NYU Center for Data Science

DS – GA 3001

Tues. 6:45 PM – 9:25 PM

Term Project Instructions

Sept. 8, 2022

Assignment

To help you gain experience working with real-world data sets, this project will require that you mine actual bank transaction data to solve a problem of interest to your client, a large Italian bank seeking to reduce bad loans. You will be work in teams of 3-5 students. Please ensure that all students participate equally.

We are asking you to design the data mining task, mine the data, and describe your results. You will also need to research existing solutions to the problem. Your own results need not be on par with actual industry results; the goal is for you to get realistic hands-on experience, given the constraints of what you have learned and your resources and the course duration.

In doing/writing up/presenting your project, think of your task as a proof-of-concept study for a possible larger project. This is very common:

- We get an interesting (from a business perspective) idea, but don't know if it will really work
- We build a proof-of-concept (applying data science best practices).
- We present the results of the POC to our stakeholders
- Based on the results, we frame a recommendation
    - continue to build out the POC to learn more
    - take the next steps to implement the idea
    - modify the direction or the analysis and research; or
    - discontinue the effort

For purposes of the Term Project, you can think of your stakeholder(s) as internal members of the bank's lending group and risk management group (in other contexts, the stakeholder might be a consulting client or a VC/incubator who may fund you). You will need both to inform the stakeholder about your project and results and provide context by reporting on what has been done to date (elsewhere) on the problem.

It is most important to develop your models and analysis study well (within the scope of what we've discussed in class) and with a good understanding of the business problem.

You should use the frameworks and methodologies we used this semester to structure your project and writeup. Keep in mind that it may be ineffective simply to proceed linearly through the steps, and this may need to be reflected in your analysis. You can interact with me and your TA from the preparation of your initial ideas through your write-up, as a consulting group would interact with a stakeholder or funding source in preparing a research report. However, given the class size, and depending on the number of groups, there may not be as many opportunities to do so as you (or I) would like. Use your imagination, prior experience, or ask us to help to fill in any gaps in your understanding.

Deliverables

#1: By **Sept 20** you will submit your choices for teams for projects to your TAs. You should self-organize your teams, but ask us if you need help. Initial ideas can simply be a paragraph or two about what you are thinking you might do. There is also a section of Piazza set up for finding teams and/or members.

#2: By **Oct 11th** you will submit a **proposal** for your project with as much detail as possible your ideas. Include: the exact (business) problem; the use scenario, the related data mining problem (and whether it is supervised or unsupervised; the unit of analysis; potential target variables (if supervised); potential features; how the results solve the business problem? etc. Feedback will be provided by Roger during office hours which each team will schedule independently.

#3: By **Nov 1st** you will submit a status report ("Project Update") including preliminary & and any issues that you have run into.

#4: By **Nov 22nd**, you will submit your final write-up and code.

Your write-up:

1. should include the information detailed in the next section, in approximately the order given.
2. be about 15 – 25, fairly dense (but readable) slides, including any appendices you would like to include. Provide clear citations and bibliography for the external sources you use.
3. An appendix detailing the contributions of each team member. All group members should contribute to the analysis and write-up is also required

Your code will include two main modules:

1. The code you used to estimate your solution (with in-line comments and documentation where necessary); and
2. A "harness" function that will allow us to upload new data into your fitted solution by calling a single function. Your harness should produce a list of results that we can use for validation.

Your project write-up should include the information detailed below, in approximately the order given. Your write-up need not have corresponding sections or bullet points, but we should be able to find the information without searching too hard. Be as precise/specific as you can.

Business Understanding (take this seriously)

- Identify, define, and motivate the business problem.
- How (precisely) will a data mining solution address the business problem?
- What has been done in the past?

NB: I'd like to see a good definition/motivation of the business problem and a precise statement of how a data mining solution will address the problem. It's less important for your ultimate results match perfectly than that you get experience working through a realistic problem definition.)

Example of how to think about reviewing previous work for stakeholders*:* For this project, we are working on loan data from a large Italian bank. One way to think about how to present previous work, is to imagine that you are pitching your idea to a principle a VC firm that has experience funding a large number of machine-learning-based FinTech startups. The VC wants you to explain how your approach is different from what the firm already knows.

You would want to discuss:
- any good ideas that you incorporated from earlier work,
- how your default probability estimation approach is different from current approaches, and
- why your approach is better (uniformly or in specific settings only) for loan underwriting

Data Understanding

- Identify and describe the data that you used to address the business problem. Include those aspects of the data that we routinely talk about in class and/or in the homeworks, including variable definitions, calculated or derived variables, and, of course, potential biases.

Data Preparation

- Specify how these data are integrated to produce the format required for data mining.
- Describe any preprocessing you did on individual features or variables.

(NB: data preparation can be time consuming! Get started early.)

Modeling

- Specify the type of model(s) built and/or patterns mined.
- Discuss choices for data mining algorithm: what are alternatives, and what are the pros and cons?
- Describe the final set of variables used in your final model.
- Discuss the economic intuition for your model and results.
- Discuss why & how this model should "solve" the business problem (i.e., improve along some dimension of interest to the firm).

Evaluation

- Discuss how the result of the data mining is/should be evaluated and how you evaluated your model.
- Discuss any benchmarks you used to evaluate relative performance and provide performance analysis.
- Suggest how a business case should be developed to project expected improvement? ROI?

- If this is impossible/very difficult, explain why and identify any viable alternatives.

Deployment

- Discuss how the result of the data mining will be deployed
- Discuss issues the bank should be aware of regarding deployment.
  - Are there important ethical or legal considerations?
  - Describe settings in which your results should not be applied.
  - Identify the risks associated with your proposed plan and how you would mitigate them.

Holdout-sample evaluation

In addition to reviewing your writeup, I have reserved a portion of the data in a vault at an undisclosed location. We will use this data set to evaluate your true out-of-sample performance using the harness code that you provide.

Presentation:

Depending on the number of teams, the top k projects will be selected and the project teams will pitch their solution to the class. Note that in most cases, $k<n$, where *k and n* are the number of teams presenting and total number of teams, respectively.

Final Thoughts

- **You will get the most out of the project if you interact with us during the development of your ideas.**
- **And please feel free to come talk to us about your ideas as time permits.**