1. Given that the possible outcomes are that the job-opportunity email is either legitimate or illegitimate, we are looking for a model that can provide a yes or no answer. This situation would lend itself to using a classification model to solve this particular data science problem.

2. The requirements for the data need to answer our question would be as follows. We would need many examples of both legitimate job-opportunity emails from recruiters as well as examples of illegitimate job-opportunities from scammers. This would allow us to determine patterns in both sets of emails from which to create a classification model. We would also need to further define the type of work (i.e. regular employee vs. independent contractor) for which we would be seeking emails, as the content and other characteristics may be significantly different. Finally, it would be good to collect emails going back between 3 and 5 years to have a sufficient sampling for hiring cycles at different times of the year.

3. For data collection, we could source job-opportunity emails from individuals (or legitimate recruiting agencies) willing to volunteer access to these kinds of emails (with the necessary privacy protections applied). We could also contact email providers (i.e. Google, Yahoo) to provide us examples of known illegitimate job-opportunity emails from which to train our classification model (since it is highly unlikely that individuals themselves would keep these emails). We would specifically be collecting not just email content, but the sender email address, name, IP address and any attachments present.

4. For data understanding and preparation, we would need to do the following. First, we would need to formally define the criteria for an illegitimate job-opportunity email (i.e. a company that does not have a physical address or a country IP address that does not match a company IP address). We would then need to filter through the data collected to make sure we had a large enough dataset from which to build a model. This step might require the collection of more emails. We could then use word-cloud and other textual-based analysis to build a library of know “legitimate” email characteristics and “illegitimate” email characteristics.

5. For modeling and evaluation, we would proceed as follows. Using the library built in the previous step, we could create a text-scanning tool that would look for both legitimate and illegitimate email characteristics, determine whether the email was from a real job-opportunity or not by distinguishing between those sets of characteristics. We could use a training set of emails (not used to build the model) known to be from legitimate recruiters or illegitimate spammers to determine the success rate of our model. Subsequently, we could plot the true positives vs. the false positives on a ROC curve based upon the number of total email characteristics used to make the determination. The idea would be to find the optimum number of characteristics that we would need to determine whether or not a job-opportunity email was legitimate or not with maximum accuracy. This step might require several iterations in order to maximize the effectiveness as shown on the ROC curve.