Data Scientist @ Vahan: Take-Home Task

About the Company & Role

At Vahan, we're helping 300M+ low-skilled workers in India find jobs using WhatsApp. We're a Y Combinator company that is backed by marquee investors such as Khosla Ventures and Founders Fund; several executives from Google and Flipkart are also investors in the company.

With 96% of smartphone users already using WhatsApp, there is no need for them to download anything else to use Vahan's job assistant. Now reaching over 100K job seekers per month across 1206 cities in India, the product generates millions of data points every month.

Data is at the heart of everything we do. We are looking for a business-savvy Data Scientist to join our team to help us draw deep data-driven insights and make more informed decisions across Product, Operations, Sales and Marketing.

Problem

A dataset generated via our tele-calling operations is available here. We have a small team of tele-callers who make calls to users who have applied for jobs on our job assistant (you can try it by sending a "hi" on WhatsApp to +91-8050-205-205). The goal is to gauge their interest in the job and then push them to go for the interview and get hired. In some cases, we might make calls to a list of leads received from a 3rd party. For instance, to a list of drivers received from a driving school to place them in a job at a cab company. Our tele-callers generally make follow-up calls to leads who don't answer the phone (RNR status) or who demonstrate interest (to get them to go for the interview/on-boarding).

Descriptions of the various tables in the dataset are below:

- leads
    - id (unique identifier)
    - userId (foreign key reference to our internal users table)
    - name (lead's name)
    - phoneNumber (lead's phone number)
    - city (lead's city)
    - state (lead's state)
    - source (source of the lead)
    - isExternal (false if the lead was acquired via the job assistant; true if the lead was acquired via a 3rd party source)

    - createdAt (date/time of creation of the table entry)
    - receivedAt (date/time when the lead was received by us)
- lead_calls
    - Id (unique identifier)
    - telecallerId (foreign key reference to telecallers table)
    - leadId (foreign key reference to leads table)
    - client (name of the client/company/employer who the call was made for)
    - status (status of the call)
    - comments (comments entered by the telecaller)
    - calledAt (date/time of the call)
    - createdAt (date/time of creation of the table entry)
- telecallers
    - Id (unique identifier)
    - name (telecaller's name)
    - phoneNumber (telecaller's phone number)
    - createdAt (date/time of creation of the table entry)

Here are the questions I'd like you to answer:

1. How efficient and effective are our tele-calling operations?
2. What are three things you would recommend we do to the efficiency and effectiveness of our tele-calling operations?
3. What are some other ways in which we can utilize this dataset to add value to Business, Operations or Product?

Please use appropriate visualization and/or modeling techniques to back your responses. Feel free to slice and dice the dataset along various features to arrive on insights.

Deliverable Expected

A zip file containing the following:

- All your code with copious comments.
- A report explaining the insights drawn as well as your approach.

The report should contain the following (the list below is by no means exhaustive):

- Detailed responses to the questions above.
- If you end up building models, rationale on the features and techniques you used.
- Any visualizations used to come up with your responses.

Tips

- Use Python or R for coding

- Rigorously follow good coding practices
- Add lots of comments to your code to make it self explanatory
- Last but not the least, back your findings with excruciating levels of detail

Have Questions?

Feel free to contact us via email on hr@vahan.co.

And finally… don't forget to have fun with the assignment. All the best!