Outline: Direct Certification Simulation

1. Welcome to our direct certification simulation
   1. Objective: share an approach to direct certification that empowers us to improve our match rates, i.e. to identify issues and develop solutions to resolve them;
   2. Purpose: our aim is to collaboratively explore the best way to think about the direct certification challenge, not necessarily to develop a solution for one particular state or school district; we hope to share some ideas, we know the specifics will have to depend on the circumstances in each state and school district
2. Our plan is to build matching models for two school districts with different levels of data quality
   1. Carefultown is a school district with reasonably good controls over the process of collecting student data;
   2. Oopstown is a similar school district in size and student body except it has less control over its collection process;

QUESTION: What are some examples you can share about good controls and not-so-good controls over the process of collecting student data?

1. Our approach to building the models today has three parts (that align with a standard framework known as CRISP-DM), we can think of them as steps though in practice its common to jump back and forth between them:
   1. Data Understanding (aka Data Exploration) : looking for patterns in the data that help us understand why records might not always match between benefits datasets and school datasets
   2. Modeling (aka Algorithms) : defining a logical, consistent way of determining what constitutes a match
   3. Evaluation (aka Match Rate) : determining how well the model works and coming up with ideas about how to fix it

QUESTION: How many of these three activities are part of your approach? How much time do you spend on these activities?

1. Let’s get started with Carefultown:
   1. First let’s try some Data Understanding
      1. Two datasets: SNAP and School
      2. Names from different ethnicities
      3. Names with different lengths (some have hyphens, some don’t)
      4. School Ages
      5. No spikes in birthmonth

QUESTION: we’ll revisit this analysis again after trying out some models, first, what else might you want to explore?

* 1. Next let’s build a model:
     1. To begin with we’ll focus on first name, last name, and date of birth, keeping the option open to add more data later
     2. We’ll also construct a scoring function that determines what constitutes a match between two records, our function will use points and sensitivity level, more points indicate more similarity between records, higher sensitivity means it takes more points in order to make a match
     3. The simplest version of this model is exact match: only identical records get points, and only records with points for first, last, and dob pass the sensitivity level
  2. Let’s run an exact match model on Carefultown and evaluate the results
     1. The printout from our model shows a match rate of <82%>
     2. The printout also has something called a confusion matrix that reports the match results in four categories:
        1. True positives -> these are records accurately matched to the SNAP eligible list
        2. False Negatives -> these are records from the SNAP eligible list that our model did not match but should have
        3. False Positives -> these are records from the school district enrollment that our model matched to SNAP eligible list but should not have
        4. True Negatives -> these are records from the school district enrollment that our model was correct not to match
     3. The graphs show how the results map to gender and group

QUESTION : Before we begin another round of data understanding on the false negatives, what is the most important element needed in order to create a confusion matrix?

* 1. Now let’s quickly explore the false negatives and see if we can get any ideas on how to improve the model
     1. Are the name distributions any different? Show graph
     2. Are the birthday distributions any different? Show graph
     3. What are the “distances” between the remaining names on the SNAP list and the enrolled students not yet matched?

QUESTION : What should we change in our model?

* 1. Within the scoring framework we’ve already developed (and the limited set of matching tools included in this scenario) let’s try using some algorithms to improve our match rate
     1. First, let’s try jaro winkler score and birthday magic with low sensitivity
        1. Match rates improve
        2. false positives explode
     2. Let’s try again with higher sensitivity
        1. Good match rate
        2. Low false positives
        3. Big improvement

1. Now that we’ve got the hang of it, let’s jump straight to evaluating an exact match model on Oopstown
   1. Match rate is about 50%
   2. False positives are higher relative to Carefultown

QUESTION: Now it’s your turn, what should we do next? Data Understanding? Modeling? Evaluation?

Go with highest voted option and repeat as much as we have time. Within a given area, ask open ended questions about what in particular they’d like to see, e.g. Modeling… ok, what do we want to try for the names, what do we want to try for sensitivity level, etc.