AI Product Management: Google AutoML Modeling Report

Report

Ben Jacobson
April 2020

Overview

Project Goal

In this project, we create four separate machine learning models using Google's AutoML Vision App. The models will be set with different underlying inputs or outputs, and we will review how this affects our evaluation metrics. In particular, we will create:

(1) a binary classifier to detect pneumonia using chest x-rays
(2) an unbalanced binary classifier
(3) a binary classifier with dirty data
(4) a three-class model with the classes "normal", "bacterial pneumonia", and "viral pneumonia"

Evaluation Metrics

We evaluate how these underlying conditions affect our metrics, namely precision and recall. We also discuss aspects such as the confusion matrix, thresholds, confidence and F1 score.

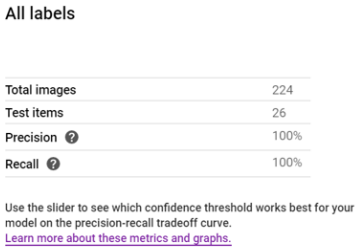
Methodologies

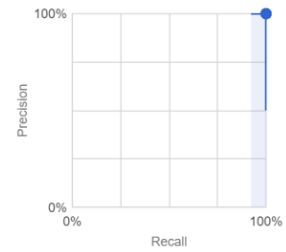
All models are trained through the Vision App on Google AutoML platform. The data provided is from the Kaggle X-ray competition is listed below in resources. The basic workflow includes:

1. Creating the data directory from the Kaggle data, for above models.
2. Upload labeled data to Google Cloud Storage
3. Train the model
4. Evaluate (As shown in Results and Analysis)
5. Deploy (make test predictions)

Results & Analysis

1. Binary Classifier with Clean/Balanced Data





Train/Test Split How much data was used for training? How much data was used for testing?	Each class (normal, pneumonia) had 125 in each class (250 total). Of this, there was a 90/10 split for training/test data, respectively. As a result, 13 of each class were used for testing (26 total).
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?	The confusion matrix shows the predicted labels and the actual labels for both the pneumonia and normal classes. It shows True Positive, True Negative, False Positive, and False Negative. Both classes correctly predicted for all test data (26 total), so the True Positive Rate is 100% and the False Negative Rate is 100%. The False Positive rate is 0%.
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** - What proportion of positive identifications was actually correct? (how correct we are; TP / (TP + FP). Our model has a precision of 100% at 0.5 threshold. When we predict an x-ray as being positive for pneumonia x-ray, it is correct 100% of the time. **Recall** - What proportion of actual positives was identified correctly? (how much left out; TP / (TP + FN). Our model has

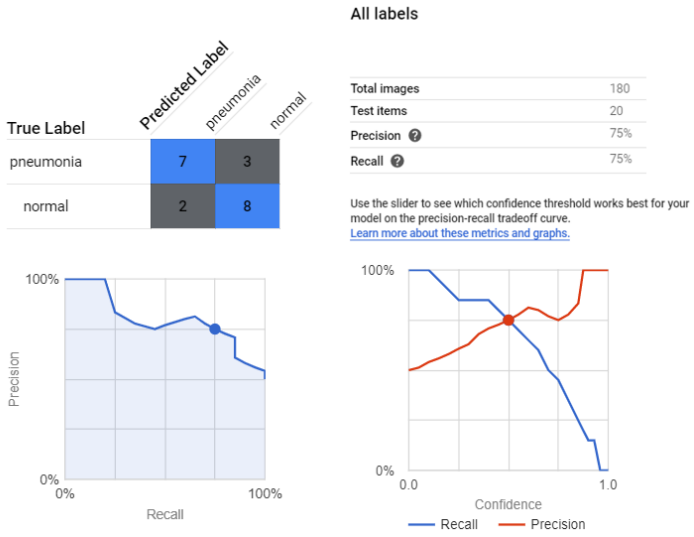
	a recall of 100% at 0.5 threshold. It correctly identifies 100% of all pneumonia cases. In this case we would want to optimize for recall, since a False Negative could be bad for a patient that does have Pneumonia.
Score Threshold When you increase the threshold what happens to precision? What happens to recall? Why?	If the threshold is high, we classify less images, but has a lower risk of misclassifying images. At threshold 0.9, recall is 96% and precision 100%. This occurs since we are more confident in our predictions (probability is higher for each input image). Conversely, If the threshold is low, the model classifies more images, but will misclassify more (decreasing precision and increasing recall, potentially). At threshold 0.1, recall is 100% and precision 96%.

2. Binary Classifier with Clean/Unbalanced Data

Train/Test Split How much data was used for training? How much data was used for testing?	The normal class had 100 images and the pneumonia class had 300 images . Of this, there was a 90/10 split for training/test data, respectively. As a result, 360 images were used for training, with 40 for testing (30 pneumonia, 10 normal).

Confusion Matrix How has the confusion matrix been affected by the unbalanced data? Include a screenshot of the new confusion matrix.	The confusion matrix, surprisingly, is not affected by the data much at 0.5 threshold. All test images were classified correctly. (This imbalance was not extreme at 25-75). However, I did think it would worsen.
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)?	At a threshold score of 0.5, it is 100% for both precision and recall. It is the same as model 1.
Unbalanced Classes From what you have observed, how do unbalanced classed affect a machine learning model?	With this model, it has had little impact, likely because it is not extreme imbalance, or there is large enough differences in the images to be discerned by the algorithm. If the imbalance were 99% to 1%, then this may be a higher impact.

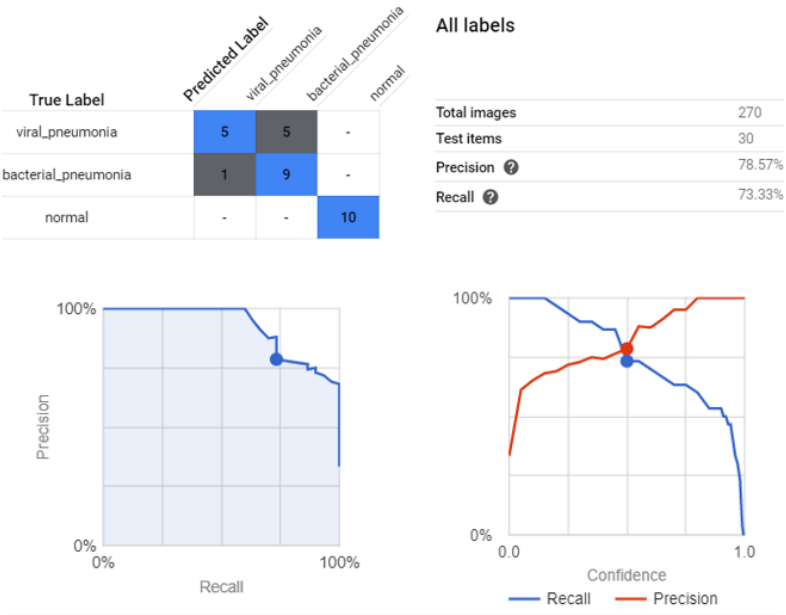
3. Binary Classifier with Dirty/Balanced Data



Confusion Matrix How has the confusion matrix been	With dirty data, we see that the confusion matrix now has values off of the diagonal, showing that there are misclassified images. 15/20 images are classified correctly,

affected by the dirty data? Include a screenshot of the new confusion matrix.	while 5 are not. This is worse than model 1 or model 2.
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?	With dirty data, we start to see the precision and recall falling, at ~75% for both at 0.5 threshold. The overall precision has dropped to 80.2 % as well, from 100% for models 1 and 2. The model misclassified one more pneumonia image as normal, than the converse. In this context, a false negative is bad for the patient. Model 1 and Model 2 still have the highest precision and recall. Model 3 is clearly worse.
Dirty Data From what you have observed, how does dirty data affect a machine learning model?	The more dirty data, the worse the model will perform. It should be limited, and shows the importance of annotating and wrangling data in the context of what outcomes we are aiming to achieve.

4. Three-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.	The model will most likely confuse the viral and bacterial cases, while correctly classifying the normal cases. After the model finished training, this was actually the case as the confusion matrix shows. The model could be improved by increasing the training data size. Inspecting images that are incorrectly classified. Perhaps a different algorithm (or ensemble) and including other fields may be an option.
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)?	At threshold 0.5, the models precision and recall are 78.57% and 73.33% respectively. This is calculated by calculating the precision for each class, and then averaging. The same is done with recall.
F1 Score What is this model's F1 score?	The models F1 score is 75.86% based on our precision and recall are 78.57% and 73.33% respectively. F1 = 2*Precision*Recall / (Precision + Recall)

Resources

- https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia
- https://cloud.google.com/vision/automl/docs/beginners-guide?&_ga=2.68477329.-1020558747.1584881242&_gac=1.250398386.1585799431.EAIaIQobChMImYXtm-vI6AIVDW6GCh1yPwTOEAAYASAAEgJZ6_D_BwE#evaluate
- https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall