Project Proposal

Ekta Bharti

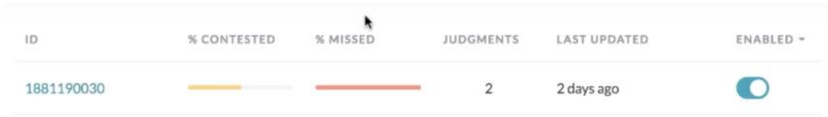
Data Labeling Approach

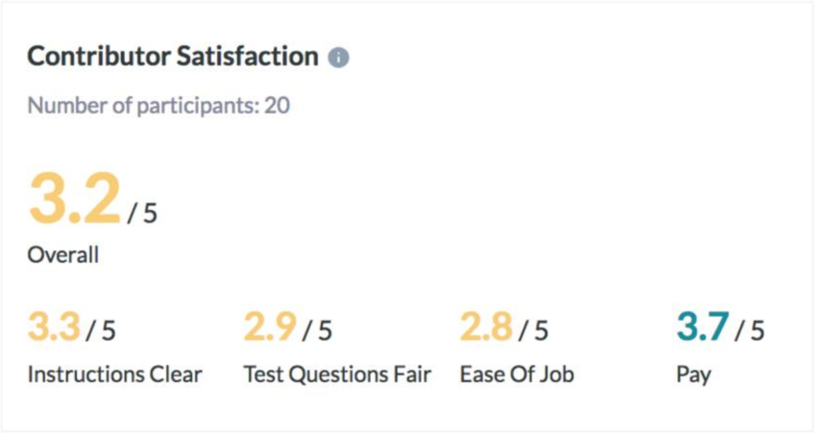
Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	I am trying to solve a typical healthcare industry problem where I am training the machine to identify a chest x-ray for pneumonia from a healthy chest x ray. I am using ML since it will help us in identifying even unknown cases of pneumonia based on cases labelled as pneumonia in the past trained data.
Choice of Data Labels What labels did you decide to add to your data? And why did you decide on these labels vs any other option?	I decided to give the following labels. 1. If the annotator thinks that the x-ray indicates pneumonia, then he marks it as Yes, otherwise No or Unsure whichever might be the case 2. If the annotator thinks the person has pneumonia, then he can choose from the options indicating which area is affected. (1-Left lung has large dense patches, 2-Right lung has large dense patches, 3- Left lung has scattered small white patches, 4-Right lung has scattered small white patches, 5-Both lungs seem affected). This question won't appear in case annotator chooses No or Unsure to the 1st question. 3. If annotator says Yes to pneumonia in the 1st question, then they can rate the severity of pneumonia by rating on a scale from 1 to 4 or else this question won't appear.

Test Questions & Quality Assurance

<table>
<tr>
<td><b>Number of Test Questions</b>

Considering the size of this dataset, how many test questions did you develop to prepare for launching a data annotation job?</td>
<td>Considering the data set of 117 images, I created 15 test questions (12.8% of the data set) . It has a good mix of all kinds of images that will help access the performance of annotator before they proceed to label rest of the images.</td>
</tr>
<tr>
<td><b>Improving a Test Question</b>

Given the following test question which almost 100% of annotators missed, statistics, what steps might you take to improve or redesign this question?</td>
<td>
<table>
<tr><th>ID</th><th>% CONTESTED</th><th>% MISSED</th><th>JUDGMENTS</th><th>LAST UPDATED</th><th>ENABLED ▾</th></tr>
<tr><td>1881190030</td><td></td><td></td><td>2</td><td>2 days ago</td><td>●</td></tr>
</table>
<ul>
<li>Simplify the question.</li>
<li>Make the journey of answering test questions intuitive.</li>
</ul></td>
</tr>
<tr>
<td><b>Contributor Satisfaction</b>

Say you've run a test launch and gotten back results from your annotators; the instructions and test questions are rated below 3.5, what areas of your Instruction document would you try to improve (Examples, Test Questions, etc.)</td>
<td>
<b>Contributor Satisfaction</b>
Number of participants: 20
3.2 / 5
Overall
3.3 / 5 Instructions Clear
2.9 / 5 Test Questions Fair
2.8 / 5 Ease Of Job
3.7 / 5 Pay
<ul>
<li>I would work on making questions simpler and more intuitive.</li>
<li>I would provide more clear examples in the instructions along with pictures to help annotators carry out labelling accordingly.</li>
</ul></td>
</tr>
</table>

Limitations & Improvements

Data Source Consider the size and source of your data; what biases are built into the data and how might the data be improved?	**To handle the given size of data** Add more images of different orientation, partial images, images of different shades of xray, different types of xray (digital, plain), images of different chest size in the training data set for labeling. **To handle the given source of data** Data labeling can be improved by using medical practitioners as 50% of annotators. This will remove personal biases of judging a pneumonitis condition by a non-medical practitioner. If multiple annotators contest the same question and a particular case is identified as pneumonia by a non-medical practitioner and identified as the opposite case by a medical practitioner, then override the label with the one given by medical practitioner. Similarly, a vice versa situation could be handled.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	Apart from the images present in the current dataset, I should also consider different past x ray color prints evolved over time. In the long run, depending on the conditions of air pollution , the definition of normal chest x ray might evolve as per medical practitioners and hence that factor could also be accommodated by training on past data.