AutoML Modeling Report

As next part of Medical project – Identification cases of pneumonia in children I built and trained a classification model. There were two steps

1. built a model using Google's AutoML Vision platform, and
2. explained how properties of data impact performance of models.

Binary Classifier with Clean/Balanced Data

**Train/Test Split**

Graphical user interface, text, application, email

Description automatically generated

**Confusion Matrix**

Confusion matrix describes the performance of a classification model. Here is a table that explains how to read a confusion matrix for 2 class model:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | predicted label | | |
|  |  | Pneumonia | | Normal |
| true label | Pn | Correct  TP for Pn | Wrong  FP for No | |
| No | Wrong  FP for Pn | | Correct  TP for No |

Pn – pneumonia, No – normal

Here is the confusion matrix for our binary classifier

A picture containing diagram

Description automatically generated

True Positive for “pneumonia” class is 100%

False Positive “normal” class is 0

**Precision and Recall**

Graphical user interface, chart

Description automatically generated

Precision is the ratio of what our model predicted correctly to what our model predicted. In other words, Precision measures how valid results are.

And it is used where the cost of getting a prediction wrong is much higher than the cost of missing out on the right prediction.

Recall is the ratio of what our model predicted correctly to what the actual labels are.

Recall measures how complete results are, how well model detect the class. It is used when the cost of missing a prediction is much higher than a wrong prediction.

**Score Threshold**

Chart

Description automatically generated

When we increase the threshold, recall will be lower as the model will classify fewer images. Precision gets higher as the model will have a lower risk of misclassifying images.

Binary Classifier with Clean/Unbalanced Data

**Train/Test Split**

For the test I used images:

“normal” 105 total | 84 for training | 10 for testing

“pneumonia” 305 total | 245 for training |30 for testing

**Confusion Matrix**

Chart, waterfall chart

Description automatically generated

The high values on the diagonal shows that labels have been identified correctly. The decreased value of “normal” gives a sign that the model is misclassifying these images. In unbalanced dataset the diagonal value of the class, that has less samples, is lower than the value of overrepresented class. In our case label “normal” has less samples (105 images) in dataset and less TP score (80%) in compare to “pneumonia” label (305 images).

**Precision and Recall**

Graphical user interface

Description automatically generated

Both Precision and Recall are lower for the unbalanced class as the model misclassified underrepresented “normal” images.

**Unbalanced Classes**

The model is to be biased towards the label with more images (in our case “pneumonia”). In general model that does not learn what makes the “normal” class “different” may fail to distinguish classes.

To solve the problem, we need to balance the training data set.

Binary Classifier with Dirty/Balanced Data

**Confusion Matrix**

Chart, waterfall chart

Description automatically generated

Confusion matrix has been affected by dirty data and shows lower values on the diagonal, it means model misclassified images for both classes, so that 30% of images identified false positive.

**Precision and Recall**

Chart, line chart

Description automatically generated

Precision and recall for the case with dirty data are the lowest among these three binary classifiers.

The highest precision was for the balanced dataset with both confidence thresholds. The highest recall was observed for the balanced dataset with confidence threshold = 0.5

**Dirty Data**

Performance of the model degraded with dirty data. We added 30% of noise data and metrics dropped 15% down. But the model is still quite robust, and we observe both precision and recall as high as 85%.

3-Class Model

**Confusion Matrix**

Graphical user interface, application

Description automatically generated

Table

Description automatically generated with medium confidence

Unlike binary classification, 3 class confusion matrix has no positive or negative classes. Values on the diagonal, intersection of predicted and true) are true positive for the related class. True negative value is not so obvious and calculated from the confusion matrix.

In my 3-class model there are two classes “normal” and “bacterial pneumonia” that are most likely to confuse. While “viral pneumonia” is most likely to get right.

I tried to remedy model’s confusion by adding additional images for training of the model.



Here is a new confusion matrix

Table

Description automatically generated

The metric True positive for “viral pneumonia” class decreased while other diagonal values for “normal” and “bacterial” increased.

**Precision and Recall**

Chart, line chart

Description automatically generated

Precision is what the model predicted correctly, for example as ‘normal’, divided by what the model predicted as ‘normal’.

Recall is the ratio of what our model predicted correctly to what the actual labels are.

There are two ways to calculate the average precision and recall of the model that has few labels

1. Micro-average method is to sum up all individual True Positives, False Positives and False Negatives and then apply then to standard formulas.
2. Macro- average method is to take average of precision and recall of all system labels.

|  |  |  |  |
| --- | --- | --- | --- |
| LABEL | PRECISION | RECALL | F1-SCORE |
|  |  |  |  |
| NORMAL | 0.80 | 0.8 | 0.8 |
| VIRAL | 0.83 | 1 | 0.91 |
| BACTERIAL | 0.88 | 0.7 | 0.78 |
| **MODEL micro\_av** | **0.84** | **0.83** | **0.83** |
| **MODEL macro\_av** | **0.83** | **0.83** | **0.83** |

Precision and recall of each label are calculated based on the system confusion matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | predicted label | | |
|  | | normal | viral | bacterial |
| true label | normal | **80%** | **10%** | **10%** |
| viral | **0** | **100%** | **0** |
| bacterial | **20%** | **10%** | **70%** |

Precision of label ‘normal’ is calculated as 80% / (80%+0+20%) = 0.80

Precision of ‘bacterial pneumonia’ is calculated 70% / (10%+0+70%) = 0.88

Recall:

Recall of ‘normal’ is calculated as 80% / (80%+10%+10%) = 0.80

Recall of ‘bacterial pneumonia’ is calculated as 70% / (70%+10%+20%) = 0.70

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | predicted label | | |
|  |  | normal | viral | bacterial |
| true label | normal | **80%** | **10%** | **10%** |
| viral | **0** | **100%** | **0** |
| bacterial | **20%** | **10%** | **70%** |

**F1 Score**

F1 score is the harmonic mean of recall and precision. It balances precision and recall score, and we get a single metric.

Model’s F1 score is 0.83