Visualizing Meta-Explanations in Early Intervention Systems for Police Departments

Damon Crockett
damoncrockett@gmail.com

Joe Walsh
jtwalsh@uchicago.edu

Klaus Ackermann
ackermann@uchicago.edu

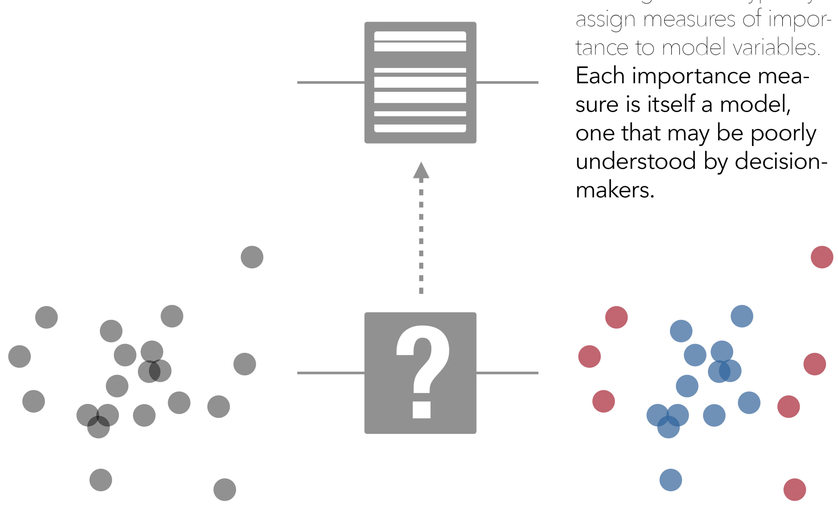
Andrea Navarrete
anavarreteriveras@gmail.com

Rayid Ghani
rayid@uchicago.edu

Center for Data Science and Public Policy, University of Chicago

ABSTRACT

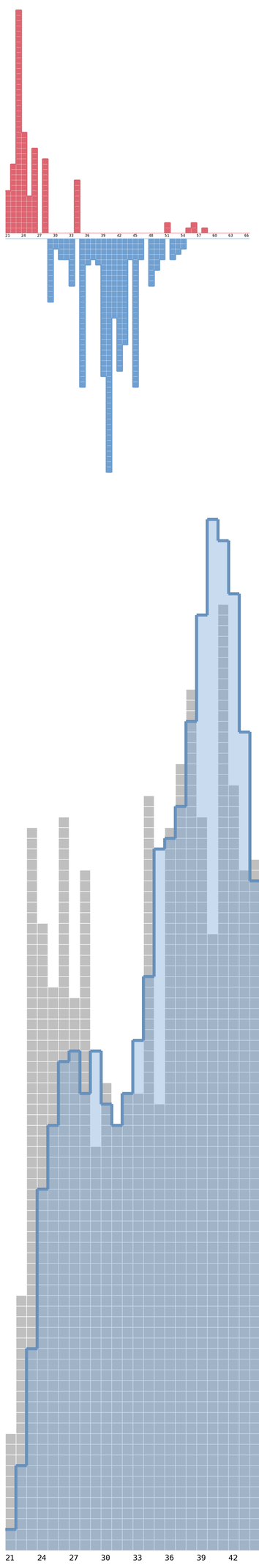
The recent spread of machine learning methods into critical decision-making, especially in public policy domains, has necessitated a focus on their intelligibility and transparency. The literature on intelligibility in machine learning offers a range of methods for identifying model variables important for making predictions, but measures of predictor importance may be poorly understood by policymakers, leaving the crucial matter unexplained: *why* the predictor in question is important. There is a critical need for tools that can interpret predictor importances in such a way as to help users understand, trust, and take action on model predictions. We describe a prototype system for achieving these goals and discuss a particular use case—early intervention systems for police departments, which model officers' risk of having adverse incidents with the public.



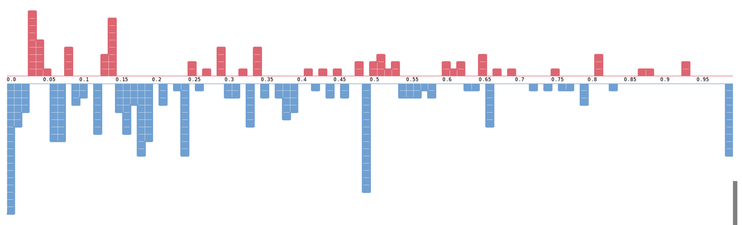
Explanations of machine learning models typically assign measures of importance to model variables. Each importance measure is itself a model, one that may be poorly understood by decision-makers.

Meta-explanations are explanations of importance measures that work by recovering and visualizing the data context that is typically ignored at this stage. The system is designed with both **analysts** and **policymakers** in mind, but rarely will the latter group use the system to recommend changes to the model or to visualize its training and test sets. Both groups are concerned with the behavior of **the model as a whole** and with its **individual predictions**, and the meta-explanations in each case are distinct, although the system will often generate views on the data relevant at both levels of scope.

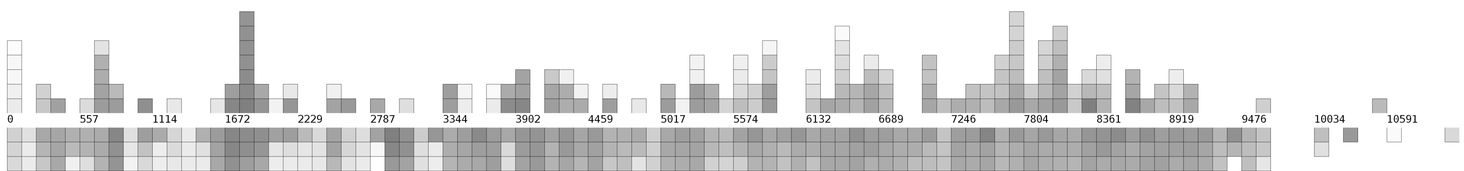
GLOBAL



The **lift-pull plot** is the concise form of meta-explanation for global predictor importances. **Lift measures the positive and pull the negative deviation in each bin from the global positive-negative ratio.** Zoomed-in lift-pull plots are unitized, making it easy visually to quantify deviation in each bin. On the left, an important predictor; on the right, an unimportant predictor.

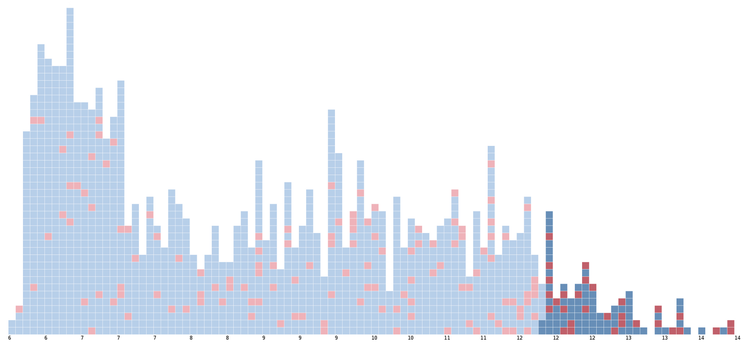


Like the lift-pull plot, **train view** is a form of meta-explanation for global predictor importances. However, while the lift-pull plot hides the predictor distribution in order to highlight lift and pull, train view recovers the distribution and presents it as a glyph histogram. This design both enables the user interactively to select individual observations in the data and visually quantifies important facts about the data in well-understood units (here, time slices of officers). On the left is train view for the predictor **officer age**, also plotted as a lift-pull immediately above. Here, we plot only the positive training observations, but the user can choose to plot the negative observations as well, below the axis. **The blue region covers what would be the negative distribution if it were the same total size as the positive distribution.** This allows us easily to identify which regions of the distribution see the greatest divergence between positive and negative groups. For this predictor, low values are proportionately more likely to be positive and high values proportionately likely to be negative.



Test and active views are like train view, but show distributions of modeled datasets. **Higher risk observations are darker and are plotted nearer the axis.** Test view plots positives above and negatives below the axis. Though shown plain, **this plot can encode additional data—e.g. 'is predictor important for this individual?'** Zeroes are common in our data; see below left.

LOCAL



Response view is a glyph histogram of the response variable on either a test set or an active set. Observations in a set are represented by glyphs whose color attributes allow us to identify true and false positives and negatives. The saturated glyphs to the right are treated by the model as positives, the de-saturated glyphs to the left as negatives. Reds are positives; blues are negatives. Response view serves both as a summary view of the behavior and performance of the model and as **an arena for the interactive selection of individual observations**.

Individual views support meta-explanations of individual model predictions. **It is common among algorithms that explain individual predictions to output a list of predictors important for the prediction**, perhaps additionally with ranks, scores, or coefficients. **Our system takes these outputs as inputs and produces views like the one above.** Four predictors important for the assignment of risk to one particular individual are presented as simple histograms, with **a yellow dot indicating the predictor value for the individual in question.** Users can select predictors for a closer look. Presently, small multiples views drill down to either test or active view on the selected predictor. In active and test views, additional data can be mapped to individual glyphs in order to fill out a substantive meta-explanation of the chosen individual prediction. At present, we are enlarging the space of individual views to include plots that locate individuals in multidimensional similarity spaces. High-dimensional similarity is difficult problem in data science, and a current challenge for us is finding appropriate data dimensions and clustering methods to produce neighbor relations that make sense for our data. **The ultimate goal is to give the user a sense of what sorts of observations are clustered nearby and how they are scored. In this, we are careful to avoid misleading the user into thinking that specific interventions have known effects on risk.**