Week 5 - Reflection

Sam D. Olson

Reflection

The reflections detail: (1) The importance of understanding a model, particularly the importance understanding differences between predicting and explaining, in addition to connecting these observations to the forests project. Additionally the reflections contain: (2) A note on the differences between statistics and data science (data mining, machine learning, etc.), (3) A note on generative art and its connection to the CwD reading *"Making Statistical Graphs"*, and (4) An exploration of the Netflix Prize example given in Shmueli (2010). The reflections conclude with: (5) An appreciative note of your assigning these readings, particularly the Shmueli article.

(1)

A key takeaway I took from the reading was the difference between predictive and explanatory modelling, and the impact this difference has on the forests project. For the sake of brevity, key quotes are included below with summaries of their impact. A conclusion of these summaries is then given before proceeding to point (2).

"Using complex predictors may be unpleasant, but the soundest path is to go for predictive accuracy first, then try to understand why"

Takeaway: This approach may be used in the ultimate model used to estimate forest variables of interest. However, there has already been preselection of the explanatory variables, e.g. forest probability.

"Multicollinearity is not a problem unless either: (i) the individual regression coefficients are of interest, or (ii) attempts are made to isolate the contribution of one explanatory variable to Y, without the influence of the other explanatory variables. Multicollinearity will not affect the ability of the model to predict."

Takeaway: We are interested in multicollinearity, in particular given the redundancy of certain variables such as two separate forest probability measures (one from plot and one from spatial).

"In contrast to explanatory power, statistical significance plays a minor or no role in assessing predictive performance. In fact, it is sometimes the case that removing inputs with small coefficients, even if they are statistically significant, results in improved prediction accuracy"

Takeaway: We may consider including some variables due to their historical importance, rather than due to their significance or impact on prediction or explanation.

"I have polarized explaining and predicting in this article in an effort to highlight their fundamental differences. However, rather than considering them as extremes on some continuum, I consider them as two dimensions"

Takeaway: It is important to incorporate, or at the very least be aware, of the research's impact on the two dimensions of explaining and predicting.

"In terms of model evaluation and scientific reporting, researchers should report both the explanatory and predictive qualities of their models"

Takeaway: It is important to be able to communicate what model we ended up using, describing the components and framework it uses, in addition to what it can be used for, i.e. how good (and not) it is at estimating forest variables of interest.

"Typically the more complex we make the model, the lower the bias but the higher the variance."

Takeaway: As variance will be a primary metric to determine a preferred forest estimator, this validates our conclusion to avoid using numerous variables to explain forest variables of interest.

Conclusion: A consistent theme in the quotes selected and the commentary on the selected quotes is the need to be mindful of the purpose of the Forests project. Though we're looking to find a 'good' model for forest variables of interest, as a key metric is variance, we'll need to be careful of issues such as multicollinearity. Further exploration is needed before jumping into the modelling portion of the research, and I will be careful to read over the initial research prompt to take into account the above points.

(2)

The Shmueli (2010) article helped me better understand the difference between data science and statistics. Though inferential on my part, in the concluding section of Shmueli's article, they note:

"While statistical theory has focused on model estimation, inference, and fit, machine learning and data mining have concentrated on developing computationally efficient predictive algorithms and tackling the bias–variance trade-off in order to achieve high predictive accuracy."

Taking machine learning and data mining to be synonymous with data science, though a rough match-up, the above quote helped me observe:

1. Data science has historically focused on application, particularly in prediction.
2. By contrast, statistics has historically not focused on prediction, at least to the extent of data science.

The above points notwithstanding, the above quote reminded me of the analogies and discussion you gave during the first day of *Math 241: Data Science*, which similarly helped me understand the differences between statistics and data science.

(3)

Though CwD frames Chapter 4 as *"Making Statistical Graphs"*, it doesn't note an emerging field of data visualization—Generative art. As defined by Wikipedia, Generative art *"refers to art that in whole or in part has been created with the use of an autonomous system."* Understandably, Generative art may more appropriately be grouped into a sub field of mathematics and art, but the field undeniably touches aspects of statistical graphs as well.

I find generative art particularly interesting, as it incorporates guidelines noted in this chapter (use of color, what visualizations to consider, etc.). I posit the omission of any reference to generative art is related to its purpose, namely its purely visual (aesthetic) use. Understandably, given the focus of CwD on writing for data science, generative art is not within the wheelhouse of this course. However, the practice and application of generative art can be fruitful in understanding different aesthetic aspects of data visualization, though perhaps this point was more relevant to last week's readings.

(4)

I am interested in an omission of Shmueli's example of the Netflix Prize (for improving their recommendation algorithm), namely the winning algorithm of 2009 not being implemented, even though the 2009 winner improved the preexisting recommendation algorithm by roughly 10%. As Netflix stated: *"We evaluated*

some of the new methods offline but the additional accuracy gains that we measured did not seem to justify the engineering effort needed to bring them into a production environment", which reflected, as noted later in press releases and media coverage, Netflix's shift to focus on their streaming platform.

This is all to highlight an observation: Application, namely the decision to apply a particular research method, framework, or finding, is a complex issue, and one that should be considered by stakeholders of a project. Though one particular method may prove more effective than other methods considered, other factors such as logistics, *"engineering effort"*, theoretical frameworks, or an ease of communicating the method, may ultimately impact a method's use.

(5)

Thank you for assigning these readings. If you had not yet connected these dots, I wish to emphasize Shmueli recommendation in noting: *"As a discipline, we must acknowledge the difference between explanatory, predictive and descriptive modeling, and integrate it into statistics education of statisticians and nonstatisticians, as early as possible but most importantly in "research methods" courses."* Bully for you!