MATH 342W / 650.4 Spring 2021
Philosophy of Modeling Draft and Final Paper

Professor Adam Kapelner

Draft Due 11:59PM Sunday by email, Mar 21, 2021

(this document last updated 2:46pm on Wednesday 21st April, 2021)

Pick a natural phenomenon below to get started. If you have a natural phenomenon in mind, contact me so I can approve it.

(a) A well-known law of physics e.g. gravity

(b) Lung cancer

(c) Creditworthiness of an individual e.g. if a loan is paid back

(d) Motor vehicle driving ability

(e) Human lifespan

(f) A disease's danger to world population

(g) Global warming

(h) Taxation by a government

(i) Success in marriage

(j) Revenue in a business

(k) Quality of an individual's ability in a specific athletic sport

The common theme among the prompts are the existence of some natural phenomenon of import which is then captured in a numeric measurement and then modeled for the purposes of future prediction. Sometimes it is difficult to separate the phenomenon from the model since the model is so common and we are so brainwashed to think it is same as the phenomenon (it is not). Before you begin ...

Preassignment: due two weeks before draft

Write me a personal message on slack providing the following information (a) the phenomenon you are interested in writing about (b) real-world reasons as to why predicting this phenomenon are important (c) the *univariate*, i.e. single-valued, response metric that measures this phenomenon (d) precisely how you would measure the response (units of measurement, time frame of measurement, etc) and (d) a justification that you can actually measure this response n times relatively easily. The reason for this preassignment is the following: if the response is not set up correctly, then your entire essay will be a failure.

Paper Draft

You will receive comments on your draft after your submission. You will have the opportunity to revise your draft once and resubmit as a PDF. Resubmit by responding to my email with your completed PDF. You will receive a final grade for this assignment assessed by the performance on your revision (not your initial submission which will only yield a temporary grade).

In order to get an A on this paper you have to demonstrate both (I) you understand the concepts in this class and (II) you can apply these concepts to a hypothetical real-world modeling project. Throughout this document I will be color coding these two considerations. I am well-aware that this class is conceptually dense as we go over the learning-from-data modeling approach from start to finish and address nearly everything that comes up in the real world. From many years of assigning this paper, I've learned that (I) is a lot easier than (II) since it involves merely a paraphrasing of class notes. Resist the temptation to not bother with (II); there is where your effort should be directed.

The following must be written in your essay. Some of these items are short and require very thought (one or two sentences). Others require a paragraph and careful contemplation.

- A title that sums what the phenomenon is and how accurately you believe you can model it e.g. "The [phenomenon] can be Modeled Well" or "The [phenomenon] Remains Poorly Understood" or "The [such-and-such model] for [phenomenon] is a [good/bad] Model".
- Definition of a phenomenon.
- An introduction of no more than 1.5 pages that talks about the phenomenon, why you are modeling it why it is important and possibly a short history of attempts to model it.
- Definition of a model.
- Definition of a mathematical model.
- Description of what a model means in the context of your phenomenon.
- Description of your phenomenon

- Description of how the prediction target is measured exactly. Do you believe this measurement can be made accurately?
- Definition of causal drivers.
- Discussion of the phenomenon's causal drivers and how measuring them would be impossible.
- Definition of what stationarity
- Is your phenomenon and response metric stationary?
- Definition of supervised learning.
- A discussion of how can supervised learning be used in the context of your phenomenon.

- Definition of independent variables (features).
- Description of your model's features and how they are measured exactly.
- Do you believe your feature measurements are practical and can be made accurately?
- In order to improve the model's predictive performance, do you think some of the features should be transformed or interacted?
- Definition of historical data observations (training data).
- How would you go about obtaining (sampling) a training dataset in your context? What would n be? Would it be possible? Expensive?
- Definition of the three sources of error.
- In your modeling scenario, which of these sources do you anticipate would be large and why? Which are small and why?
- Definition of prediction using models.
- How can your model be used to predict?
- Who will be predicting using your model and for what purpose?
- Definition of prediction error metrics.
- Which error metric would you employ in your modeling context?
- What is the threshold for "usefulness" in your context?
- Definition of interpolation and extrapolation.
- When your model will be used in the real world, will its users be interpolating or extrapolating?

- What is an algorithm and candidate set? What is machine learning?
- What are some algorithm choices in this modeling context?
- Definition of the model section problem and how it arises during modeling.
- How would you select a model from those set of choices?
- Which model do you think will ultimately get selected and why?
- For the algorithm you are considering after selection, what is the null model you seek to outperform?
- Comment if you have enough sample in your historical data to fit this selected model.

- Define underfitting.
- Could your chosen model be underfit?
- Define overfitting.
- Could your chosen model be overfit?
- Define validation using concepts such as *in-sample* and *out of sample*.
- How would you validate your chosen model?
- Conclude: is the title of your essay correct and why? Tie the answer to what you believe will be your chosen model's predictive performance.
- Throughout the essay you must use all the following notation where appropriate:

$$t, f, g, g_0, h^*, \delta, \mathcal{E}, e, \mathbb{D}, \mathcal{H}, \mathcal{A}, t, z_1, \ldots, z_t, n, p, X, x_{\cdot 1}, \ldots, x_{\cdot p}, x_{1\cdot}, \ldots, x_{n\cdot}, \mathcal{X}, y, \mathcal{Y}$$

You are welcome to bring outside sources about philosophy of modeling as well as sources which help make your arguments in support of a prompt. Please cite them appropriately using natural text citations e.g. "The measurement device is accurate (Johnson et al., 1999)" or "Johnson et al. (1999) demonstrate the measurement device is accurate" and enter them into a bibilography. Format the bibliography in APA style.

Specs: Your essay must be typed and must be at least 10 pages double-spaced with one inch margin, 12pt Times (or Computer Modern if using LaTeX) and be appropriately organized. No need for a separate title page. Sectioning is at your preference and highly recommended. The bibliography does not count towards the page limit. Keep footnotes to a minimum and do not use endnotes. You must email me a PDF of your paper (Microsoft Word allows saving as PDF).

Revision is Due 10 days after draft is graded (as a PDF by email)

You will receive comments on your draft after your submission by email. You will have the opportunity to revise your draft once and resubmit as a PDF. Resubmit by responding to my email with your completed PDF. You will receive a grade for this assignment assessed by the performance on your revision (not your initial submission).