Joe - Data & Engineering Rotation

During your continued time with the data team, you'll define and refine the project question, find two approaches, work with me to refine these, code it, and then present your findings. The details of these approaches are left intentionally vague, as a big part of data-related projects you may encounter will require communication with different stakeholders to discover what they're actually trying to find, and also to leverage their subject matter expertise to ensure that the definitions you are using are accurate. As you're working solo, I will be working with you more closely to aid on coding and logic.

This week, I want you to set the scope of this project. How do you expect the timeline for data exploration, cleaning, researching algorithms, refining, presenting, etc to look? I'd love to get a sense of the benchmarks + deadlines you think are reasonable. With all of that said, you've already done a lot of work with the data. How can you build on the features you'd chosen (or better examine the features you excluded)?

Churn

Customer churn is the percentage of customers that stop using a product. For most services, the definition is straightforward: in most industries, it means that the customer stops using. Our current definition is startups that *started* consuming OCI (>$0 total consumption), and then had 30 days or more of $0 usage. You and the other analysts began to answer this question, so your work would deepen the question.

The questions are as follows:

1. Are there unstudied or uncalculated characteristics that might impact the probability of a startup churning?
2. Do we need to adjust our definition of churn (e.g., no usage for 15 days or adding an additional condition)?

Consumption/Lifetime Value Calculation

The question:

1. Can we predict how much a startup will consume based on their characteristics?

Expected Output

- Deck outlining process and results
- Call out explicit traits indicating churn or VIP status (i.e., Financial startups in JAPAC have a 45% probability of churning; Startups that complete 50% of their Portal profile within 20 days are likely to spend more)

A couple notes for this project:

1. Which project(s) do you want to focus on? What's your expected timeline, and what deliverables do you expect to produce?
2. We'll be focusing more specifically on feature selection within data cleaning. It will be helpful to research how these might (or might not) aid in choosing features *given what you know about the data*.

a. Some potential candidates:
    i. Calculate metrics' correlation to each other
    ii. ANOVA
    iii. chi-squared tests
    iv. linear discriminant analysis
    v. Forward selection/backward selection/Recursive feature elimination
b. Might it be helpful to do the following? Why or why not?
    i. Combine and bucket the 3 Technology and 3 Industry columns to reduce features?
    ii. Look at different features that measure the time between x and y (e.g., the time between enrollment and Startup Portal profile creation or Market Connect step completion)?
    iii. Source of the enrollment?