Applecart

Memorandum

From: Applecart
To: Prospective Data Analyst
Re: Data Analyst Practical Exercise

We are pleased to invite you to continue the interview process for our Data Analyst position! This interview round is a practical exercise – a simplified version of a project that you might expect to work on at Applecart.

You will have 24 hours to complete the project. You may use any and all resources at your disposal, but the response should be written by you alone. If you use outside sources, cite them. Feel free to contact us with questions.

Background

We have clients across the political, advocacy, and commercial spheres and we offer our clients a range of data products. These offerings stem from our core technology: the Social Graph. The Social Graph is a representation of a community of individuals and their relative "connectedness." For example, my graph illustrates how I am connected to my friends, colleagues, peers, and the people I know from various organizations of which I'm a member. The Social Graph algorithmically determines the relative strength, relevance, and origin of those connections. From a slightly more technical perspective, a graph is a collection of nodes (people or organizations) and edges (the connections between people).

The graph provides valuable information on individuals, including:

- How a given person knows another person
- How strongly two people are connected
- Of two people, who is more influential on a specific population
- How information and influence flow between people

In addition to the Social Graph data we generate ourselves, we frequently work with demographic data, modeled attitude scores, and consumer data. We also receive a lot of client-provided data, both individual-level and aggregate. We use these to generate contact and no-contact universes at both the individual and household level. For instance, if we're helping an organization target citizens with a phone bank, we'll want to rank order individuals to contact based on a score. By contrast, if we're sending mail to a household, we'll want to rank order households by score *after* removing any household containing a resident opposed to our clients' goals.

Instructions

We'd like to see you conduct some of the most common tasks required of an Applecart Analyst. Specifically, we want to see you use SQL to match datasets, analyze data, and build contact universes. We also want to see you communicate with our hypothetical client. To this end, we're providing you with a wire-frame slide deck and five data sources.

In this hypothetical client case, an education reform group needs to know whether or not the electorate in a particular district will support its policy goals. We're looking to provide them with a viability study of the territory in question. NOTE: there is not a single "right" answer to this question – we'd like to see you work through the data with the broader problem in mind.

You will receive a data dictionary and the following five datasets:

- Phone bank totals
- Voter file
- Teachers union member list
- Charter school parent list
- Modeled scores

When the following tasks are completed, you will return:

- Your SQL queries
- The completed deck
- A summary of the approach you took to complete the assignment

In order to complete the project, you will need to host the datasets that we send you. If you don't have access to a SQL database, here are two ways you can host the data on your hard drive:

1. Use Databricks Community Edition: https://databricks.com/product/faq/community-edition

2. Use pandas in Python: http://blog.yhat.com/posts/pandasql-intro.html

If you have a different method that you prefer, please feel free to use whichever tool is easiest for you.

Your task is as follows:

- Slide 1: Population overview

Summarize the overall electorate on the following variables: age, income, party, ideology, and charter school support. Present the client with both counts and percentages.

- Slide 2: Two target populations for a mailer

The client is interested in sending a piece of direct mail to charter school supporters, but the budget for the mailer is flexible. We have a list of charter school parents, who we want to contact for sure. Additionally, we have produced models to estimate the likelihood that each person in the dataset is a charter school supporter. Joining different tables as necessary, generate two potential target lists, one larger than the other. Describe your criteria for inclusion in each list. Be sure to de-duplicate each list to ensure that we're only sending a single piece of mail to each household. Provide final household-level counts for both target lists. Using your best judgment, explain why the client should choose one or the other list.

- Slide 3: No-Contact list

We have a list of members of teachers unions, who we want to avoid contacting, but the list contains only partial profiles. Match across the appropriate datasets to generate a final list of no-contact individuals and no-contact households. Then give a demographic overview of the no-contact list.

- Slide 4: Phone bank progress report

We have a dataset reflecting daily activity in our phone banks. The dataset indicates how many calls have been made, how many individuals have answered, and then counts for how many supporters and how many opponents of charter schools have been identified for each day. We need to visualize this information for the client. Come up with a simple, easily interpretable visualization and provide any necessary explanations.

We look forward to seeing your completed project. Contact Kathryn@applecart.co if you have any questions throughout the day. Please email your results back to Kathryn@applecart.co when complete. You have 24 hours… GO!