Reflections on “To Explain or to Predict?” by Galit Shmueli

STAT 5010, Prof. Brian Zaharatos,

University of Colorado Boulder

Student: Sushma Akoju

Read “To Explain or to Predict?” by Galit Shmueli and answer at least three of the following questions. Each answer should be typed, and roughly one paragraph (5-10 sentences) in length. The goal of this assignment is to provide the opportunity to think more rigorously about some of the statistical models that we will encounter in this course. What is their purpose? How can we get the most out of our statistical analysis?

1. What is explanatory modeling? Where is explanatory modeling most often used? What does explanatory modeling have to do with causality?
    a. Explanatory modeling is the application of statistical models to data for testing causal hypotheses about theoretical constructs. Psychology is one domain which is rooted in theory. Research in Psychology is founded in hypothesis and experiments before statistics was introduced as a metric to measure the psychology experiments by researchers. Thus psychology seems like a suitable example for Explanatory modeling. The well defined theoretical constructs, causal hypotheses have existed in Psychology since many centuries. Most branches of science evolved from philosophy. Similarly psychology is considered a branch of philosophy. It was not explicitly mentioned in the paper, but where there are well-defined theoretical constructs from domain expertise, it seems explanatory modeling plays a

vital role in testing causal hypotheses. Causality here suggests that the theoretical constructs cause the hypotheses and thus serves as an evidence. For example, there are types of personalities that can be studied in Psychology. Theory in psychology suggests internal locus of control and external locus of control that contribute significantly towards personality psychology. [Refer: https://en.wikipedia.org/wiki/Locus_of_control#Personality_orientation] Over several years, psychologists various types of tests to measure the locus of control in humans. It is still an unresolved mystery which is widely challenged among various researchers, even though there are several successful causal hypotheses that measure the locus of control based on various theoretical constructs. [Refer:https://en.wikipedia.org/wiki/Locus_of_control#Measuring_scales]

2. What is predictive modeling? Where is predictive modeling most often used? What does predictive modeling have to do with causality?
    a. Predictive modeling is the process of applying a statistical model or a data mining algorithm to data for the purpose of predicting new or future observations. Predictive modeling is used in applications requiring to predict an output value given their input values or in temporal forecasting such as time series forecasting for predicting output at a future time . The author recommends that predictive modeling be used beyond utility and instead be pursued as a scientific endeavor. Thus for new, large and complex datasets, that often contain complex relationships, but are hard to hypothesize, predictive modeling can uncover new causal mechanisms and helps to generate new hypotheses. But it does not necessarily result in a causal relationship.

       For example, when there is a large set of health data collected from devices, applications such as Apple Watch and Health applications, it can certainly provide new theoretical constructs such as the combination of heart

rate, cardio workout, nutrition, dietary habits, family history can point towards a likelihood of a heart disease or not. While heart rate, medical tests can suggest likelihood of heart disease from explanatory modeling, but predictive modeling could suggest new theoretical constructs such as dietary habits, nutrition that can change existing theoretical constructs surrounding heart rate, LDL cholesterol, blood pressure and DNA markers that may be expected to indicate heart disease. [Refer: https://www.cdc.gov/heartdisease/facts.htm and History of Coronary Heart Disease]

3. How are explanation and prediction different? Is it universally recognized that they are different?https://www.ncbi.nlm.nih.gov/pm
    a. The author discusses several definitions from various domains to distinguish between explanation and prediction since he attempts to relate the definitions more rooted in philosophy in science with an intersection of the main context of the paper i.e. data. Explanation, in the context of data, as defined in the paper, given a construct ***X***, and a theory that postulates X causes another construct ***Y*** which occurs via function ***F***, s.t. ***Y*** = ***F***(***X***). If X and Y are measurable variables, then there exists a function ***f*** s.t. **E(Y) = f(X)**. Explanation is more about the theory and basic science, such as describing about the world while prediction is about providing explanations of phenomena on measurable variables. It was at first not agreed that explanation and prediction were different, but it became apparent they are different as scientists, philosophers from different domains described each of them more. As discussed in the paper, explanation encounters different type of uncertainty which are different from that of prediction, in the context of modeling. Thus explanatory modeling is about matching ***f*** to ***F*** as closely as possible for statistical inference to apply to theoretical hypothesis, while predictive modeling is about X and Y as entities of interest and thus f is constructed from data.

Disparity Aspects	Explanatory modeling:	Predictive modeling:
Causation–Association	***f*** represents the underlying causal function. X is assumed to cause Y.	***f*** captures association between ***X*** and ***Y***
Theory-data	***f*** is constructed based on ***F***, so that it supports the relationship between X and Y and testing hypotheses.	***f*** is constructed from data, without expecting X to cause Y.
Retrospective–Prospective	Retrospective, ***f*** is used to test an existing hypothesis.	Forward looking. ***f*** is constructed to predict new observations.
Bias–Variance	Focus is on minimizing bias for better accuracy of theoretical hypothesis	Focus is on minimizing bias and variance, irrespective of

4. Describe how, according to Shmueli, the statistical modeling procedure might change based on whether the goal of the modeling is explanation or prediction.
   a. First, Shmueli defined generic process for statistical modeling as follows:
      i. Define Goal
      ii. Design Study and collect data
      iii. Prepare Data
      iv. EDA (Exploratory Data Analysis)
      v. Choose Variables
      vi. Choose Methods
      vii. Evaluate, Validate & Model Selection
      viii. Use Model & Report

   b. In the above steps, the goal can be explanatory or predictive, which decides the subsequent steps in the process.
      Given: given a construct ***X***, and a theory that postulates X causes another construct ***Y*** which occurs via function ***F***, s.t.

Y = ***F***(***X***). If X and Y are measurable variables, then there exists a function ***f*** s.t. **E(Y) = f(X)**.

Steps	Goal: Explanatory Modeling Estimate theory-based ***f*** as close as possible to ***F***	Goal: Predictive Modeling Given ***X*** and ***Y***, determine ***f***
.Design Study and collect data	Since f has to be estimated to match F, data has to be sufficient to match the precision as well as reduce the bias. Thus sampling methods also should be selected to better represent the data. **Example:** collecting health data would require we collect data that addresses bias w.r.t gender, age, race etc. Thus reducing bias is important to estimate appropriate ***f.*** **Preferred method:** Experimental data is preferred, i.e. by conducting experiments based on existing theoretical constructs, so data is reliable.	Need large samples for lower bias and variance as well as for cross validation/hold out data and test data (validation and testing new hypothesis ***f***). Thus sampling methods also should be selected to better represent the data. **Example:** collecting health data would require we collect data that addresses bias w.r.t gender, age, race, cardio workout, nutrition, dietary habits, family history. Reducing bias as well as variance are important here. **Preferred method:** Observational data is preferred by collecting data by observing and without much restrictions or well-defined settings and hence data is expected to be natural yet have nice measurement quality. (example: we do prefer a heart rate 60-100 beats per minute as acceptable, anything extreme like 10, 0 are clearly errors generated by a fitness watch or health application indicative of a watch being removed from wrist, but held in the palm with contact).
Prepare Data	**Missing values:** the values cannot be replaced by dummy variables. Experimental data containing missing values would need other domain specific mechanisms “to fill in the blanks”. **Data Partitioning:** for smaller datasets, which is usually the case in experimental data collections, cross validation is often used. (example: cement strength dataset, would have as	**Missing values:** can be a blessing in disguise, since it helps to understand the dependency or impact on **Y**. Thus, this does make for an interesting case. If it does not, can use some data imputation techniques. **Data Partitioning:** for smaller datasets, which is usually the case in Observational data, bootstrap is a common, successful method. Small data

	low as 30 rows of data collected from experiments in the lab.)	can have higher bias for estimating ***f***. Thus data partitioning is often a crucial step for predictive modeling.
EDA (Exploratory Data Analysis)	Involves summarizing, visualizing data and reducing dimensionality. The relationships between variables will have to match that of theoretical constructs depicting causal relationships. Visualizations, numerical summaries are compared, analyzed with that well-defined formal hypothesis, relationships. **For example,** for measuring cement strength for a regression problem, collinearity is known to exist, from theory. Requires a certain amount of water, sand to add to concrete to get sufficient strength. More or less water will reduce/increase the strength and other physical properties of cement. Dimensionality reduction is used to measure survey quality.	Involves summarizing, visualizing data and reducing dimensionality. Relationships between variables are not formal and/or unknown. Visualizations provide unfamiliar insights into data. **For example,** the president's approval ratings saw a significant drop since September 2021 since the US withdrawal of troops from Afghanistan and further drop in ratings occured from highest inflation in 4 decades. By looking at the raw tabular data from pollster websites, it would only summarize so much. We would know the decreasing trend from the initial visualization and numerical summaries of the pollster data. The comparative trends with that of other presidents from past data, shows the approval ratings trends during the first year of presidency. [fivethirtyeight polls] Dimensionality reduction such as Principal Component analysis or Non-Negative Factorization are used.
Choose Variables	For example, for cement data, data also needs air pressure, temperature values. By omitting temperature, if it was a constant for the entire data, there could be endogeneity i.e. features X could be reverse determined from Y i.e. X correlates with error in ***f***. Thus this is not acceptable, but helpful information in this type of modeling to "***explain***" the existing theoretical constructs and causality.	In predictive modeling, association is more important than causation. Thus advanced techniques in feature selection such as Permutation Feature importances, Correlation, Ordinary Least Squares Fit are considered to study the associations. For example, if there is a spotify playlist, with audio features such as tempo, energy, acousticness, instrumentalness etc with popularity score as

		response variable. From feature selection mechanisms, accousticness does not contribute to popularity score.
Choose Methods	Models need to be interpretive. For example, the same cement data is known to follow Bayesian ridge regression with some domain specific additional features included. Thus neural networks or k-nearest neighbors which are often not interpretive may not suit for demonstrating existing causal explanations. For example, consider an example of testing cement in external temperatures for climate change impact in manufactured materials. We can explain why having different temperatures would “cause” or “explain” change in strength of cement that is conducted in an open-industrial area, but we cannot predict cement strength for a future experiment since the temperature of the environment of the same open-industrial area settings is unknown.	Models do not need to be interpretive. Thus, vastly popular methods such as neural networks are often used by most data analysts. Thus the goal here being accurate predictions over unseen data matter more. For example, a neural network model may not shed light on an underlying causal mechanism ***F*** or ***f***, it can associate complicated associations and accuracy in predictions. For example, the vastly popular BERT neural network, is known to predict many challenging Natural Language processing tasks. Thus a small, but very fastly evolving field of explainability in neural networks is also becoming popular, to make neural networks more explainable. Also algorithmic modeling is more suitable for predictive models. Decision trees, neural networks are examples of algorithmic models.
Evaluate, Validate & Model Selection	Model validation i.e. if ***f*** adequately fits ***F*** and model fit i.e. if ***f*** fits the data $\{X, Y\}$. For example, for the resulting model, it is then used to generate random X values i.e. generate X values based on model generated coefficients for concrete, temperature, water, pressure etc and domain experts test measurements by mixing the ingredients by quantifying the coefficients for the ingredients in the lab and test strength of cement and check with hypothesis and results with experimental data collected and test this	Here validation is more focussed on generalization i.e. ability of ***f_hat*** to predict {X_new, Y_new}. Overfitting is a most common, dangerous impact of generalization. Performance of f_hat is measured over training as well as hold out data. If performance is too good on training data, it implies overfitting. As quoted from paper “...Multicollinearity is not a problem unless either (i) the individual regression coeffi-cients are of interest, or (ii) attempts are made to

	against statistical model validation. R^2 and F measures provide level of association for model evaluation.	isolate the contribution of one explanatory variable to Y, without the in- fluence of the other explanatory variables. Multicollinearity will not affect the ability of the model to predict.“ Predictive power is measured by Akaike Information Criterion (AIC) and Bayesian Information criterion (BIC). AIC estimates prediction error while BIC estimates goodness of fit.
Use Model & Report	The focus is on theory, causality, bias and retrospective analysis. The aim is to test or compare existing causal theories. Thus reports, scientific papers with explanatory modeling have statistical inferences and include theoretical constructs and unobservable parameters.	The emphasis is on data, association, bias-variance considerations and prospective aspects of study. Thus reports here include theory building aspects such as new hypothesis generation, practical relevance, discussion about error and predictability level.

5. What are some suggestions that Shmueli gives to the statistical community based on his analysis of prediction and explanation? Do you agree with these suggestions?
    a. Shmueli suggests following with two examples:
        i. To be aware of how statistical models are used in research outside of statistics, why they are used in such a way and in response to develop methods that support sound scientific research. By exchanging knowledge from such researchers at statistical conferences, workshops and requiring graduate students to read and present papers from other disciplines, this will help in gaining awareness on said issues.
        ii. To first acknowledge the difference between explanatory, predictive and descriptive modeling; to integrate into education on the aforementioned topics, include in research methods courses, and make this material available for non-statisticians; finally to advocate both predictive and explanatory modeling, clarify their differences, distinctive scientific and practical uses, and disseminate knowledge on both tools and knowledge on implementing both.

I agree with the suggestions. While reading the paper, I found that most of the information and discussion indeed was due to some gap between what each method meant to non-statisticians. In recent years, there is a sudden increase in the number of published scientific papers more focussed on modeling as an application or solution to a domain specific problem. It is indeed good to read this paper. Having such clarity on each type of modeling, how and when to use each type of modeling with a generic set of steps defined by a statistician seemed like much needed reading. I think this paper should be shared with the Data Science community in CU Boulder to be of better help to researchers, data analysts and data scientists. Perhaps a short talk or discussion about this paper from a professor would be even more helpful than sharing this paper to a community.