CAIS 380: Intro Machine Learning

Spring 2024

Final Project: Applying Supervised Machine Learning

Assignment milestones are DUE as indicated on the course schedule.

*This is a **group assignment**. Work with 2-3 classmates, and submit as a group on GitHub and Gradescope. **Individual assignments will not be accepted unless prior approval is obtained.***

Overview

For your final project you will investigate a research question of interest to you using supervised machine learning techniques. This project will be a large portion of your final grade, and is broken up into milestones, described below. Be sure to submit each milestone on time and to put your best effort into all pieces. Your final grade will be based on your grades for all milestones.

All milestones must be well formatted and readable. Text must be proofread, use good grammar, and communicate clearly. Code must be modular, well documented, show output, and run reliably.

Milestone 1: Proposal – 20 points

For your project proposal you will identify your project group, a plan for working together, your topic of interest, a research question, and potential roadblocks. Type up a document that answers the following questions. Your document should be 1 – 2 pages long and well formatted. **Submit as a group on Gradescope before class on the proposal due date**. Your proposal will be reviewed in class so that you get quick feedback.

1. Group
    a. With whom do you plan to work? Groups must be 2 – 3 members (you must speak with me if you would like to work with a different size group).
    b. Talk about your schedules. Detail a plan for coordinating your work throughout the project.

2. Topic
    a. What topic area will you work with? Why is investigating this area important/interesting?
    b. What is the broader impact of your analysis? (What will other people use it for?)
3. Building Blocks
    a. What topic area will you work with?
    b. Broadly speaking, what is the overarching research question for your analysis? Remember, your research question needs to be a question that you can investigate using methods we have learned in class.
    c. What type of modeling (classification, regression, etc.) will you perform to answer your research question?
    d. What variables will you need in your dataset to answer your research question? Which will be predictors, and which will be the response?
4. Roadblocks
    a. What roadblocks do you anticipate as you perform your analysis? (Name at least three.)
    b. Detail your plan for dealing with these roadblocks. How will you overcome them, or if you cannot, how will you modify your analysis?

Points will be awarded for answering each question above fully (each question is 2 points, 0 points will be awarded for missing or nonsensical answers, 1 point for partial answers, and 2 points for complete and correct answers).

Milestone 2: Check-in 1 – 20 points

Your first check-in will focus on finding a dataset to work with, understanding the data, and cleaning the data.

1. Data Source
    a. Find data that you can use to perform your analysis. You may need to combine multiple datasets, so having more than one is okay. Provide a link to your data source(s).
    b. If you plan to use more than one data source, how will you join your datasets? What is the common key across datasets on which you can join?
    c. Investigate the source of your data. Who collected it? Who provided the funds for data collection? Who published it? What possible data biases do the answers to these questions raise?

2. Exploratory Data Analysis
    a. What variables are included in your dataset? What is the type of each variable (type here means mathematical type, not computer science variable type)?
    b. Which variables will you use for your analysis?
    c. For each individual variable you plan to use, perform EDA. What is the distribution of the variable? Is it what you expected? Do there appear to be any missing values or distribution issues that could affect your analysis?
    d. Continue performing EDA to investigate any patterns between variables in your dataset. Do you notice anything that impacts your analysis plan?
    e. What data issues and biases did your EDA reveal? Ex. Do some variables have a lot of missing data? Is that missing data concentrated to specific observations?
3. Data Cleaning
    a. Which of the issues and biases that you found during EDA will you correct for your analysis? Why are these important to correct?
    b. Correct the issues / biases you identified above. Clearly record the steps you take so that your analysis is replicable by other researchers.

Points will be awarded for answering each question above fully (each question is 2 points, 0 points will be awarded for missing or nonsensical answers, 1 point for partial answers, and 2 points for complete and correct answers).

Milestone 3: Check-in 2 – 20 points

Your second check-in will focus on implementing your analysis. **The work you submit is not expected to be complete**; it is expected to show that you have made significant progress towards completing your analysis.

1. In-progress code
    a. Submit the code you have so far; it does not need to be perfect but should be readable and well documented.
    b. Roughly half to two-thirds of your analysis should be completed at this point.
2. Revised plans
    a. Have you run into any roadblocks that significantly changed your analysis plan? Briefly explain what happened and how you adapted.

In-progress code will be awarded up to 15 points as follows:

	Missing / Not Complete (0)	Approaching (1-9)	Meets (10-14)	Exceeds (15)
Code	No code is included in the assignment, or the code included is unreadable.	Code is missing one or a few key elements such as documentation, attributing sources, modularity, or appropriate variable and function names. Or code includes these elements but significant improvement could be made.	Code includes documentation, attributes sources, is modular, and has appropriate variable and function names, but minor improvements in one or more of these areas could be made.	Code is well done. It includes appropriate documentation, attributes sources, is modular, and has appropriate variable and function names.

Revised plans will be awarded up to 5 points as follows:

	Missing / Not Complete (0)	Approaching (2)	Meets (4)	Exceeds (5)
Plans	No explanations included, or most are incorrect or missing.	Text demonstrates some conceptual understanding of machine learning theory and techniques, but multiple details are incorrect.	Text demonstrates a conceptual understanding of machine learning theory and techniques, but one or a few details are incorrect.	Text demonstrates a deep conceptual understanding of machine learning theory and techniques. All explanations are accurate.

Milestone 4: Final Report and Presentation – 20 points

Your final milestone includes all the final deliverables for your project. Your submissions should we well formatted, proofread, and clear. **Submit on as a group Gradescope before class on the day presentations begin. No late submissions will be accepted. There will be no exceptions to this rule.** You will submit a final report, that includes code and text, and give a presentation to the class.

1. Your **final report** should include:
    a. Interspersed text and code chunks.
    b. Text chunks should:
        i. Be well formatted.

ii. Be proofread.
iii. Include an introduction to your project with your research question and a brief explanation of why your analysis is interesting and important.
iv. Include a data overview explaining your data source (who collected the data, who funded collection, who distributed the data), and any biases, ethical issues, or general issues with the data.
v. Include an overview of your data cleaning process. How did you clean the data and why did you choose to do it that way?
vi. Walk the reader through the analysis. What does each code chunk relate to? Why is it included? What do outputs indicate?
vii. Include a conclusion. What did you find with respect to your research question? What are the implications of this finding?
viii. Include citations for any sources referenced.

c. Code chunks should:
    i. Be well documented with comments.
    ii. Run without errors.
    iii. Show pertinent output.

2. Your **presentation** should include:
    a. An introduction to your analysis, including motivation for your research question.
    b. An overview of your data.
    c. An overview of your analysis. Do not show your code, rather, explain the steps of your analysis at a high-level and what you found at each step.
    d. A conclusion. What did your analysis show? What are the implications of this?
    e. Future work. What new questions do you have? How would you build on or follow up on this project?
    f. Appropriate visualizations.
    g. Contributions from all group members (each person must talk).
    h. Time for Q&A. [In total, your presentation should be ~15 minutes.]

Points will be awarded for addressing each bullet point above fully (each question is 1 point, 0 points will be awarded for missing or nonsensical work, 1 point for complete and correct work).

Reflection

Your reflection is an individual portion of the project that you will submit on Gradescope individually. The purpose is to reflect on your own work, and how your group worked together.

Write a few short paragraphs that address these points:

1. Your specific contributions to the project.
2. Your teammates' specific contributions to the project.
3. Whether you navigated any conflict or discrepancy in workloads with your teammates.
4. How you navigated those conflicts or redistributed work.

You will not receive points for your reflection; however answers may be used to adjust individual's project grades if the distribution of work was not even.