Lab 5: Final Project

This lab is due at 11:59 PM ET on Sunday, November 18, 2018. Late submissions will be penalized as described in the syllabus. See also “Other Penalties” at the end of these instructions.

You may work individually or in groups of two. If working in a group, only one student needs to submit the group’s work via Carmen. As with previous labs, you’ll do all your work in a Jupyter notebook, but you will submit a written report in Word or PDF format as well.

This lab is worth 200 points.

This lab will give you an opportunity for in-depth exploration on a topic of personal interest in the area of data mining. You will apply techniques and methods learned in the course to a real-world problem domain. This project will consist of a comprehensive, end-to-end analysis of data, driven by questions that you want to answer using data mining. The goal of your project is to give you hands-on experience applying data mining techniques to one or more real-world datasets, going through the following steps:

- Identifying a problem domain and dataset(s)
- Determining what questions you want to answer via data mining
- Choosing appropriate techniques and algorithms
- Implementing and testing your methods
- Evaluating your techniques on your datasets(s)
- Reporting conclusions

You have the flexibility to choose a topic of interest to you, but you should NOT recycle work from previous classes or assignments. You should choose your topic based on an application domain and data mining approach that interest you.

1. Proposal (10 points)

Please submit, via Carmen, a proposal by 11:59 PM on Sunday, November 4th. The proposal should not be more than a page, may be of any format that I might reasonably be expected to read, and include an outline of your proposed final project, including the following details:

- What is the topic of your analysis? Be specific about the problem domain and provide a bit of background.
- What data source(s) will you use? If using publicly available data, cite the source including the URL(s) where the data can be obtained. If using data of your own devising, describe the process by which you created or collected the data, and the date(s) on which the data were collected. Do not use synthetic datasets.
- What data mining approach (classification, clustering) will you use? You are encouraged to explore other algorithms and capabilities of the Scikit-Learn package.

The proposal counts as 5% of your grade for this assignment. Your grade will be based on submitting the proposal on time and on how well reasoned your proposal is.

I will provide feedback on your proposals, and you may be asked to make modifications. Your final submission should not deviate significantly from your proposal.

2. Business Understanding (30 points)

Write a paragraph or two describing in detail the problem domain that your analysis is drawn from. Include background, such as the current state or problem that motivates your analysis. Explain the real world questions that your analysis will attempt to answer.

3. Data Understanding (30 points)

Discuss any observed patterns or data quality issues, such outliers, missing values, and attribute correlations. Visualize those that are interesting or important. Include a dataset description and summary statistics, as always. Describe your observations.

4. Data preprocessing (30 points)

Determine whether any transformations or data preprocessing (data reformatting, feature creation, data integration, and so on) are needed and make the appropriate changes as needed. Describe and justify your decisions—i.e. what did you do and why?

5. Modeling (90 points)

Select and describe the data mining approaches and algorithm(s) you intend to use. Define your evaluation approach (e.g. cross-validation) and generate your test design (e.g. create training and test sets if appropriate). Build and assess your models and describe your results.

Scoring for this section is as follows:

• Select modeling approach/algorithm(s):	10 points
• Generate test design:	20 points
• Build model:	10 points
• Assess model using quantiative evaluation measures:	50 points

6. Conclusion (10 points)

Describe your conclusions, including the answers or explanations that your analysis provided to the original business question(s) of interest. Describe any problems that you encountered and how you dealt with them, and any gaps or special cases that your analysis did not address. Provide suggestions for future research that could be pursued to close these gaps.

7. Submit your work

Your final report should be typewritten and include appropriate visualizations, data tables, references, and meta information (title, names, page numbers, headings, and so on), in addition to all the written content that previously would have appeared in markdown blocks of your lab notebooks. Word or PDF documents are preferred. (Let me know if you have something else in mind.)

Save your files using descriptive filenames, as always. You may split your work into multiple notebooks if you wish but be sure to provide consistent and meaningful filenames. Submit your notebook(s) via Carmen, under Lab 5. Submit data files only if they are not publicly available.

Other Penalties

In addition to the usual penalties for late submissions, your work will also be subject to additional penalties, up to a maximum of 20% (40 points) for the following:

- Missing meta-information from your written report. Names, page numbers, title, and section headings at minimum are expected.
- Sloppy formatting – You needn't be a professional print designer, but content that is pasted together without much regard for the final presentation will not be well received.
- Explanations and descriptions that are not clearly articulated and well-reasoned.
- Analysis that is not thorough or is too trivial, or a project goal that is not sufficiently ambitious. Your project should be challenging, and it is okay if you do not achieve perfect results. Thorough and thoughtful planning, analysis, and evaluation of results will win the day.
- Gross omissions, misapplications, or logic errors. You don't have to be perfect, but your work should reflect at least a best effort to apply standard data mining practices (e.g. cross-validation) appropriately.
- Other errors or omissions, at the grader's discretion.