VANDERBILT | **M.S. DATA SCIENCE**

Capstone Development (DS-5999)

Project Proposal

Please answer the following questions to help guide you in scoping out your Capstone project. Each answer should be about a paragraph for questions 3-7.

1. Who are you working with on this problem? This could be a company, a faculty member, research group, etc. If it is just you, that's fine, just say "NA" here.
   I will work with Professor Haslag, Professor White, and Professor Blocher. I will meet with Professor Haslag, Professor White for about once a month. And I will meet with Professor Blocher once a week.

2. Will you need to meet with me during the semester? If you said "NA" above, the answer to this question is "Yes." Not meeting with me requires that some other faculty member or practitioner is volunteering to provide mentoring and/or oversight. If you say "No" here, please let me know who this person is who will perform this task. It is perfectly fine for you to meet with both me and someone else if you prefer that, and you can change your mind later.
   I can provide the following:
   a. Accountability on timelines, planning and meeting deadlines.
   b. Help in how to solve difficult logical coding/analysis challenges.
   c. Assist in formulating and calibrating metrics that effectively communicate results.
   d. Scoping out your problem and ensuring it is not too big or too small.
   e. Anything having to do with analytics, statistics, and classification/regression ML tasks.
   f. I will be of less help in the following areas: Image recognition, NLP as well as specifics in deep learning.

   Yes! I will meet with professor Blocher.

3. Describe the problem you are solving. Be sure to include why this problem is unique or novel. If you are working in a research group or team, be sure to detail your precise contribution in relation to what the whole group is doing.
   I will scrape data from ***indeed*** website (www.indeeed.com) and combine these data with the "firm-level shocks" data. Then I will analysis the correlation of employee satisfaction with the firm or employee outcomes. This problem is novel because it will combine with the "firm-level shocks" data and give a Macro analysis of employee satisfaction. The companies can use the result to figure out how to improve the employee satisfaction and then get more benefits. Also, applicants can use the result to choose the companies.

4. Describe the data you need for your project. Be as detailed as you can be here – even including key column names you require is encouraged.
   NOTE: You should be sure that you have the necessary access to this data before the beginning of the semester. Access to data is a key point of delay in these projects!

   I will need two parts of data in this project. The first part of the data is the "firm-level shocks" data, which will be provided by Professor Haslag and Professor White. The second part of my data will be scraped from the Indeed website. I will collect these variables from the Indeed website: Employee satisfaction, the time of the position posted, reviews of the positions, etc.

5. Describe your approach or primary task. What are you going to do with the data above to answer the question above?

   I will first scrape the data from Indeed and then join it with the "firm-level shocks" dataset. Then I will do EDA to answer the question above.

6. Describe what you have done so far, along with an estimate of how much time you have invested in this project already.

   I already wrote some web scraping code and build the GitHub repository of this project. And I watched some teaching videos to learn how to use Selenium. I already invested about 4 hours in this project.

7. Describe what you think will be the biggest challenges you will face in executing this project. Identify 1-3 challenges.

   I think I will meet the challenges below:

   (1) There are a lot of companies on the Indeed website and one company will post a lot of positions. So, it will be very time consuming to directly collect all the job positions data on Indeed. This might be solved after I have the "firm-level shocks" data. And scrape the job positions based on the "firm-level shocks" dataset.
   (2) Some data on the Indeed website is not very "structured". Some companies are very small, which might don't have job reviews for them. This will induce some "None" value to the dataset.
   (3) Same job positions might be posted several times. I think this is also a problem that I should deal with.

You should simply download this file, save a copy for yourself, and insert your answers above. You may use images or plots if they are helpful to answer the questions.