AutoML Modeling Report

Jimmey Jose

Binary Classifier with Clean/Balanced Data

Train/Test Split

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

A total of 200 x-rays were used, of which 180 was used for training and 20 for testing. The exact breakup based on type of x-rays is given below:

Break Up of Data Used

	Training	Testing	**Total**
Normal chest x-rays	90	10	**100**
Pneumonia chest x-rays	90	10	**100**
Total	**180**	**20**	**200**

Confusion Matrix

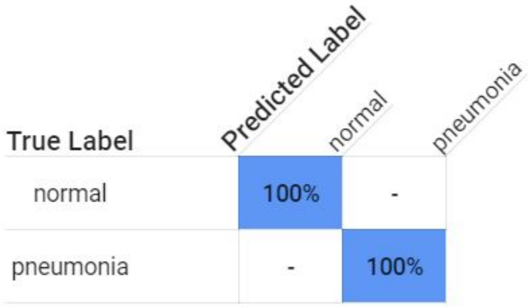
What do each of the cells in the confusion matrix describe? What values did you observe (include a screenshot)? What is the true positive rate for the “pneumonia” class? What is the false positive rate for the “normal” class?

Each cell in a confusion matrix captures the number of predictions that were either correct or wrong when testing an ML model.

The four different types of values in a confusion matrix from the perspective of this binary class model are described below. (‘pneumonia’ is being considered as the positive class for the purpose of the below descriptions.)

1. True Positives - Test cases in which the true label is ‘pneumonia’ and is correctly predicted to be the same.
2. False Positive - Test cases in which the true label is ‘normal’ and is predicted wrongly to be ‘pneumonia’
3. True Negatives -Test cases in which the true label is ‘normal’ and is predicted to be the same
4. False Negatives - Test cases in which the true label is ‘normal’ and is predicted to ‘pneumonia’

Confusion matrix for this model



True Label	Predicted Label normal	pneumonia
normal	100%	-
pneumonia	-	100%

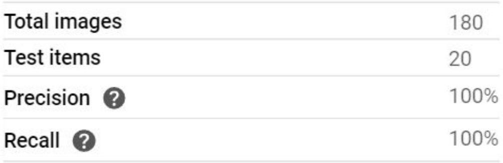
- True positive rate for “pneumonia” class is 100 %
- False positive rate for “normal” class is 0 %

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 measures the ratio between the number of true positives and the number of predicted positives(True positives + False positives). This value gives a confidence level about how accurate true positive outcomes are.
- Recall measures the ratio between the number of true positives and the number of actual positives(True positives + False Negatives). This value gives a confidence level about how many of the true positives were identified.

The precision and recall were both 100 % at a threshold of 0.5 is this model.



Total images	180
Test items	20
Precision ?	100%
Recall ?	100%

Note: This Precision and recall are calculated using Micro-averaged method.

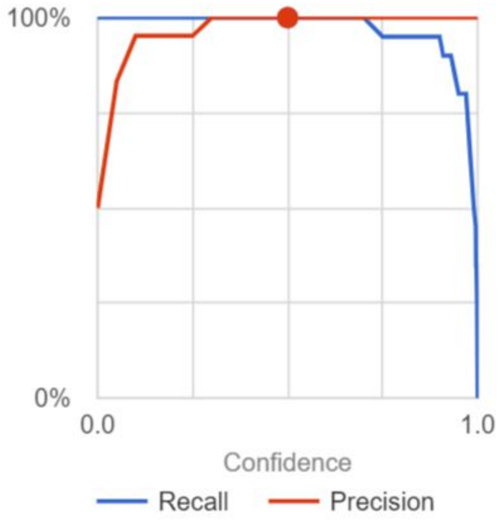
Score Threshold

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

There will be a trade-off between Precision and Recall in almost all models unless the model is perfect (which generally doesn't happen). In this context higher precision leads to lower recall as the threshold increases and a higher recall will lead to lower precision as threshold decreases.

With a low threshold, a model will classify everything even though the risk of getting it wrong is high; while with a high threshold a model might not classify anything because of the risk of getting it wrong.

In this model the precision and recall are 100 % when the threshold is in between .3 and .7; below .3, precision will drop and above .7, recall will drop.



Binary Classifier with Clean/Unbalanced Data

Train/Test Split

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

A total of 400 x-rays were used, of which 360 was used for training and 40 for testing. The exact breakup based on type of x-rays is given below:

Break Up of Data Used

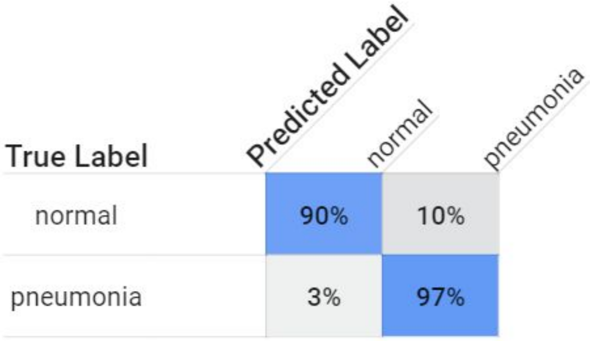
	Training	Testing	**Total**
Normal chest x-rays	90	10	**100**
Pneumonia chest x-rays	270	30	**300**
Total	**360**	**40**	**400**

Confusion Matrix

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

The unbalanced data has resulted in higher accuracy for 'pneumonia' class which had the bigger data set. The confusion matrix shows the same through the following.

1. The true positive rate is higher by 7 % for 'pneumonia'. If you consider the size of the test data set, you will observe that for both classes,the model inaccurately predicted 1 record each which doesn't seem high; but if the test data were more equally distributed we may have seen a bigger true positive rate for 'pneumonia'
2. 'Pneumonia' class has a higher false positive of 10 % compared to 3 % for 'normal class which shows that there is a higher chance of being classified in the 'pneumonia' class

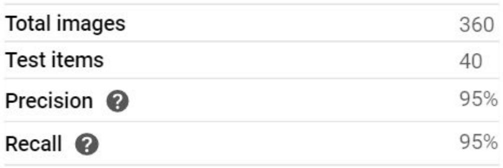


True Label	Predicted Label: normal	pneumonia
normal	90%	10%
pneumonia	3%	97%

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 precision and recall at threshold of .5 are both 95%. The unbalanced data seems to give the same 95% precision and recall within a wide threshold between 0 .03 and 0.99 which appears to be very skewed and unreliable.



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

Note: The above Precision and recall is calculated using micro-averaged method.

Unbalanced Classes

From what you’ve observed, how do unbalanced classes affect a machine learning model?

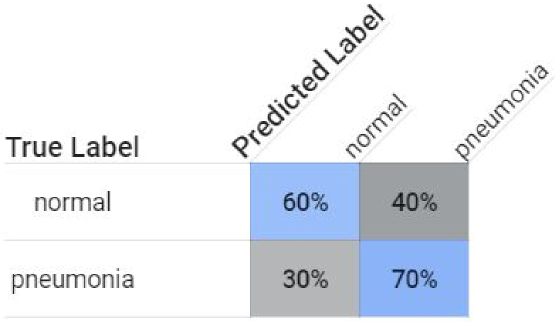
1. Unbalanced classes seem to skew predictions to the class with a larger data set which is ‘pneumonia in this case
2. In this model, the precision and recall for ‘pneumonia’ is 96.67 each and for ‘normal’ is 90% each; this is a difference of only around 6% and for that reason, the combined precision and recall of 95% is accurate. There is generally a high chance to have a sizeable difference between the individual precision and recall because of unbalanced data which can make the model unreliable.

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 dirty data has brought down the True Positives and increased the False Positives substantially compared to the first model which had the same amount of data. This huge drop in reliability of data is because the dirty data resulted in the model training on the wrong parameters and noise.

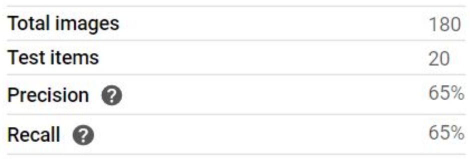


True Label	Predicted Label normal	pneumonia
normal	60%	40%
pneumonia	30%	70%

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 reca*ll?*

The precision and recall has dropped from 100% each in the first model with the same amount of data to 65% each at a threshold of .5 for this model because of the dirty data.



Total images	180
Test items	20
Precision ?	65%
Recall ?	65%

- The first model with Clean/Balanced data has the highest precision and recall of 100% each at a threshold of .5 among the binary classifiers

Summary of Precision and Recall in Binary Classifier Models

Binary Classifier	Precision	Recall
1. Clean/ Balanced Data	100 %	100%
2. Clean/Unbalanced Data	95%	95%
3. Dirty/Balanced Data	65%	65%

Note: The above Precision and recall is calculated using micro-averaged method.

Dirty Data

From what you've observed, how do dirty data affect a machine learning model?

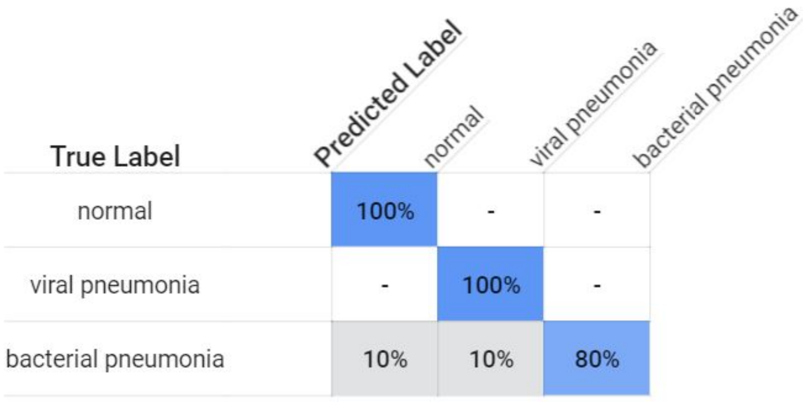
1. Dirty data confuses the machine learning model resulting in a substantial decrease in accuracy, precision and recall
2. The ML model may have fitted in more of the noise in the data making predictions unreliable
3. Equal amount of dirty data in each class does not result in correct output in the models. This model had an equal amount of bad data in both classes, but there is no relation between the precision and recall for each class

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.

This model is most likely to confuse the 'bacterial pneumonia' class and get 'normal' and 'viral pneumonia' classes correct as seen in the confusion matrix.



True Label	Predicted Label normal	viral pneumonia	bacterial pneumonia
normal	100%	-	-
viral pneumonia	-	100%	-
bacterial pneumonia	10%	10%	80%

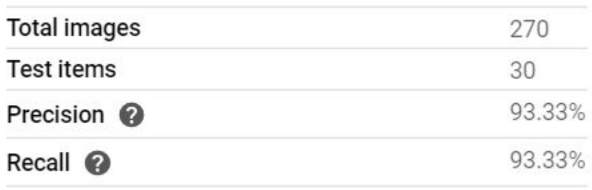
The following can be done to improve the model's confusion:

1. Use more data with all 3 sets
2. Improve quality of data through preprocessing, augmentation or any other technique
3. Look for opportunities to adjust the threshold levels to improve results
4. Look for opportunities to change or improve the modeling algorithm to improve results

Precision & Recall

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

This models precision and recall are 93.33% each at a threshold of 0.5



Total images	270
Test items	30
Precision ?	93.33%
Recall ?	93.33%

Note: The above Precision and recall is calculated using micro-averaged method.

There are 2 main ways to calculate Precision and Recall.

1. Precision and Recall for each class
2. Combined Precision and Recall using micro-averaged method

Precision and Recall for each classifier

1. 'Normal' Class Precision and Recall calculation

   True Positives (TP) = 10
   True Positives + False Positives (TP+FP) = 10+1+0 = 11
   True Positives + True Negatives(TP+TN) = 10+0+0 =10

   Precision = TP / (TP+FP) = 10 / 11 = 90.91 %
   Recall = TP/ (TP+FN) = 10/10 = 100 %

2. 'Viral pneumonia' Class Precision and Recall calculation

   True Positives (TP) = 10
   True Positives + False Positives (TP+FP) = 10+1+0 = 11
   True Positives + True Negatives(TP+TN) = 10+0+0 = 10

   Precision = TP / (TP+FP) = 10 / 11 = 90.91 %
   Recall = TP/ (TP+FN) = 10/10 = 100 %

3. 'Bacterial pneumonia' Class Precision and Recall calculation

   True Positives (TP) = 8
   True Positives + False Positives (TP+FP) = 8+0+0 = 8
   True Positives + True Negatives(TP+TN) = 8+1+1 = 10

   Precision = TP / (TP+FP) = 8 / 8 = 100 %
   Recall = TP/ (TP+FN) = 8/10 = 80 %

Precision and Recall using micro-averaged method

True Positives (TP) = 10 + 10 + 8 =28
True Positives + False Positives (TP+FP) =11+11 + 8 = 30
True Positives + True Negatives(TP+TN) =10+10+10 = 30

Precision = TP / (TP+FP) = 28 / 30 = 93.33%
Recall = TP/ (TP+FN)= 28 / 30 = 93.33%

F1 Score

What is this model's F1 score?

F1 scores can be calculated for each classifier and combined using multiple ways. Below are F1 scores by classifier and combined F1 scores in 2 popular different methods.

Formula: F1 Score = 2 *(Precision * Recall)/ (Precision + Recall)

F1 Score by Class

'Normal' class F1 Score = 2 * (90.91 %* 100%)/ (90.91%+100%) = **95.24%**
'Viral pneumonia' class F1 Score = 2 * (90.91 %* 100%)/ (90.91%+100%) = **95.24%**
'Bacterial pneumonia' class F1 Score = 2 * (100 %* 80%)/ (100%+80) = **88.89%**

Macr0-averaged F1 Score= (95.24%+95.24%+88.89%)/ 3= **93.12%**

	Micro-averaged F1 Scoro=Micro-avg. precision/recall l= **93.33%**