Project Proposal

Ishika Jain

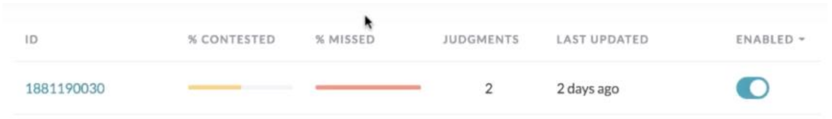
Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	This project helps doctors to quickly identify cases of pneumonia symptoms in children. This project is trying to solve the problem of the Medical Field. Machine Learning is used here since if by following some rules and tips we are reducing the time our doctors invest and we are using the machine to do the work so that those who are affected by pneumonia can be treated faster.
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?	There are three labels: “Yes”, “No” and “Unknown”. Yes is used when we find symptoms in the given image. No is used when the given image is symptom-free. The third label is used in cases when there is uncertainty if this image shows symptoms or not. These labels are simple to understand so I used them. There could be another option of choosing “Healthy”, “Pneumonia” or “Unknown”.

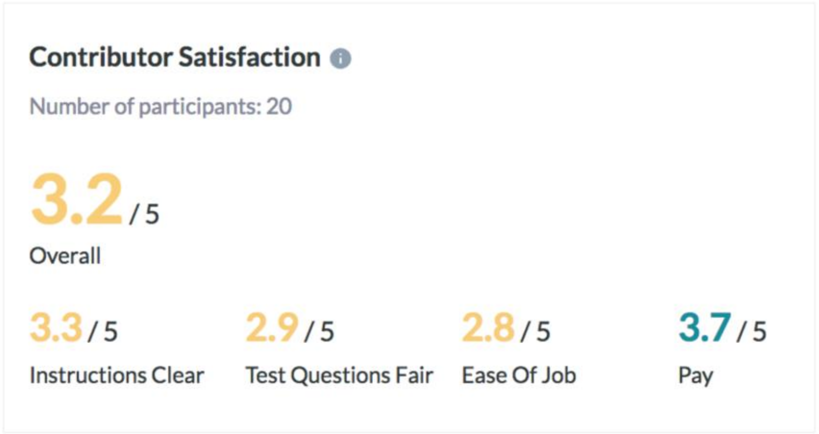
Test Questions & Quality Assurance

<table>
<tr>
<td>**Number of Test Questions**

Considering the size of this dataset, how many test questions did you develop to prepare for launching a data annotation job?</td>
<td>I developed 12 questions to prepare for launching a data annotation job. There were 4 for each label so that no bias can be generated towards any label.</td>
</tr>
<tr>
<td>**Improving a Test Question**

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>

The description provided should be clear so that annotators f ind appropriate reasons for the correct labeling. Rules and ti ps should be revised to remove ambiguity.</td>
</tr>
<tr>
<td>**Contributor Satisfaction**

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>
Contributor Satisfaction
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

More examples and tips should be included. The rules stated should be clear.</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?	Dataset Size is not large. The given data has 16 labels, 8 for yes and 8 for no. So there is no bias on the upper level. If biases are present in the data set then it should be improved by choosing the correct ML model. Pictures from different conditions should be included in the dataset to build a more powerful model.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	We may need to change the annotation job and update data to include more relevant definitions or examples.