|  |  |
| --- | --- |
| AutoML Modeling Report |  |

*Ognen Loparski*

Binary Classifier with Clean/Balanced Data

|  |  |
| --- | --- |
| **Train/Test Split**  How much data was used for training? How much data was used for testing? | * 80 images in each class used for training * 10 images of each class used for testing |
| **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 represents the percentage of correctly and incorrectly classified images by the model. Horizontally are the ground truths for the two labels, while vertically are the predicted labels.  These are the values for detecting pneumonia:   * TP – 100% * FP – 20% * FN – 0% * TN – 80%   The true positive rate for the “pneumonia” class is 100%. The false positive rate for the “normal” class is 20%. |
| **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 measures how many images out of all the ones with a certain predicted label, should have really been predicted in such a way.  Recall measures how many images were predicted to have the correct label out of all the images that really have that label.    For a score threshold of 0.5 the overall values were:   * Precision of 90% * Recall of 90% |
| **Score Threshold**  When you increase the threshold what happens to precision? What happens to recall? Why? | When increasing the score threshold between 73% and 77% both measures go slightly down. Above 78% precision goes higher than 90%, while recall continues dropping.    In general, increasing the confidence threshold increases precision and decreases recall. Increasing the threshold, decreases the number of false positives and since precision is calculated as TP / (TP + FP), logically it goes down. On the other hand, it also increases the false negatives and that affects recall which is calculated as TP / (TP + FN), making it go down. |

Binary Classifier with Clean/Unbalanced Data

|  |  |
| --- | --- |
| **Train/Test Split**  How much data was used for training? How much data was used for testing? | * 80% of the data is used for training   + 80 for normal   + 240 for pneumonia * 10% of the data is used for testing   + 10 for normal   + 30 for pneumonia |
| **Confusion Matrix**  How has the confusion matrix been affected by the unbalanced data? Include a screenshot of the new confusion matrix. | The confusion matrix did not change and shows the same percentages.    The only difference is in the item counts for the “pneumonia” class which jumped from 10 to 30. |
| **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)? | Both precision and recall have increased when looking at both labels together and stand at 95% for a score threshold of 0.5.    When looking at each class individually precision has increased for both classes, while recall has increased for the “pneumonia” class and decreased for the “normal” class.   |  |  |  | | --- | --- | --- | | 0.5 threshold | Precision | Recall | | pneumonia | 93.75% | 100% | | normal | 100% | 80% | |
| **Unbalanced Classes**  From what you have observed, how do unbalanced classed affect a machine learning model? | In general, having unbalanced data creates a tendency for the model to incorrectly classify more images in the class that has a higher number of samples.  In this case, as recall for the “normal” label has decreased, there is a higher chance for an image to be incorrectly classified in the “pneumonia” class. Still as the number of images this model is trained on is twice more than the first model, it performs better.    It is interesting to note that this model is a lot more sensitive to score threshold levels when compared with the previous model. |

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. | By introducing 30% dirty data in the dataset, the model’s performance has decreased in every aspect.    These are the values for detecting pneumonia:   * TP – 60%, down from 100% * FP – 30%, up from 20% * FN – 40%, up from 0% * TN – 70%, down from 80% |
| **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? | Both precision and recall for a confidence threshold of 0.5 have decreased to 65%.  Of all the binary classifiers the binary classifier with clean, but unbalanced data had both the highest precision and recall where both measures were 95%.   |  |  |  | | --- | --- | --- | | 0.5 threshold | Precision | Recall | | Clean / Balanced | 90% | 90% | | Clean / Unbalanced | 95% | 95% | | Dirty / Balanced | 65% | 65% | |
| **Dirty Data**  From what you have observed, how does dirty data affect a machine learning model? | Dirty data significantly affects the predictive power of a model, making it perform much worse than a model with clean data. |

3-Class Model

|  |  |
| --- | --- |
| **Train/Test Split**  How much data was used for training? How much data was used for testing? |  |
| **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? What might you do to try to remedy the model’s “confusion”? Include a screenshot of the new confusion matrix. | The confusion matrix shows us that the model performs perfectly for the “bacterial pneumonia” and “normal” classes while correctly labeling 80% of the “viral pneumonia” class. The “viral pneumonia” class is the most likely to confuse, while both the “normal” and “bacterial pneumonia” classes are most likely to be labeled correctly.  To remedy the model’s confusion, we could train the model with an improved dataset by providing an equal number of new images from all three classes. By adding 200 new images in each class to train the model on, the new confusion matrix is:    Unfortunately, the measures did not improve. This could be because there is too much similarity between bacterial and viral pneumonia or the quality of the images themselves is not great. Further analysis of the incorrectly labeled pneumonia images could indicate what features should be present in an improved dataset, but medical expertise is needed in order to perform that analysis. |
| **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)? | The model’s precision is 93.1%, while recall is 90%, both with a confidence threshold of 0.5.  Individual class precision and recall are calculated as:    The model level precision and recall are calculated as average of the individual class ones. On our case that is:    Google AutoML does the model calculation slightly differently, by using micro averages. |
| **F1 Score**  What is this model’s F1 score? | The model’s F1 score is 91.52.  It is calculated as: |