IEOR 4650: Business Analytics
Spring 2020 Professor Adam Elmachtoub

Business Analytics Group Project

1. Important Dates

1. **October 23** Form a team of 1,2,3,4,or 5 people. Each person in your team must be registered in the same class time. Notify your TA of your (partial) team by this date, you must cc all your teammmates. We will potentially assign more people to your team at our discretion, and do our best to balance them appropriately. If you do not send an email to your TA, then you are comfortable with us assigning you to a team. The TAs will complete and finalize the teams.

2. **November 9** Submit a one-page description of your project on Canvas. You should describe the problem(s) you are interested in tackling, why it is interesting, what data you will use, and a general plan going forward. The person in the group whose last name is first in alphabetic order will submit this under their name.

3. **November 11-12** Consult with Adam for 15 minutes on what you plan to do and make sure it makes sense. About 1 in 4 teams tend to walk away having to change their idea, most likely because it was not well thought out, no plan for evaluating success, or it was too simple/boring. Not all team member need to be present. Meetings will start and end on time. Sign-up spreadsheet will be available the week of.

4. **December 7** Submit your presentation slides, 4-6 page report (with appendices if needed), code (all languages acceptable), and other supplementary materials via Canvas. The person in the group whose last name is first in alphabetic order will submit this under their name.

5. **December 8 and 10** Deliver your presentation. Each team will have 10 minutes including questions, and will present on one of the two dates. Please email me if you have a time preference. You may have 1-5 people present depending on what your group wants to do. Questions may be asked during the presentation, so everyone should understand all the components of the project.

2. Project Details

The project is for each group to identify a business problem and answer it by analyzing data using techniques from class. Using one or more of the regression, classification, simulation, and optimization tools we learned in class, you should generate valuable insights from the

data. Projects will be evaluated mainly on whether you managed to generate value to a potential stakeholder (director at a company, government official, industry reporter, etc.). In other words, someone should be willing to pay to see your presentation! We encourage you to be creative in finding interesting and relevant questions, finding the right data to answer your question, and being creative in implementing the right analytics necessary.

In addition to the presentation itself, the main deliverables will be a set of presentation slides (at most 10 slides that should take no more than 10 minutes to present, the content is more important than formatting), a 4-6 page report with appendices as needed, and supporting files (datasets, code) to fully document your work. *You may use any programming languages you desire including R and Python.*

Students should work closely as a group in defining their project, collecting data, doing the analytics, developing the project presentation, and writing the report. Any programming languages and libraries are fine, you are not restricted to R. Please make it clear the work involved in combining, cleaning, and processing your data. In explaining your project, be concise and clear. Keep in mind that plots, tables, and other visual representations can be effective in conveying your ideas and findings.

You will be assessed along the following three criteria, with roughly equal weight.

1. **Value Opportunity and Creativity**–Where is the potential to use analytics to capture value? Specifically, what benefit can be achieved in terms of added value and/or cost reductions using analytics? What data is available that can be leveraged to achieve these benefits? Where and how was the data collected? What non-obvious insights or decisions were you able to generate? The ideal project setting would be one in which there is a large value opportunity and novel data that can be leveraged to capture this value.

2. **Analytical Rigor**– Does the proposed data, model, and methodology do a good job of capturing the value identified above? That is, are the data available appropriate and sufficient? What work was required to clean the data? Is the methodology used appropriate? Are the models and methods applied correctly? Are the models well validated? Is the model performance good enough to deliver the anticipated benefits?

3. **Clarity and Interpretability**– Can you explain the landscape you are working in and how your problem fits into this landscape? Is the problem clearly defined? Are your final outputs easily interpretable? Is it easy to understand what you did to the data and how it was done? Can someone reproduce your results based on your presentation and report? Can someone use your results to capture value?

We encourage you to find a topic that will be personally relevant to you. You can base your project idea on current or previous work experience. You may want to think of various topics you encountered as a student at Columbia (the class blog has many interesting references). Your project could be based on an idea you have for a new business venture based on analytics.

In all cases, we are not looking for a fully functioning system. Rather, think of your project as a proof-of-concept prototype. Specifically, we are looking for a good problem idea, and then sample data and analyses that are sufficient to validate the potential of your idea. View your project and presentation as something you might use to demonstrate your idea to a potential investor, government agency, nonprofit, customer, consulting client, or senior executive. They should be willing to pay for the report and presentation! Ask an interesting question(s), gather one or more interesting sources of data, do proper analytics, and communicate it all effectively!

3. Data Sets References

Groups are encouraged to collect data by themselves. Some references of notable sources of data are listed below. These are just examples, and groups are encouraged to search beyond this list, as there many sources. *You may not use data from a competition, be original!*

1. Fairness: `http://fairness-measures.org/Pages/Datasets`
2. Gapminder: `https://www.gapminder.org/data/`
3. Human Rights Data Analysis Group: `https://hrdag.org/data-publication/`
4. New York City open data: `https://nycopendata.socrata.com/`
5. U.S. open government data: `http://www.data.gov/`
6. World Bank: `http://data.worldbank.org/`
7. US Census Bureau: `http://www.census.gov/main/www/access.html`
8. Google trends: `http://www.google.com/trends`
9. Google finance: `https://www.google.com/finance`
10. Million song dataset: `http://aws.amazon.com/datasets/6468931156960467`
11. Data from academic papers, for example see data links in Esther Duflo's website: `http://economics.mit.edu/faculty/eduflo/papers`
12. USA Facts by Steve Ballmer: `http://usafacts.org`
13. Google Datasets: `https://datasetsearch.research.google.com/`
14. Open data on Amazon: `https://registry.opendata.aws/`
15. Enigma data exchange on Amazon: `https://aws.amazon.com/marketplace/search/results?filters=vendor_id&vendor_id=46c64acb-20c1-41fe-a495-a364f64d0083`
16. COVID-19: `https://delphi.cmu.edu/blog/2020/10/07/accessing-open-covid-19-data-via-t`
17. PropPublica: `https://www.propublica.org/datastore/datasets`