Notice: All items on this sheet are subject to change. However, this document will always be the single source of truth.

Objectives

This course will require students to complete an end-to-end data science project on a realistic task, using real-world data sets. The project should involve rigorous analysis of a data using extensive code that the group constructs. The objectives of this project are:

- Allow students to utilize all of the skills they have learned throughout the semester.
- Simulate the types of projects students will encounter in the "real-world" upon graduation for the MSDS program.
- Provide students an opportunity to work on a large-scale data science project in a controlled environment, before interacting with real-clients in their capstone project.
- Help students identify where their future professional or academic interest may lie.
- Offer an opportunity for the class to understand what types of skills the students find necessary or interesting, so that future learning opportunities can be tailored to those gaps.
- Give students the opportunity to prepare materials that they can show to future employers or academic programs to highlight their skills and abilities.

Requirements

1. Students should pick a project in a domain that interests them and for which sufficiently a complex and large enough data sets exist.
2. Projects will be conducted in groups of 4 people.
3. Unless otherwise approved by the instructor, projects must be done in R or Python.
4. All projects must use Git to properly version and track all code. Github and Bitbucket both offer easy ways of hosting this. The instructor must be given access to the repository to review the code.

5. Students must actually deploy their model. This means that during the presentation, students must show that their application has a way to input new data and get some type of scored model result or business result. This could either be done via a front-end or via a command line.
6. All project written deliverables (ie NOT the Group Formation document, the code, or the project plan IF you are using a commercial project planning tool) should be submitted via Google Doc (one per group) to adam.m.mcelhinney@gmail.com with comment access enabled. Feedback will be given via comments. The docs should be named in the format of [group name]_[deliverable]. For example, if my group was called "Soccer Analytics", then I would submit my proposal as a file called soccer_analytics_proposal. Failure to submit in the proper format will result in the deduction of points.
7. Students are required to leverage at least one secondary data set that provides additional contextual data. This should be joined with their other data sets. Examples of this might be:
    a. Weather data
    b. Macro economic data
    c. Stock ticker symbol data
    d. Web-scrapped ancillary data
    e. New data sets
8. Students should show the difference in statistical performance between their model based on their primary data set and the model based on their primary plus additional contextual data sets.
9. Remote students will be expected to video conference in and present their respective materials.

Timelines

Timelines for the project are updated in the syllabus.

Deliverables

Group Formation

One student per group should submit their project group name and members using the form located here.

Proposal

The objective of this proposal is to ensure alignment between what the student group plans to research, the objectives of the course and the appropriate level of difficult for the group. The proposal should be a 1-2 page document that states:

1. A description of the problem to are researching.
2. The data sources that are available for this project and that you intend to use.
3. An outline of the steps you intend to take to complete the analysis. Utilize the process for data science projects outlined in the class and in the Myatt book.
4. The success metrics, statistical models, and KPIs you are utilizing to judge the effectiveness of the project.
5. A description of the final deliverables. Deliverables could include a website, graphs, report, etc.

An example proposal can be found here.

Project Plan

The objective of the project plan is to ensure that the necessary steps to complete the project outlined in the Project Proposal are understood in detail and assigned to specific stakeholders. Note that I highly recommend your team's track the project using a professional project management tool (typically available for free for small projects), such as Asana, or Basecamp. This will be valuable experience for how large-scale projects are actually managed in the "wild".

Project plans that are insufficiently thorough will LOSE points.

The project plan should include:

1. An expanded list of steps from the Project Proposal (note that this should easily be at least 50 items)
    a. Every data source you intend to incorporate
    b. Every KPI you plan to measure
    c. Every graph you plan to make
    d. Every model you intend to test
2. One dedicated owner from the group for each step.
3. Agreed upon internal deadlines to complete the steps listed above.

Example project plan can be found here. If you are planning on using Asana or Basecamp or similar, you may simply add my email (adam.m.mcelhinney@gmail.com) to your project, but please make sure you name your workspace and project equal to your team name.

Project Presentations

Project presentations should last approximately 20 minutes. All members of the group are required to speak during the presentation. The presentations will be driven based off the student's laptops. Make sure that you have redundancies in case of technical issues. Also, please email the slides to the instructor prior to the presentation. The presentation should roughly follow the format below:

1. Identify your team and each member on it.
2. 2-3 slides discussing the problem you are trying to solve and providing all necessary background information to frame the context and explain why the problem is important and why your group chose it.
3. 2-3 slides discussing the results of your initial data analysis and data cleaning.
4. 1-2 slides discussing the techniques you used for you analysis and why.
5. 1-2 slides explaining the results of your analysis. Ensure you frame the results around impact to the area of study.
6. 1-2 slides on lessons learned and possible next steps.
7. Groups are required to have deployed their analysis in a productionalizable fashion and do a simple command line demo during their presentation. Suggestions for tools to demonstrate the productionalized insights include:
    a. Shiny
    b. Flask
    c. Plumbr
    d. OpenScore

Project Reports and Code

1. Students are expected to submit final reports and all of the associated code.
2. Groups are required to store the code using a version control tool, such at Git (via Github or Bitbucket) and grant instructor access for reviewing proposes.
3. Final paper should be 6-10 pages in length. It should contain the same sections as the project presentation, though greatly expanded in detail.

Project Ideas

Students are encouraged to come up with their own project ideas that are tailored to their interests. However, if students are looking for some possible inspiration or project ideas:

1. Pretend you are working for a bank and considering offering loans to prospective customers. Analyze the Lending Club data sets to build models that predict whether or not someone is likely to receive a loan, whether or not he is likely to default on that loan and determine the ultimate profitability of that loan. Explore any changes in acceptance rates, default rates and profitability over time. Overlay this with macroeconomic data from the Federal reserve to see if that yields additional predictive power.
2. Explore what factors are associated with a given patent being approved. Explore what factors are associated with a patent being litigated. What interesting insights can you glean from the network of authors and patents? There are numerous other research questions that could be developed based on the electronic patent records. Merge this data set with Yahoo Finance data to determine what sector a given company belongs to and figure out how that affects the probability of them receiving an issued patent. Does this relationship change over time?
3. Predict the sender of a given email from the Enron scandal. Using the email text, recipient information, date, time and other features, can you determine who was likely to send a given email? What interesting patterns can you determine about the emails of those who were convicted of illegal activity versus those who were not? Can you use additional 3rd party spell-checking tools or data sets to clean up the emails?
4. Can you predict the number of stars an establishment is likely to receive in Yelp? What factors are associated with restaurants that have high ratings versus restaurants that have low ratings? How does geography affect the number and quality of reviews an establishment receives? How do Yelp ratings change seasonally? Will coffee shops get higher reviews if its an especially cold winder? Will smoothie shops be more prevalent in historically warm climates?
5. Interested in sports? There are a HUGE number of sports datasets available for your analysis. Can you predict the career longevity of baseball pitchers? Can you design a program to automate your fantasy football draft? Can you merge player data from college with pro data to better predict sports outcomes?