MATH 342W / 642 / RM 742 Spring 2025
Philosophy of Modeling Paper
(First draft and final submission)

Professor Adam Kapelner

Draft Due 11:59PM April 2 by email

(this document last updated 3:37pm on Wednesday 7th May, 2025)

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

(a) If an individual pays back a loan

(b) Human lifespan

(c) If an individual gets lung cancer

(d) Global warming

(e) Success of taxation by a government

(f) Revenue in a business

(g) Ability of an individual 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

Create a document called "modeling_essay_proposal.txt" (could be in any raw text format e.g. txt, md, tex/PDF) and push to your repository under the folder "writing_assignments".

In this document, provide the following information (a) the phenomenon you are interested in writing about (b) your 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), (e)

a justification that you can actually measure this response n times relatively easily and (f) a preliminary listing of features / proxies that you can use to model the response.

The reason for this preassignment is the following: if the response and predictors are not set up correctly, then your entire essay will be a failure so we want to catch this now.

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.
- An introduction of no more than 1.5 pages that talks about the phenomenon, why you are modeling your phenomenon (why it is important) and a short history of attempts to model your phenomenon.
- Definition of a phenomenon. A description of your phenomenon.
- Definition of a model. A sentence saying we seek to create a model for your phenomenon.
- Definition of a mathematical model.
- What metric do you use for your response, y? Are there any details needed to measure your y? Do you believe the measurement of y can be made accurately?
- What is the model target (this is usually y, but for probability estimation, it is the probability y = 1)?
- Who will be predicting using your model and for what purpose?
- Definition of causal drivers. What are your causal drivers? Discussion of how measuring them would be impossible
- Definition of supervised learning. How supervised learning is used to create your model
- Definition of independent variables (features).

- Definition of historical data observations (training data).
- Definition of prediction using a model.
- 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 error metrics. Which error metric would you employ in your modeling context?
- What is the threshold error for usefulness in your modeling context?
- Definition of interpolation and extrapolation. When your model will be used in the real world, will its users be interpolating or extrapolating?
- Definition of stationarity. Is your model stationary for y? Discuss.
- Definition of the model section problem and how it arises during modeling.
- Definition of underfitting. Could your chosen model be underfit?
- Definition of overfitting. Could your chosen model be overfit?
- Definition of validation using concepts such as in-sample and out of sample.
- Definition of algorithm
- Definition of candidate set
- Definition of machine learning
- Definition of null model - what is your null model for your response?
- Description of your model's features and how they are measured exactly. What is the value of p? Do you believe your feature measurements are practical and can be made accurately?
- How would you go about obtaining (sampling) a training dataset in your context? What would n be? Would it be possible? Expensive?
- What are some other algorithm choices in this modeling context for you response, your modeling target and your features?
- In order to improve the model's predictive performance, do you think some of the features should be transformed or interacted?
- Regardless of what you wrote for the previous question, write about how you can use forward stepwise with an expansive set of non-linear transformations of the raw features and interactions of the raw features. Write about how you would select the prediction model from the stepwise procedure.
- Do you have enough samples in your historical data to do forward stepwise regression? Discuss.

- How would you validate your chosen model?
- How can your model be used to predict?
- What is the threshold for usefulness in your context?
- 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 by email send:

1. PDF of the final draft *and*
2. a PDF with tracked changes.

To create the "PDF with tracked changes" for

MS Word: follow instructions at `https://www.youtube.com/watch?v=PH5wrLDH_wY`

LaTeX: follow instructions at `https://www.overleaf.com/learn/latex/Articles/How_to_use_latexdiff_on_Overleaf`

You will receive a grade for this assignment assessed by the performance on your revision (not your initial submission).