Capstone Development (DS-5999)

Ty Painter - 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.

   Alliance-Bernstein

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.

   I have scheduled weekly update meetings with Andrew Chin and Michael Li from AB. I would also like to meet with you maybe bi-weekly to get another perspective of my work.

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.

   The main question I will be answering for my project is 'Can news data be predictive of stock/sector/market performance?' I will apply NLP to various news sources to determine stock alphas. The problem itself could be seen as novel, but the various approaches to solve to problem is what makes it unique.

The two main factors that can uniquely affect the outcome of the project are the news sources from which data is collected, and the NLP sentiment analysis package used to produce score ratings. The three main ways I plan to quantify news is by looking at the sentiment of company specific news along with microeconomic news, and also evaluating the volume of articles concerning the company or its specific industry. Finally, I would like to evaluate multiple models to predict various period returns based on sentiment ratings along with performance ratios and metrics.

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!

   The data source I will use to collect news data is The GDELT Project. GDELT is free and is said to be "the largest, most comprehensive, and highest resolution open database of human society ever created" as it is sponsored by Google, Yahoo, and other research initiatives. The data is somewhat messy so some data parsing will be necessary. Initially the most important features seem to be the keywords extracted from the article and the source of the article. Additional useful columns will be determined once more research is completed.

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

   The approach I will take will be to first collect periodic data (daily, weekly, etc.) and then apply various NLP techniques or 3rd party APIs to construct a sentiment analysis score. I will use the sentiment scores to rank stocks and assess how those scores relate to stock returns of a specified period of time. Next would be to see the volume of news covering each stock along with any other related factors (ratios, metrics, etc.) and how that could affect stock performance. Last, would be to evaluate various predictive models regarding stock performance and interpret their results.

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

   So far, I attended a kickoff meeting with AB representatives to establish the scope of the project right before winter break. I have worked periodically over break on gathering data and cleaning/parsing data values and establishing column names. I have spent about 8-10 hours working on this project so far. I have another meeting scheduled for this week with AB to answer any additional questions and gather some more details before the beginning the semester.

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

   The challenge I have initially faced will be to clean and organize data into a format that is interpretable. However, the biggest challenge I believe will be applying or developing sentiment analysis and how those result should be interpreted in terms of a stock's alpha. I think interpreting the results will be where I will rely heavily on Ab to provide some subject matter expertise, and also where I believe I will learn the most from and gain the most industry experience.

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.