Explanation vs. Prediction & Data Models vs. Algorithmic Models

In the world of data science there are two major goals, explanation, and prediction. Many consider these two to be very similar but, in this article, Shmueli discusses how different these two goals are. A model whose goal is to explain must begin with a hypothesis in order to build or test a theory. The data in these models are used to describe how one variable causes another through a given function. In predictive modeling, the data is used to predict new observations that are not part of the given dataset. The distinctions between the two leads to better model designs and higher utility.

One of the major differences between explanation and prediction that is highlighted in this article is the retrospective-prospective disparity. This suggests that predictive modeling is prospective in that it looks to predict things that have yet to happen. In contrast, descriptive models are retrospective, they aim to explain what has already happened.

Considering the goal of the model, aspects of the design will differ such as the choice of methods. Models that aim to explain the data are more fitted to use models that are easily linked to the theory. Statistical models and regression-type are best for these whereas neural networks or k-nearest-neighbors would not fit as well. In predictive modeling, the output is generally unknown which makes methods such as algorithm based neural networks fit better. Depending on the goal of prediction or explanation, other aspects such as variables used in the model would also differ. A model that aims to predict will contain variables that would make an explanatory model too busy and inaccurate. If one aims to predict what rating a subject would give a movie, variables such as actors in the movie might not be useful. Whereas if the goal is to explain a subjects particular rating of a movie, knowledge of the actors would make for a stronger model. Knowing the differences between explanation and prediction will help exponentially advance data science. The facts provided in this article should be taken into consideration by anyone who aims to build mathematical models that provide the hidden answers behind data.

An alternative article that discusses this topic in a different light is Leo Breimans “Statistical Modeling: The Two Cultures”. He argues that the two cultures in statistics deals with extracting information using data models on one hand and algorithmic models on the other. In the data model he explains it as a box with one end having the x variable and the other end having the output, y, variable. Inside the box contains the regression functions for example. “The values of

the parameters are estimated from the data and the model then used for information and/or prediction" he explains. Model validation in this context requires things such as goodness of fit tests. In the algorithmic culture, the insides of the box are considered to be unknown, therefore decision trees, neural nets, and other algorithmic methods are places outside of the box to predict the unknown. Model validation for these methods is measured by predictive accuracy.

In the case of data models that use goodness of fit tests and R^2 for model validation, there are fallacies. Yes, the coefficient of correlation does show that the mathematical model is correct and you can reject the null hypothesis. But there are cases where you can have a significant correlation and the mathematics suggests a good model, but due to the variables chosen and other factors- the effect just does not make sense. Examples are discussed within the paper. In data models, Breiman explains that the conclusions are about the model's mechanisms and not about natural causes. This can lead to inaccurate conclusions and a model that is not emulating nature correctly.

The best way to see if the model is emulating nature is, instead of using goodness of fit tests that give a yes-no answer, deal with error. Put an X down and see what the output Y is. Then do it again but with Y'. The closeness of Y and Y' is a measure for how good the model emulates nature. What is observed is a set of X's that go in the "box" and an output of Y's is the result. Then an algorithm is found, such as F(X), that can be a good predictor of Y. These algorithms include methods such as decision trees and neural nets instead of the data models given by regression-based analyses.

Shmueli and Breiman's discussions can be compared in several ways. Shmueli is more concerned with explanation versus prediction and how prediction should be used more for theory building methods while Breiman is concerned that people rely on data models such as regression much more than they should and should rather move towards a more algorithmic modeling future. Both authors are proponents of prediction rather than the alternative. Breimans data models that he speaks of is equivalent to the explanation Shmueli speaks on. Instead of predicting, these models are trying to explain what factor caused which. In the end, it can be agreed upon through the evidence provided that these predictive methods have many more advantages that their counterpart as far as accuracy, being able to emulate nature, and reliability.