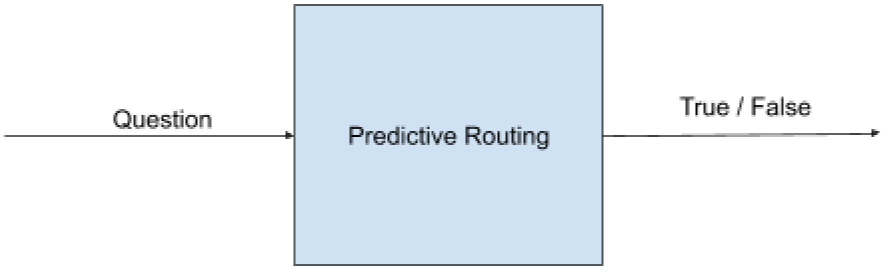
Machine Learning Engineer Take Home Exercise

As a customer support automation company, we deliver machine learning based triage classifiers called Predictive Routing. Your task is to create a Predictive Routing Algorithm given a subset of real questions from Directly's Expert network.

The model you are creating is a binary classifier where:

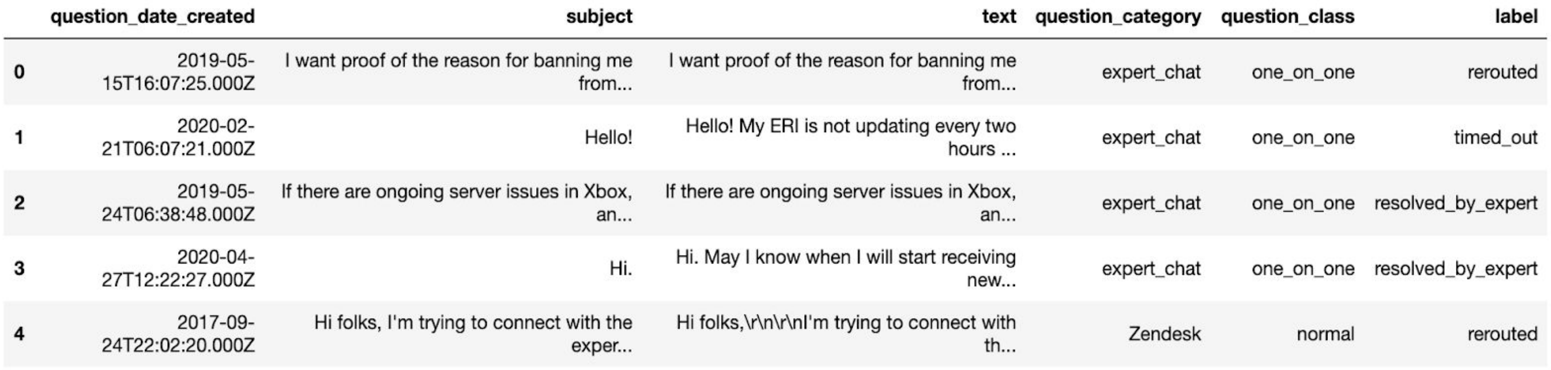
- True == This question will be resolved by one of our experts
- False == This question will not be resolved by one of our experts



The Data

The training data has 36,552 rows of questions submitted to our expert network, looking for resolution. Each row represents a single question with a feature set described in the next section.

First 5 rows of the training data:



	question_date_created	subject	text	question_category	question_class	label
0	2019-05-15T16:07:25.000Z	I want proof of the reason for banning me from...	I want proof of the reason for banning me from...	expert_chat	one_on_one	rerouted
1	2020-02-21T06:07:21.000Z	Hello!	Hello! My ERI is not updating every two hours ...	expert_chat	one_on_one	timed_out
2	2019-05-24T06:38:48.000Z	If there are ongoing server issues in Xbox, an...	If there are ongoing server issues in Xbox, an...	expert_chat	one_on_one	resolved_by_expert
3	2020-04-27T12:22:27.000Z	Hi.	Hi. May I know when I will start receiving new...	expert_chat	one_on_one	resolved_by_expert
4	2017-09-24T22:02:20.000Z	Hi folks, I'm trying to connect with the exper...	Hi folks,\r\n\r\nI'm trying to connect with th...	Zendesk	normal	rerouted

We have a testing set of 4,062 rows that we will use to evaluate your model. We will evaluate your model on the following 3 metrics:

1. Accuracy
2. Weighted Precision
3. F1

Features

1. question_date_created - date the question was opened to the system
2. subject - The subject of the question
3. question_category - The category of the question (may be null)
4. question_class - Either “one_on_one” or “normal”
    a. one_on_one - question is marked as answerable by only one expert
    b. normal - question is marked as answerable by anyone
5. subject - The subject of the question
6. text - The body of the question
    a. This is the only required feature to use in the model. All other features are optional to use in the model.

Label

The label column is the response of the machine learning problem.

1. timed_out - this question was never claimed by an expert, expired, and was rerouted back to the client
2. open - this question is still open and doesn’t have a final label yet
3. resolved - this question was resolved by an expert
4. rerouted_by_expert - this question was picked up by an expert, and was explicitly rerouted back to the client by the expert

You will notice that the label is not binary. In our test set, we made the following label mappings to make the problem binary:

1. timed_out -> False
2. open -> IGNORED entirely (neither True nor False)
3. resolved -> True
4. rerouted_by_expert -> False

Deliverables

We are looking for three outputs:

1. All of your training code that was used to ingest, parse, analyze the data and create your final model.
2. Testing code that loads your final model, and creates labels for a sample of test data.

```
# Sample testing code
import pandas as pd

testing = pd.read_csv('testing_sample.csv')

model = load_up_model(**kwargs, *args)
predictions = model.predict(testing)

# eg. predictions == [True, True, False, False, True]
```

3. A slide deck of **minimum 3 slides** that you will present to **a panel of engineers and non-engineers** explaining your process. You may assume that every person on the panel is familiar with the problem at hand and the data. Your main job is to walk us through how you approached the problem and through any assumptions you made. This presentation should be no longer than **10 minutes (we will leave 10-20 mins for Q&A and discussion afterwards)**.

Thank you for taking the time to work on this exercise. If you have any questions, please feel free to reach out to Sinan Ozdemir <sozdemir@directly.com>.