Context

One of the primary use cases for Eventbrite is as a self-service platform for creating a webpage to sell tickets to an event. While in general this opens up our platform to many more great event creators, the drawback is that nefarious parties will attempt to post unrelated and unwanted content.

Goal

Your goal in this assignment is to produce a new spam classifier based on the the provided features and data, provide a brief justification for your model, and answer some questions about how you would adapt your model to some additional constraints. See the **Results** section for more detail.

Data Description

The feature set is split between attributes of a user (event creator) and attributes of an event. These will be provided in two csv files. Each event is owned by exactly one user but users can own multiple events. Since users tend to produce either entirely spam events or entirely non-spam events, the target variable "is_spam" is an attribute of the user. Feel free to make your predictions at the event level if you prefer, but note that this choice will affect your evaluation process.

To come as close to reality as practical while remaining good stewards of our users' data, I've provided real data for sample of features that we use to identify spam on Eventbrite but not the code or data used to produce those features, or any identifying information about the events, etc. Even so, please consider this data to be sensitive and do not share it in any form.

Since this is a portion of the real feature set of a real model, the feature extraction methods used were tailored to that approach and may or may not fit well with your preferred modelling technique. Do your best to account for this early in the process. If you end up spending a lot of time on a model that simply will not work because of this, please include an explanation of why you believe that's causing you problems and how you would change in the feature set to address this given access to that layer.

Results

You will need to present your work in two stages. Results from the first portion will be the code that you used to train and evaluate your model. Results from the second portion will be an answer to a question about how you would adjust your approach given a real world complication.

1) Model Training

During this stage, it's important that you show your work. Include in your response all code, plots, and technical explanations required to reproduce your results and to understand your process. This should not be formal - a jupyter notebook with a bit of markup would suffice as would a few python scripts, a few plot images, and a brief text document. Please reserve some of your time for making sure to put your best foot forward on this front. I am far more interested in your process than your results, so clarity is imperative.

There are just two requirements on your results for this phase:

- You must use all data provided at some point in the feature extraction or model selection process. If you cannot find a way to use a field, please make a note of it, and provide an explanation of why you could not use it. If you run out of time, provide a sketch of how you would use that field.
- A comparison of two distinct models. The definition of distinct is up to you but the distinction should be larger than a slight tweak of hyperparameters. Results from each of these models and an explanation of why you consider one to be performing better or more promising than the other.

2) Labeling Errors

The spam operations team is receiving many reports that the model is failing to identify a lot of spam because it is scoring below the threshold on the existing model. You know this is a problem because anything not explicitly labeled as spam is assumed to be good in your training data. How would you adjust your approach to account for this? Write a paragraph outlining your approach.

When you're finished, please send your results to joshv@eventbrite.com.

Reminder: this assignment if expected to take around three hours and should be returned within 24 hours of receipt.