BAIT 509 Final Project Instructions

Contents

1 Overview

1.1 Goals

For your final project, you will work in groups of 4-5 that have been randomly assigned on Canvas. Your goals will be to:

1. **Develop a concept for a startup / product / tool with supervised learning as its core.** This doesn't necessarily have to be profit-driven, but the practical application of your idea (and therefore its success criteria) must be clear.

2. **Identify a suitable dataset, and use it to create a proof of concept for your idea**

Please read this document carefully! There are a lot of tips in here on how to make your project successful. The teaching team is here to help – Please communicate via Piazza if you have questions.

1.2 Deliverables

1. **Week 2, Monday**: A list of a few datasets or ideas your are playing with. Not graded, but the instructors will help you rule out non-starters and tell you if you're heading in the right direction.

2. **Week 3, Monday**: A **project proposal**, along with some exploratory analysis for your proposal, to show that you have verified that your proposal is viable.

3. **Exam week**: Your final deliverable will be a **report** for your stakeholders (think: investors, board of directors), in the form of a **written report and short presentation**. Your audience includes both generalists and technical experts. This means that you must make a case for the importance of the problem you are solving, the technical soundness of your proof of concept, and the practical significance of your findings.

2 Generating ideas for your project

Start looking for ideas early. If you have questions about the viability of your project, or how to approach it, please reach out to the instructor team via Piazza – we are here to help!

2.1 Datasets

What's a good dataset?

A prerequisite for a good project is to have an interesting dataset. You can start by looking at one of the repositories listed below for possible datasets, or bring your own dataset and problem.

- Your dataset should be sufficiently realistic to serve as the basis of a proof of concept for your business idea. Your project proposal and exploratory analysis due in week 3 will serve as a check-in point to help you determine whether this will be true.

- Your dataset will unavoidably be imperfect for this purpose: dirty data, sampling bias, etc. This is OK, but in your final report, **you must be sure to note how your dataset affects the interpretation / validity of your findings**, and what data you might collect instead for the next iteration of your project.

- Some dealbreakers for datasets:
  - Purely synthetic data
  - Datasets containing only anonymous features (e.g., simply called feature1, feature2, etc., where we don't know what the features represent).
  - Excessively simple learning datasets, with few samples
  - The following datasets are **blacklisted** – these have inherent problems or have yielded low-quality projects in previous years.
    - Diabetes risk prediction (This topic overall is blacklisted)
    - `https://www.kaggle.com/datasets/laotse/credit-risk-dataset` Synthetic data
    - `https://www.kaggle.com/datasets/ankitpatel2100/corporate-stress-dataset-insights-in` Synthetic data

- Don't just take the first ideas offered by ChatGPT – usually this leads to multiple teams selecting the same boring idea. Do something that interests you!
- Later in the course we will learn how to process image and text data, so you should feel free to use these data sources in your project. Doing so will require a bit of additional work, but the TAs and instructor are happy to support you in this. **Incorporating these kinds of rich data into your modeling will be worth extra points** – see the rubric.
- If you bring your own data and problem from the real world, this will **be worth extra points**. One way to do this is to scrape your own dataset from the web, but while rewarding, be warned that this can be time-consuming and you may run into rate-limiting issues – start early if you want to do this.
- You may not re-use a project from another class.

Where to find datasets?

There are many open repositories for machine learning datasets, for example:

- ML Datasets by Sebastian Raschka: If you're just exploring, start with this list
- Kaggle: a platform for hosting machine learning competitions. Active or recent competitions tend to be very good projects. Be careful though – outside of competitions, dataset quality is here is quite variable.

Some other types of real-world datasets for inspiration:

- Platforms like Yelp sometimes release data; you can also look for scraped data (e.g., for Airbnb listings, Amazon reviews)
- Many cities have open data portals, e.g., Vancouver open data portal, NYC open data portal
- There are many good transportation datasets, such as bike-sharing (Vancouver, Chicago, NYC), taxis (NYC), bike traffic (Vancouver)
- Sports analytics datasets tend to be quite rich

2.2 Example: Flood Forecasting

- Flooding can have a catastrophic impacts on people's lives and lead to substantial resources needing to be spent on rebuilding communities. We want to develop an early-warning system, which can predict floods early enough to warn residents. This will either be provided to local governments or directly to residents.
- Proof of concept: we will focus on the Bow River at Banff, Alberta, for which we have a dataset. The initial learning problem will be to predict water levels 2 days in advance, based on all data available up to that point.
- The data consists of daily observations of river discharge starting March 14 1984, to the end of 2014. Data are from the Water Survey of Canada (Environment and Climate Change Canada).

3 Deliverable 1: Project Proposal and Initial Analysis

At the beginning of week three, you'll be asked to submit a project proposal, and exploratory analysis. This is worth 15% of your overall grade (same as an assignment). **The goal here is to demonstrate that you have found an interesting problem with a viable dataset, and to receive feedback on how to proceed**. If you have doubts about this, please discuss with the teaching team before submitting!

In this proposal, work with your team to answer each of the questions below:

3.1 Proposal

1. What is your proposed business idea / product / tool? Who are the intended users, and what will they use it for? What information / features will they have available at prediction time?

2. What is the supervised learning problem that you want to solve for your business proposal? What are the features you want to use and the outcome you will predict?

3. What metrics will you use to assess the success of your proof of concept?

4. What dataset will you use? How does it support the goals of your project, and what are its limitations?

5. What methods and analysis do you plan to perform? Feature generation, models to try, interpretation, etc. This is primarily for you to receive feedback.

3.2 Initial analysis

1. Plot the distributions of features and outcomes in your dataset.

2. Make a first attempt at solving your problem without ML, to serve as a baseline (e.g., simply predict the majority class). How well does this perform on the metrics you have defined?

3. Make a first attempt at solving your problem using a linear model (linear or logistic regression). How well does this perform on the metrics you have defined?

3.3 Grading

You will be graded on the following criteria, worth 5% each.

1. Is the learning problem clearly defined? Would solving it with the chosen dataset achieve the practical goals of this project?

2. Are the metrics clearly defined? Do they reflect the practical goals of the project?

3. Is your initial analysis technically sound?

You may need to try out a few different project ideas to find one that is viable. If there are serious concerns about the project you propose here, you will be asked find a new problem and redo this. Communicate early and often with your TA if you have doubts to prevent this!

4 Deliverable 2: Final report and presentation

Your final deliverable (due exam week) will be a report for your stakeholders (think: investors, board of directors), in the form of a **written report and short presentation**. These deliverables together are worth 30% of your overall course grade.

- This should be considered a **report on a proof of concept**. You must have completed training and analyzing some model, but that means that it's okay if your current iteration doesn't achieve absolutely stellar performance (the problem may be fundamentally hard or your dataset may be imperfect) but you must **know what metrics you would like to achieve, clearly state the limitations of your current approach, and present a plan for getting there on the next iteration.**
- You must have iterated and improved on initial analysis you provided in your Project proposal. This can be in the form of more advanced models, more advanced features, interpreting your model's output, checking for robustness – whatever is necessary for your application.
- Your audience includes both generalists and technical experts. This means that you must make a case for the importance of the problem you are solving, the technical soundness of your proof of concept, and the practical significance of your findings.

4.1 Submission contents

Your submission should contain all files necessary to reproduce your analysis, as well as your final report. All submissions should include the following (or equivalent) documents:

1. Report (in .pdf format)
2. Presentation file
3. Clearly documented code (e.g., Python files or Jupyter notebooks (upload as both .html and .ipynb)).
4. If you have multiple files, include a `README.md` explaining how to navigate your repository. [1]
5. Data file(s), if the license permits distribution; otherwise please provide a link where it is obtainable

4.2 Final report

Your final report should contain the following:

1. **Problem definition**: Motivation, uses, and use cases for your proposed supervised learning tool, and the specific supervised learning problem you want to solve.
2. **Dataset**: exploratory analysis, provenance, limitations

[1] A short, clear README file is important for reproducibility. In it, you should aim to orient a hypothetical data scientist who stumbles upon your work but doesn't know anything about the project. After reading the README, they should know what the project is about, what files are what, and how to run your code to reproduce your results. More on README files [here](`https://www.makeareadme.com/`).

3. **Methodology**: Describe how and why you chose your pipeline: Models, losses, feature preprocessing, model selection, feature importance, interpretation.

4. **Results**: How well did your model perform on key metrics? How did they compare to simpler baselines?

5. **Practical implications**: What do your results imply about the viability of your project? Are there any key limitations of your analysis, ethical concerns, or any other practical considerations that your stakeholders should know about? How will your model be used after deployment and what potential caveats are there?

6. **References**: Where the data was obtained and any other resources you used to guide your analysis.

Ideally, you should try fitting several models using supervised learning to answer your statistical question. You can use any supervised learning method - not just the ones discussed in BAIT 509, but make sure you understand the model you are using, **do not use code that you do not understand**! It might not be relevant to pick one model, per se. For example, it's common to fit several models to see that they all point to the same conclusion. And if one (or some) don't, it would be useful to discuss why this might be, and whether we should take this to heart when we try to draw overall conclusions. Once you've selected your final model(s), you should include a discussion based on different scoring metrics as well as model characteristics (Is the model interpretable?).

4.3 Final Presentation

During exam week, you and your group will be asked to give a seven-minute presentation on your project, with three minutes for Q&A. This represents an end-of-project review meeting with both your hypothetical client, and with your BAIT 509 teaching team.

You should aim to briefly cover each section of the report, with a focus on the business and statistical questions, the methodology, and your recommendations for your client. As in the report, it should be clear why you chose the methods that you did. You should be forthcoming about the successes as well as the limitations of your approach.

4.4 Rubric

- **Presentation (20%)**
  - Presentation is clear and engaging, covering the major aspects of your project within the allotted time limit.
  - Slides are easy to follow. Usually this means that they should be heavy on visuals and light on text, and have an obvious flow.
  - All members of your group should participate roughly equally in presenting.
  - You should be able to respond clearly and concisely to questions about your data, your choice of methodology, and any limitations of your approach.
- **Report (80%)**
  - **Clarity**:
    - Takeaways are clear and prominent in the report.

    - ∗ Language is largely clear for a generalist audience, while technically precise when necessary.
    - ∗ Tables and plots are clear.
    - ∗ Overall, writing is concise and clear.
  - **Learning Methodology**:
    - ∗ The report discusses clearly and accurately the procedure and reasoning behind major methodological choices, including:
      - · Feature engineering
      - · Reflection on selected performance metrics
      - · Hyperparameter tuning
      - · Model selection
      - · Model interpretation
    - ∗ The methodology is well-suited to the problem and free of technical errors
  - **Analysis**:
    - ∗ The motivation for the project and choice of statistical question is suitable.
    - ∗ Discussion about the statistical question is insightful.
    - ∗ Discussion on the model choice is sensible and insightful. Machine learning concepts are well understood.
    - ∗ Other discussions pertaining to the business question are useful and insightful.
    - ∗ Note that the weight of various discussions within this category will vary, but generally, the discussion of machine learning models will carry the highest weight (depending on the complexity of the business question).
    - ∗ Real-world implications of your results are clear and logical.
    - ∗ Limitations of the methodology, dataset, metrics, learning problem, etc., are discussed and analyzed
    - ∗ For more details on what we're looking for, see the MDS reasoning rubric.

- **Extra credit (Up to 20%)** Exceptional work is possible for these projects. This credit will be provided at the discretion of the teaching team, but some ways to earn it:
  - Exceptional feature engineering: use rich text / image datasets, requiring the deep learning tools we cover later in the class.
  - Exceptional problem selection: Choose a real-world problem related to your own expertise / experience, with a real-world end user, where you will be involved in the deployment of your tool (i.e., not just a Kaggle competition).
  - Exceptional data sources: acquire your own real-world data, e.g., via scraping or by engaging with a partner.

5 Other Policies

5.1 Group grading

All group members will receive the same grade unless there is evidence that one member has contributed very little to the project. In case any group-related issues arise, please first try to resolve them via discussion within your group. If they can't be resolved, e.g. if one member of the group is barely contributing, please discuss with the instructor as early as possible. If a group member has caused significant conflict in the group, or barely contributed to the project, they run the risk of having their final project grade scaled down significantly. Please avoid this situation by being a good teammate, communicating early and often, and **making each team member's responsibilities clear**!

5.2 Plagiarism

- Make sure you are attributing not only written sources but coding sources as well.
- As always, plagiarism will be taken **extremely seriously** for this assignment.
- This includes code plagiarism. See the heading "Code Plagiarism" in the course syllabus (linked on the landing page for the course website).
- You must cite all original sources that you use in your assignment.
- If you're unsure about whether something constitutes plagiarism or not, please speak with the instructor.

5.3 Policy on use of generative AI

In the group project, you are permitted to use generative AI tools like ChatGPT. However, this is subject to the following exclusions:

1. You may not use generative AI to generate the text of your report or presentation. Any reasoning / verbal analysis must be your own.
2. You may use models and other tools suggested by AI that we have not covered in the course. However, when doing so, **you must describe why these tools are suitable for your problem, how they work at a high level, and their pros and cons.** This description must not be AI-generated.

While AI tools can assist you, they should not replace your own critical thinking or original work. Clearly separate your contributions from AI-generated content. In addition, AI is a tool, not a substitute for your own problem-solving abilities. Use it to enhance your work, but ensure your project is primarily the product of your efforts.