AutoML Modeling Report

Svajune Klimasauskaite

Binary Classifier with Clean/Balanced Data

Train/Test Split

How much data was used for training? How much data was used for testing?

The AutoML Vision uses the 80% of your content documents for training, 10% for validating, and 10% for testing.

- **Train** - Use the image to train the model.
- **Validation** - Use the image to validate the results that the model returns during training.
- **Test** - Use the image to verify the model's results after the model has been trained.

AutoML Vision automatically places images it in one of the three sets to ensure that there is enough training, validation, and testing content.

Here was the status for my dataset:

Training images	147
Validation images	25
Test images	28

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?

Confusion matrix represents the percentage of times each label was predicted for each label in the training set during evaluation.

True label \ Predicted label	Pneumonia	Normal
Pneumonia	76.9%	23.1%
Normal	12.5%	87.5%

- Within Pneumonia Images, the model labeled as Pneumonia (**true positive**) on the **76.9 %** of the cases, however on the **23.1** cases the model failed to label correctly (**false negative**).
- Within Healthy Images, the model labeled as Healthy (**false positive**) on the **87.5 %** of the cases, however on the **12.5** cases the model labeled as Pneumonia (**true negatives**).

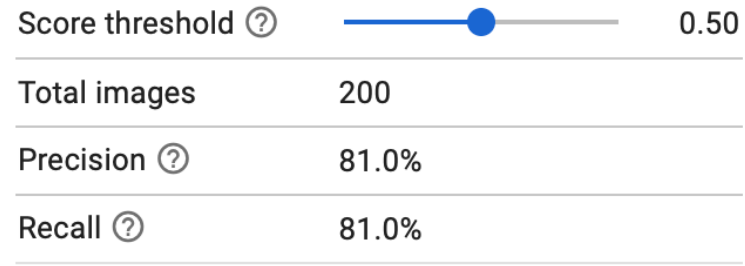
Precision & Recall

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 just different metrics for measuring the "success" or performance of a trained model.

- **precision** is defined as the number of true positives over all positives, and will be the higher when the amount of false positives is low.
- **recall** is defined as the number of true positives over true positives plus false negatives and will be higher when the number of false negatives is low.

Both take into account true positives and will be higher for high, positive accuracy, too.



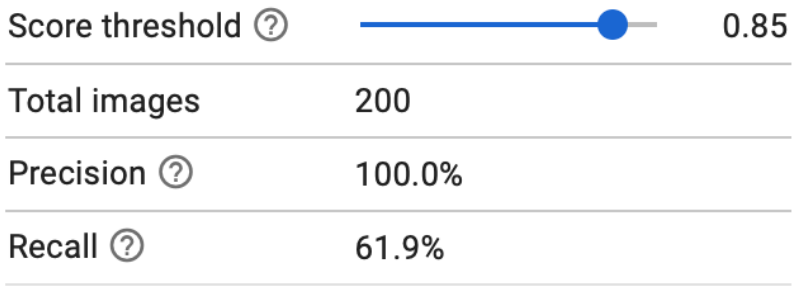
Score Threshold

When you increase the score threshold, what happens to precision? What happens to recall? Why?

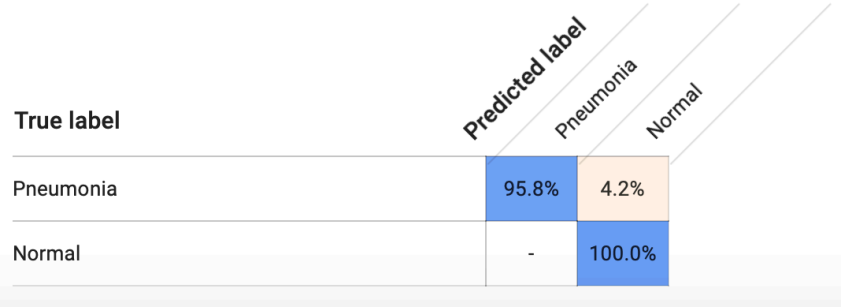
The score threshold refers to the level of confidence the model must have to assign a Pneumonia case.

- If the score threshold is low, the model classifies more images, but runs the risk of misclassifying a few images in the process.
- If the score threshold is high, the model classifies fewer images, but it will have a lower risk of misclassifying images.

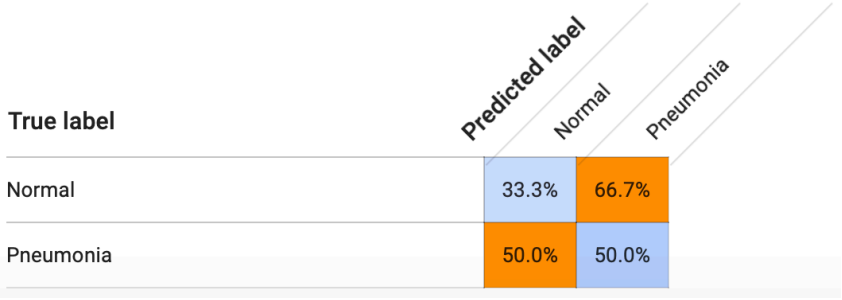
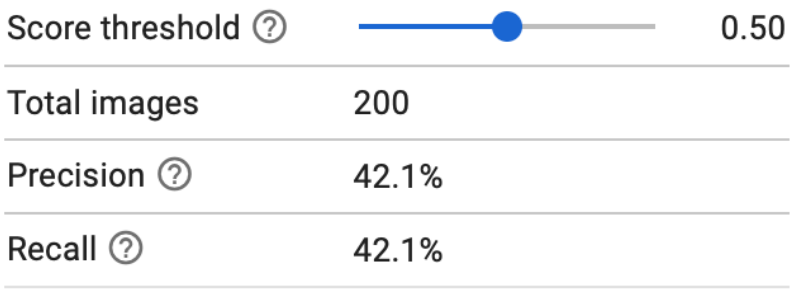
In this case, we want to ensure that Pneumonia case will not get labelled as Healthy, therefore we should be interested in lower confidence level and higher recall.



Binary Classifier with Clean/Unbalanced Data

Train/Test Split How much data was used for training? How much data was used for testing?	**Training images** 318 **Validation images** 42 **Test images** 39
Confusion Matrix How has the confusion matrix been affected by the unbalanced data? Include a screenshot of the new confusion matrix.	Healthy cases had 100% success, while Pneumonia images had 4.2 % false negative. This way of unbalanced data showed a good output and it seems that it didn't bias Normal images to lean towards Pneumonia label. However, I assume the opposite training scenario, i.e. with 300 Normal images and 100 Pneumonia, would give completely opposite results. 
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.)	The machine learns very well when we increased Pneumonia training data by 300%. Score threshold 0.50 Total images 399 Precision 96.9% Recall 96.9%
Unbalanced Classes From what you've observed, how do unbalanced classes affect a machine learning model?	The machine gave a better output when we increased Pneumonia training data by 300%. Could be that Pneumonia images are much more complex than a Normal ones, therefore the training, with all kind of Pneumonia variations, gave such a good output. Even though the bias helps accuracy it might not be suitable for a practical application

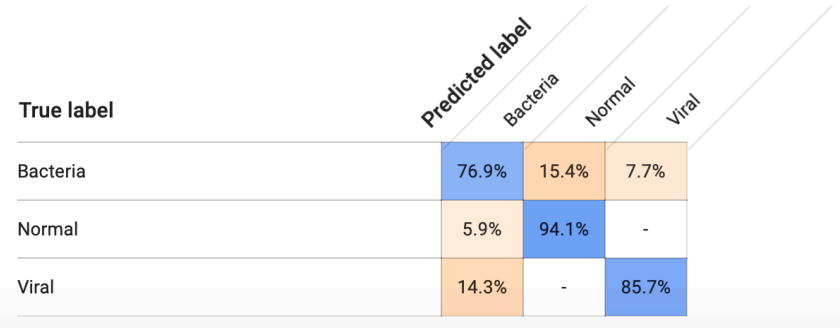
Binary Classifier with Dirty/Balanced Data

Confusion Matrix How has the confusion matrix been affected by the dirty data? Include a screenshot of the new confusion matrix.	The confusion matrix says it all - machine didn't learn at all. 
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?	
Dirty Data From what you've observed, how do dirty data affect a machine learning model?	It shows the implication of having 30% dirty data. Machine didn't learn anything from this dataset.

3-Class Model

Confusion Matrix

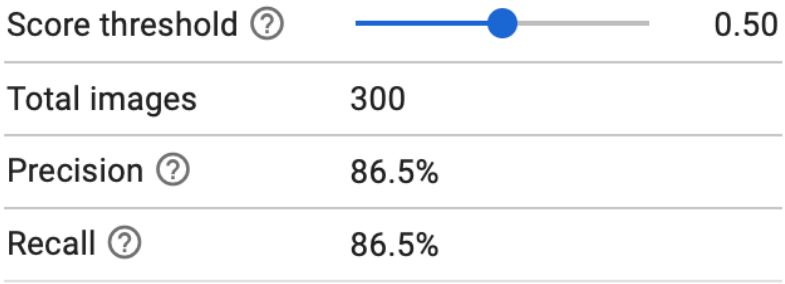
Summarize the 3-class confusion matrix. What classes are the model most likely to confuse? What class(es) is the model most likely to get right? What might you do to try to remedy the model's "confusion"? Include a screenshot of the new confusion matrix.



- Class Normal has the highest chances to get labelled correctly. It got confused only with the Pneumonia - Bacteria by 5.9%.
- Class Viral has 85.7% chances to get it right and got confused with Pneumonia - Bacteria by 14.3%.
- Class Bacteria is most likely to get confused with Normal or Viral.

Precision & Recall

What are the model's precision and recall? How are these values calculated? (Report the values fora score threshold of 0.5.)



- First, the recall and precision are calculated for each class based on the predicted labels (absolute numbers are taken from the confusion matrix).
- When we have the calculation for each class, then we calculate the average to get the precision/recall of the whole model.

F1 Score

What is this model's F1 score?

F1 measures how well the model performs across all score thresholds. In AutoML Vision, this metric is called Average Precision and is equal to **0.972**.