threshold we get visibly reduced recall.

Unbalanced Classes
From what you have observed, how do unbalanced classed affect a machine learning model?

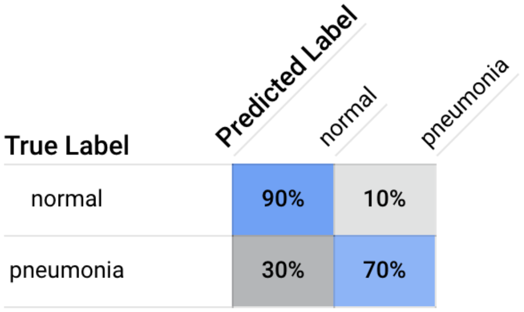
Unbalanced data causes the model loose the "confidence": it becomes more difficult to label the data correctly.
I would expect the bias: e.g., drop in recall for "normal" (many false negative) and high stable precision for "pneumonia". I can see this from graphs but not from the confusion matrix.

Binary Classifier with Dirty/Balanced Data

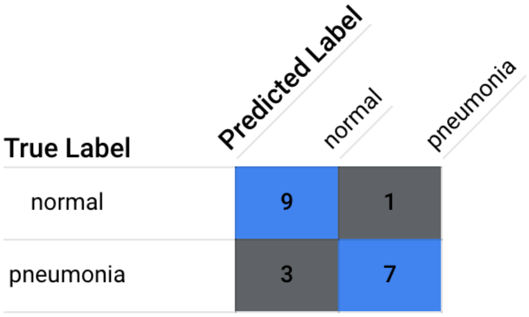
Confusion Matrix
How has the confusion matrix been affected by the dirty data? Include a screenshot of the new confusion matrix.

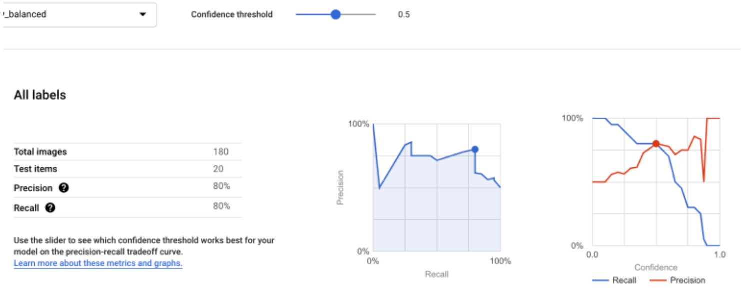
NOTE: I keep balanced mix of bacterial and viral pneumonia when using pneumonia images:

- Normal – 70 clean and 30 dirty data (15 viral and 15 bacterial)
- Pneumonia – 70 clean (35 viral and 35 bacterial) and 30 dirty data from "normal" set.



True Label	Predicted Label normal	pneumonia
normal	90%	10%
pneumonia	30%	70%

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

<table>
<tr><td>True Label</td><td>Predicted Label
normal</td><td>pneumonia</td></tr>
<tr><td>normal</td><td>9</td><td>1</td></tr>
<tr><td>pneumonia</td><td>3</td><td>7</td></tr>
</table>
Confusion matrix shows us that we get a significant drop in recall and precision comparing to clean balanced data. The model begins making more mistakes, increases the number of false positive.
</td>
</tr>
<tr>
<td>
<b>Precision and Recall</b>
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>
Dirty data leads to drop in precision and recall. The numbers drop to 80% (comparing to 96% for clean data).



_balanced

Confidence threshold 0.5

All labels

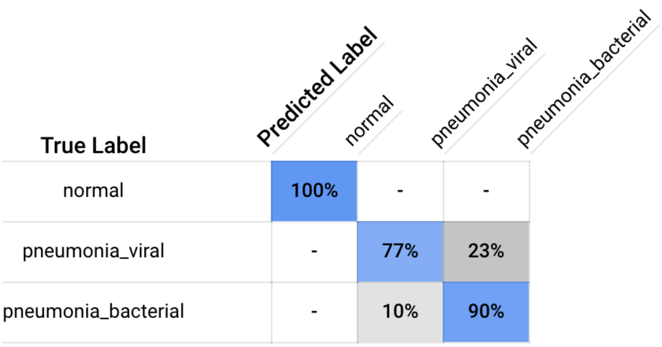
Total images	180
Test items	20
Precision	80%
Recall	80%

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

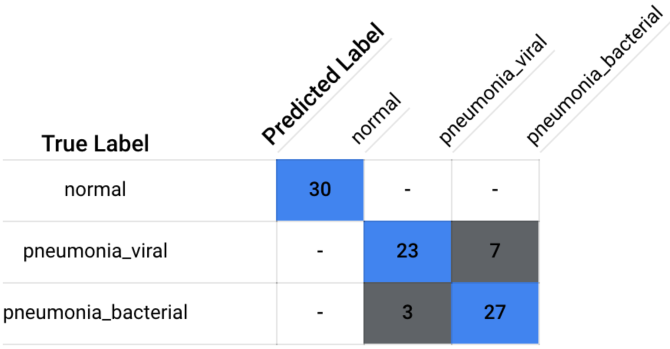
From the graphs we can see that with increasing threshold recall drops sharp and precision becomes less predictable.
Confusion matrix shows us:
<ul>
<li>highest precision is for “pneumonia” (0.87 comparing to 0.69 for “normal”)</li>
<li>highest recall is for “normal” class (0.9 comparing to 0.7 for pneumonia)</li>
</ul>
</td>
</tr>
<tr>
<td>
<b>Dirty Data</b>
From what you have observed, how does dirty data affect a machine learning model?
</td>
<td>
This one is a game changer: dirty data leads to lower precision, lower recall and quite unpredictable confidence of the model (comparing to clean data). The result is worse even comparing to the unbalanced data. It will be difficult to expect good results with threshold close to 0.85.
One more thing that I noticed is a bias leaning toward “normal”. I cannot tell that this is predictable bias: for me it is difficult to say which bias the model can get with dirty data – it depends on the nature of this data and how “dirty” they are; but it is definitely a side-effect.
</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? Why might you do to try to remedy the model's "confusion"? Include a screenshot of the new confusion matrix.



True Label \ Predicted Label	normal	pneumonia_viral	pneumonia_bacterial
normal	100%	-	-
pneumonia_viral	-	77%	23%
pneumonia_bacterial	-	10%	90%

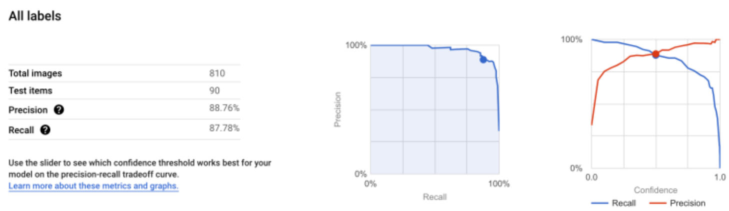


True Label \ Predicted Label	normal	pneumonia_viral	pneumonia_bacterial
normal	30	-	-
pneumonia_viral	-	23	7
pneumonia_bacterial	-	3	27

Confusion matrix for 3 classes shows us perfect precision and recall for "normal" label and that the model is able to differentiate between the "normal" x-ray and pneumonia – in this term results are better than for clean balanced binary class dataset. However, when it comes to the difference between the type of pneumonia, we see that the model gets confused. Despite the data set was balanced (300 images for normal, viral and bacterial classes each), we can see a slight bias toward bacterial pneumonia (34 images were called "pneumonia_bacterial").

The classes the most likely to be confused:
"pneumonia_viral" and "pneumonia_bacterial"
The class the most likely to be right: "normal".

To reduce the model's confusion in this case I would probably increase the number training data in balanced way. (I see this as a weak side of my knowledge and would like to get some advice.)

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)?	 All labels Total images 810 Test items 90 Precision 88.76% Recall 87.78% Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve. Learn more about these metrics and graphs. 100% Precision 0% 0% 100% Recall 100% 0% 0.0 1.0 Confidence Recall Precision The model's precision value is 88.76% and the model's recall value is 87.78% - both are quite lower than for binary model. The model's precision calculation: (30/30 + 23/26 + 27/34) / 3 = (1 + 0.88 + 0.79) / 3 = 0.89 The model's recall calculation: (30/30 + 23/30 + 27/30) / 3 = (1 + 0.76 + 0.9) / 3 = 0.88
F1 Score What is this model's F1 score?	Using data from the screenshot above: F1 = 2 * 0.8876 * 0.8778 / (0.8876 + 0.8778) = 1.5582 / 1.7654 = 0.88