Heinz 95-845: Article Title on Applied Analytics: the Machine Learning Pipeline

Firstname Lastname ANDREWID/EMAIL@ADDRESS.EDU
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA, United States

Firstname Lastname ANDREWID/EMAIL@ADDRESS.EDU
Department of Research
Another University
City, State, Country

Abstract

The abstract is the summary of the article. Your potential readers will glance at the abstract to decide if the article is worth reading. Make it good–this is your most-read text!

1. Project overview – for reference – remove for submission

The project will be conducted from late March to May, with the **code and paper due on May 8th.**

1.1 Objectives

The primary objective of this project is to bring together the elements of the machine learning pipeline covered in class to a problem that interests you. By reinforcing these concepts, you will perform a rigorous analysis that uses common data science tasks and that may has impact in a particular applied domain. Towards this goal, additional objectives include:

- Statement of one or more specific use cases of machine learning for your application
- Presentation and evaluation of machine learning techniques appropriate for the task
- Description of the analytic pipeline with comparison to existing analyses
- Description of results
- Description of the implications of your analysis

1.2 Grading

The project is worth 40 points total: proposal 10 points, code 5 points, paper 25 points.

Code (5) - Your code should be commented and organized. You should be able to hand your code to a third party and they should be able to run it. The code should be clear enough to modify it. That may mean you need to provide basic details about how to run your code, *e.g.* in a README file. Your code should be submitted in the form of a git repository (private or public is your choice) and should not contain any private data.

Paper (25) - Your paper will be assessed for clarity, completeness, and conciseness of the Sections in this template. Your analysis will be assessed for largely for appropriate (1) study design, (2) implementation, and (3) reporting of results. Other Sections in the template will also be graded, e.g. is the task appropriately motivated, are appropriate related works identified, do the conclusions faithfully reflect the results of the analysis, etc.

(1) Study design - Likely you will not be able to control the actual study design, because the project will be based on a study already conducted or data set already collected. Nonetheless, documentation and review of the design steps is important to assess the study for correctness, impact, etc. Your particular analysis also has elements of a study design and those should be clearly described.

(2) Implementation - The steps of the implementation should be documented. Here are some questions you may want to consider. What was the form of the raw data? What data cleaning did you conduct? How did you conduct it? Was there missing data and if so how did you approach those features? Did you conduct cross-validation? Which models did you select for the study (e.g. random forests)? How did you select the parameters of your model? Did you hold out a final test set? What was your a priori hypothesis? How did you plan to test it? What are your outcomes of interest? Why are those appropriate outcomes to evaluate? This list is incomplete, but remember to include details pertinent to the scientific question and your approach to answering the question. Those details should be for sufficient to match your code to your write-up. If there are many such details, consider moving them to an Appendix.

(3) Results - Your report of results should include baseline characteristics, primary and secondary endpoints and visuals to illustrate the outcomes fo your studies. See Section 6 for details.

Please use this template for your write-up; you will find helpful information about the content of each of those sections. Feel free to change/re-order the sections below as you see fit. Please turn in both your TeX file and your PDF. **The main text of the paper should be 6 to 8 pages, single spaced, not counting References or Appendices; any additional material should go into an Appendix.**

What follows are section headers with guidance on points to include about your analysis.

2. Introduction

In your own words, tell your audicence about the problem. For example:

"Recent advances in {subtype of machine learning} (Name and Name, 2016) have resulted in substantial progress in {application domain}. However, {describe limitation} .... In particular, Name1 et al. (2016) describes a novel technique with the potential to improve {outcome}. In this work, we ..."

Keep in mind who your intended audiences are, e.g., computer scientists, the public, statisticians. In the later sections, you might use some terminology that some reader may not be familiar with. That's okay. But the introduction should be approachable from a wider audience.

Be sure to motivate and describe this paper's contribution to the literature. In order to do this, you must summarize existing approaches and/or fields (almost certainly multiple ones), and describe in what ways your approach should be an improvement.

Let me re-emphasize: the introduction discusses how your work builds upon and relates to the literature. While you had to be brief in your abstract, the introduction is the place to describe why this paper contributes substantively towards your objectives. Convince us.

At the end of the Introduction, provide the layout of your paper, e.g., "In Section 2, we provide background on {subject}, ...", et cetera.

3. Background

Background is a presentation of the underlying concepts necessary to understand the details of your approach. Often times there are subsections for key math concepts. For application papers this section often provides the context that your research addresses. Figures are helpful.

4. Method: Your Model Name Here

This section describes your model and references the notation you introduced in the Background Section. **Figures are definitely helpful here**, so that someone who is in your area can visualize how your approach is novel, and someone who is not in your area can visualize what you are doing.

If you introduce new mathematical or statistical methods, use the terminology you defined in Section 3 and define your model. Give the technical details and remember: do be precise and do be concise.

If you are combining existing methods, then you don't need to provide a ton of detail: feel free to just cite other packages and papers and tell us how you put them together.

If you developed new code that does not (should not) contain sensitive or private information, include a reference, e.g.:

" Code is available at `http://my.github.page.com` "

5. Experimental Setup

Note: if the paper is more about the application than the method, this Section may be entitled Methods and appear before Section 4

By reading the Experimental Setup, your reader should have the information necessary to replicate the study.

Describe the cohort/data. Provide information about the population, the inclusion and exclusion criteria, what data were extracted, how features were processed, etc. In fact, you may want the following headings. **A flow chart can be very helpful** to illustrate the experimental setup, study design, inclusion/exclusion process, etc.

For more applied papers, each of the sections above might be several paragraphs or pages because we really want to understand the setting.

5.1 Cohort Selection

Describe how the samples you used were selected to form your cohort and also to provide cohort descriptive statistics. In methodologic papers, the "Table 1" describing the population by covariate summary statistics goes here. In application papers, "Table 1" leads the Results Section. Relevant information about the study design, such how cases and controls were identified, goes here. See Section 6 for an example of how to build a table in LaTeX.

5.2 Data Extraction

Describe the pipeline from raw data to processed data. Figures can be helpful. What assumptions did you make? How did you deal with missing data? Do not place interpretations here except possibly for short justification phrases. Longer discussions about the assumption you made go in the Discussion Section.

5.3 Feature Choices

What features were used? What conversions were necessary? What assumptions (e.g. i.i.d.) are made? with how you might have converted the raw data into features that were used in your algorithm.

5.4 Comparison Methods

To evaluate your model, often times you will compare against existing models. If so, include them here with a brief description, citation, and any tweaks you made for your experiment.

5.5 Evaluation Criteria

Evaluation methods belong here as well. Perhaps you used accuracy and the AUROC– explain why these are most useful measures of the outcome.

6. Results

Present the results here. Do not describe how the results were obtained. Those descriptions belong in Section 5.

Typically there are multiple parts and subparts of your study. Use subsections to report the results.

6.1 Results on Application A

Give us some numbers about how well your method works, especially in comparison to some baselines. You should provide a summary of the results in the text, as well as in tables (such as table 6.1) and figures (such as figure 1).

You may use subfigures/wrapfigures (LaTeX packages) so that figures don't have to span the whole page or multiple figures are side by side.

6.2 Results on Subanalysis B

Did you conduct more than one analysis? If so, provide the details here.

Method	Outcome (%)
Us	20.1
Baseline	18.2

Table 1: Outcome by method used. These are our results.

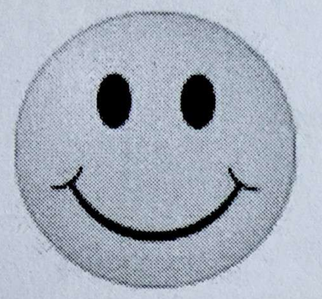

Figure 1: Example smile graphic.

7. Discussion and Related Work

This is where you characterize the outcomes of your method and draw conclusions from your experiment. The discussion will build upon the Introduction and the Results sections to synthesize where your contribution brings the field. Discuss any implications of your work. Discuss limitations of your work. Are there situations where you should and should not use your method. What implications are there on policy making, decision making, or future research activities? Remember to contextualize your work with respect to related work and provide references.

8. Conclusion

Summarize your work one more time, this time assuming the reader has read your paper. Build suspense for what your next extension to this method would be. How does the machine learning analysis fit in with the analytic pipeline?

References

Her Name and His Name. Aamlp project. In *The Journal Where It Was Published*, 2016.

Name1, Name2, and Name3. The title of said publication. In *Publishing Venue*, pages 123–345, 2016.

Appendix A.

Some more details about those methods, so we can actually replia them.