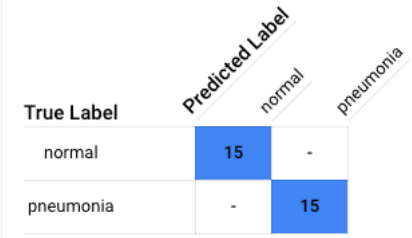
AutoML Modeling Report

Alex

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>150 for normal, of which 120 for training, 15 for validation, 15 for test
149 for pneumonia, 15 for validation, 15 for test</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>I think in this dataset, the confusion matrix is flipped – it would be better to have the true positives in the upper left corner, so I’m putting that here:
<table>
<tr><td></td><td>Pneumonia</td><td>Normal</td></tr>
<tr><td>Pneumonia</td><td>15</td><td>-</td></tr>
<tr><td>Normal</td><td>-</td><td>15</td></tr>
</table>
(this is the original I got from google:
 

)

To answer the questions:
I see a total of 30 as the sum of all cells. This is the total size of the dataset (which matches the label stats from the first question above).
What we see is that in the False Negative and False Positive cells the value is 0 – this is great, it means we have no wrong predictions (for both classes even).
Hence, the True Positives and True Negatives are at a 100% each (15 normal and 15 pneumonia images, which were used for testing).</td>
</tr>
</table>

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 and Recall are both at 100%.
Precision is the ratio of actual positive cases to all positive predictions (including the false positives). So, a low rate of precision would mean that some patients that were classified as pneumonia cases were actually healthy patients. The formula is: TP / (TP + FP)

Recall is the ratio of actual positive cases to actual positives. This answers the question 'how many cases slipped through that were actually positives? For our model, a low value would mean that some pneumonia cases were labeled as healthy cases – which is more problematic than having a low precision, so I guess this model should be optimized for a high recall rate.
The formula for recall is: TP / (TP + FN)

Score Threshold
When you increase the threshold what happens to precision? What happens to recall? Why?

A low confidence rate will mean a lower precision. Why is that? If the confidence is lower, it means that an image will more likely be classified as positive (pneumonia). So, looking at the confidence matrix, this would raise the number of false positives. Our matrix would start to look like this in an extreme case of confidence = 0:

	Pneumonia	Normal
Pneumonia	15	-
Normal	15	-

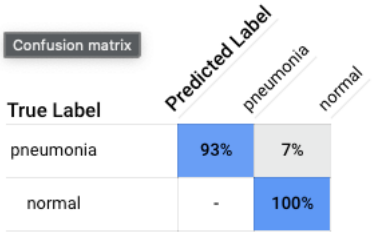
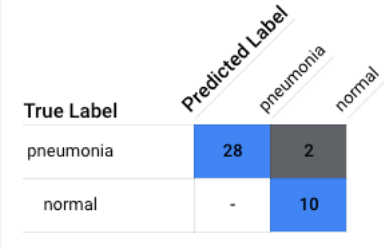
We'll then get a 50% precision – which makes sense, because we would then just accept every picture as a pneumonia case, and still get half of them right. It would also render the model useless though.

In turn, a high confidence level leads to a lower recall. Why is that? If the confidence level is high, it means it's hard for an image to get accepted as a pneumonia case. Hence, there will be more images that are false negatives.
So, our table might look more like this for a higher confidence rate:

	Pneumonia	Normal
Pneumonia	15	10
Normal	-	5

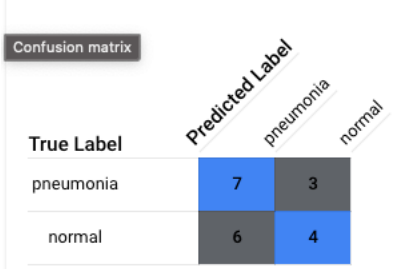
That would be a 15 / (15 + 10) = 0,6 recall value. Many patients who are ill would be ignored, so it's not a good idea to set the confidence level very high.

Binary Classifier with Clean/Unbalanced Data

Train/Test Split How much data was used for training? How much data was used for testing?	100 images for 'normal', of which 10 were for validation and 10 for test 298 images for pneumonia, of which 238 for training, 30 for validation, 30 for test
Confusion Matrix How has the confusion matrix been affected by the unbalanced data? Include a screenshot of the new confusion matrix.	This is the new confusion matrix:  In numbers:  We see that some false negatives have appeared!
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)?	It says 95% for both precision and recall. I thought it was strange, because when I calculate them from the numbers in the table, I get: Precision = 28 / (28 + 0) = 1 Recall = 28 / (28 + 2) = 0,93 I haven't found any explanations on why this happens in the forums. There are lots of posts, but they are either about the 3-class case or don't provide an explanation of why this happens. When I flip the matrix and do a calculation of the values for the minority class, I get Precision = 10/12 = 0,83 Recall = 10/10 = 1

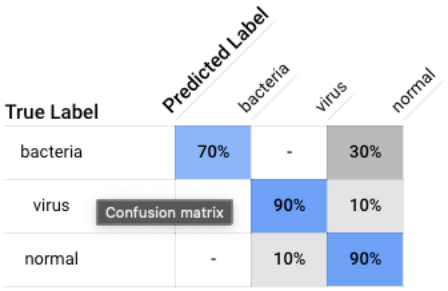
Unbalanced Classes From what you have observed, how do unbalanced classed affect a machine learning model?	Unbalanced classes influence the model in a negative way, as we can see in precision and recall dropping. We have far more images for pneumonia cases, so I would have assumed that there would be misclassifications for the normal cases, not for pneumonia. Because my assumption would be that through the larger number of images, the model would be more 'skilled' in choosing pneumonia cases correctly. On the other hand, we had many more test images for pneumonia than for normal (30:10). So, a reason for having more misclassifications for pneumonia images could be that we did three times more tests than with normal images.

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.	 We see that the matrix now shows large numbers in the false negatives and false positives cells. This is already an indicator that the model has not been trained very well on the given dataset. This was expected – the 'dirty' data makes it very hard for the model to learn.
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?	Precision and Recall dropped down to 55% each – which makes sense, because via the 'wrong' images that were mixed in, it's very hard to train the model accurately. Of the binary classifiers, the clean/balanced ones has the highest recall and precision. From the above matrix, precision and recall are: Precision = 7 / (7+6) = 0,54 Recall = 7 / (7+3) = 0,7 Again, these values are different from what google tells

	me.
Dirty Data From what you have observed, how does dirty data affect a machine learning model?	Dirty data makes it harder to train a model accurately. Precision and recall drop through dirty data – hence I suggest that when thinking about creating a machine learning model, a lot of care has to be taken about the data being properly managed in the first place (I've heard rumors from data professionals that grooming their data is the largest part of their job).

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.	Here's the confusion matrix:  It works the same way as the 2-class matrix, but with three classes. In the top left cell (70% here), we still have the true positives, and on the diagonal axis from that cell we have the true negatives. Above and below that diagonal axis, we have the false negatives and false positives – so for the false predictions, we have more cells than in the 2-class matrix. ~~Which classes is the model most likely to get right? That depends on what 'getting it right' means.~~ Question to the reviewer: Is the calculation of precision and recall correct? I've calculated the precision and recall for each class: | | Precision | Recall | |---|---|---| | Bacteria | 100 | 70 | | Virus | 90 | 90 | | Normal | 69 | 90 | If we would average precision and recall, then the bacterial class would be most likely to be recognized correctly. On the other hand, if recall is what matters

	most, then the model would perform best for the virus and normal classes. To answer the question: Which class is easily gotten right, and which is confused? The virus and normal classes score at 90% true positives, so they are gotten right quite easily. The bacteria class only scores at 70% percent, which is not as good as the other two classes. I'd rather have expected the normal class to score at the top, and the bacteria and virus classes to both score lower – because it's much harder to distinguish between the two. What can we do to remedy the confusion? We could include more images, especially for the bacteria and virus classes. If we'd really want to dig deeper, we'd need to have a close look at the quality of the images. Are the images very 'diverse' in a sense that they show a wide range of different features that might influence a decision? Data quality is important here.
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)?	Google returns a recall and precision of 83.33%. According to https://cloud.google.com/automl-tables/docs/evaluate#micro-average, the calculation is done via micro-averaging. So, our precision would be: (7+9+9) / (7+9+9 + 1+1+3) + 83.33% And the recall (same actually): (7+9+9) / (7+9+9 + 1+ 1+ 3) = 83.33%
F1 Score What is this model's F1 score?	The F1 score is calculated as 2 * ((precision * recall) / (precision + recall)) So: 2 * 0.83 * 0.83 / (0.83 + 0.83) = 0,83 As I understand, a higher F1 score is better, so this model seems to be performing not so bad.