Project Proposal

Prathamesh Morde

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	The goal of this product is to help doctors quickly identify cases of pneumonia in children. It will help to flag serious, healthy cases and act as a diagnostic aid. The system will actively learn and improve what its confident in to improve the business outcomes.
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?	For this I have selected four labels. **Yes with High confidence**–If you find cloudiness pattern around lungs, heart and diaphragm or one large area then we will use this label. This label uses two selections and provides the accurate result. **Yes with Low confidence**–Sometimes it's hard to differentiate between abnormal spots and vasculature therefore it helps an annotator to assess the uncertainty. **No with High confidence**–Here we use two filters to solidify our findings: a) Check for a clear diaphragm shadow b) No cloudiness (abnormal spots) around heart, lung. **No with Low confidence**- Here we use two patterns : a) There is not a clear diaphragm shadow b) No abnormal spots around heart, lung. This data label could help us to identify false positive scenario. **Unknown-** When it's hard to see a clear diaphragm shadow and hard to differentiate between vasculature and abnormal spots. This label requires more manual efforts to complete the assessment.

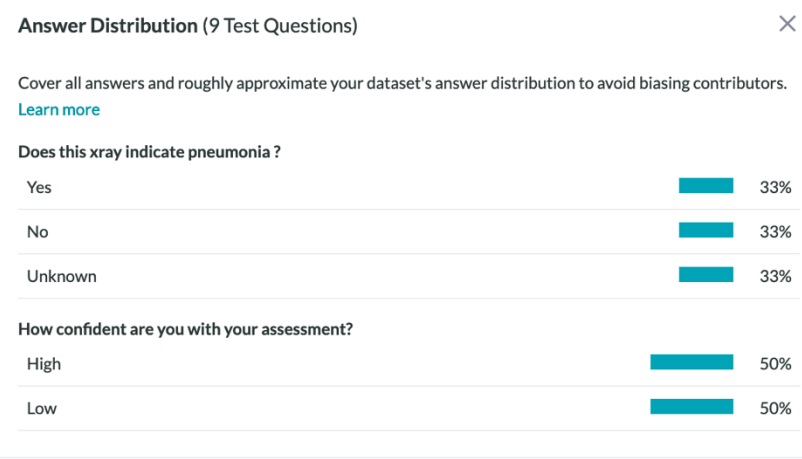
Test Questions & Quality Assurance

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?

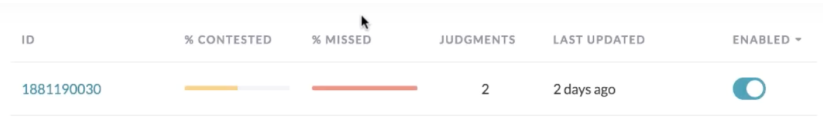
I have developed around 9 training questions for a data annotation job.

Here is the distribution of my quality.



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?



In case if annotators misunderstood my question, I would consider following steps
1. Augment the instructions.
2. Include more examples. (Take a particular image which annotators can't understand and bring it into the instructions as an example)
3. Update the data label design If it's necessary.

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.)	 1. Provide user friendly details for Steps, Rules and Tips. 2. Update the design label. 3. Create more test questions. 4. Add more Examples.

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?	1.The data source improved if we provide the good quality of the images so it would be better for annotators. 2. The majority of the images biased towards Pneumonia, so having more scenarios will help.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	1. In current data label design for No with Low confidence and Unknown labels might confuse annotators. To improve annotator's accuracy there would be only No with Low confidence label, which would tell us the annotator is unsure about the classification. 2. Other option would be instead of Low, High confidence we will ask our annotators how confident they are by supplying 1-5 scale where 1 is least confident and 5 is very confident. 3. Data annotation job also able to identify pneumonia in adults.