Using Visual Analytics to Interpret Predictive Machine Learning Models

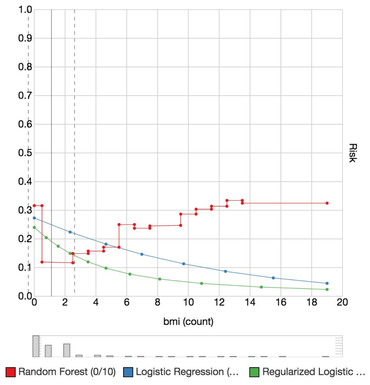
Josua Krause*, Adam Perer+, Enrico Bertini*
*New York University Tandon School of Engineering; +IBM T.J. Watson Research Center

Why and when is interpretation needed?

1. Data understanding and discovery.
2. Trust building and accountability.
3. Model comparison and diagnostics.

Model Transparency, Representation, and Interpretability

- *Model structure* does not imply *model representation* and *interpretability*.
- Models can be understood by looking at behavior without structure (**Prospector** and **Class Signatures**)
- Interpretability vs. accuracy false dichotomy?



Regression models can express only a single slope (downwards or upwards) whereas the random forest can model the strong decrease in predicted risk going from no BMI measures to one as well as the later increase again if a patient has several BMI measures. The distribution of input values in the histogram below the plot, however, hint the model might be overfitting as most of the observed values are 2 or less.

The Role of Visual Analytics in Interpretation

- Visual Analytics enables human involvement.
- Human involvement needed for interpretation.

Two approaches for model interpretation with visual analytics:

1. Visualizing Model Structure (**White-Box**).
2. Visualizing Model Behavior (**Black-Box**).

Interpretation with In/Out Model Behavior

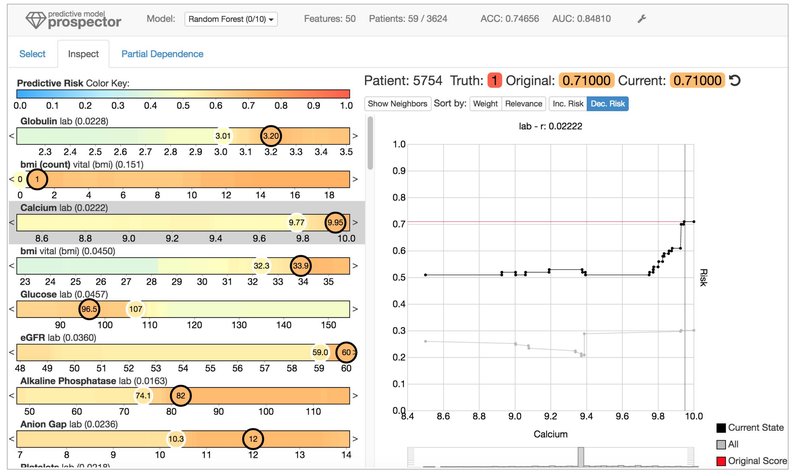
- Extremely flexible and generic.
- Independent from model representation.
- Input/output behavior can be obtained using: *training data*, *test data*, or *simulated data*.

Three main mechanisms for I/O model behavior:

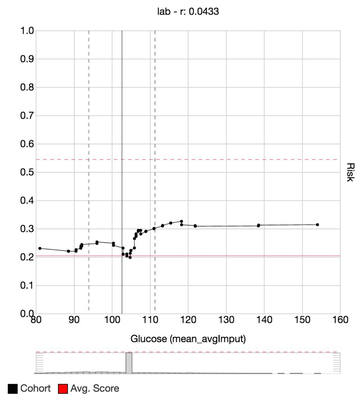
1. Item(s) to outcome
2. Single feature to outcome
3. Multiple features to outcome

Example 1: Prospector

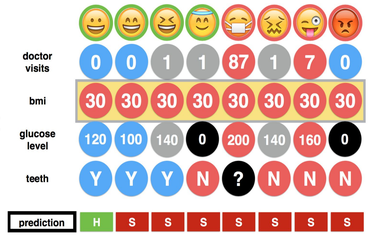
- Designed to help understand predictive models.
- Making partial dependence (Friedman, 2001) fully interactive.
- Localized inspection to understand prediction results.
- Interactively tweak feature values and see the prediction respond.
- Find the most impactful features using a novel model agnostic local feature importance metric that only depends on partial dependence.

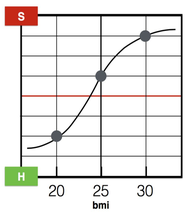


Identifying changes to features that reduce the risk of a high risk patient. The features (sliders) are sorted by decreasing impact to make large sets of features manageable. The background color of each feature indicates the predicted risk for this value. The system suggests the most impactful changes which appear as white circles.



Instead of a direct relationship between higher Glucose lab values (x-axis) and higher risk scores (y-axis) the model predicts a low risk for average Glucose lab values. Since missing values are imputed using the average Glucose value the valley in the plot can be explained by the outcome independence of this value due to the high number of missing values.





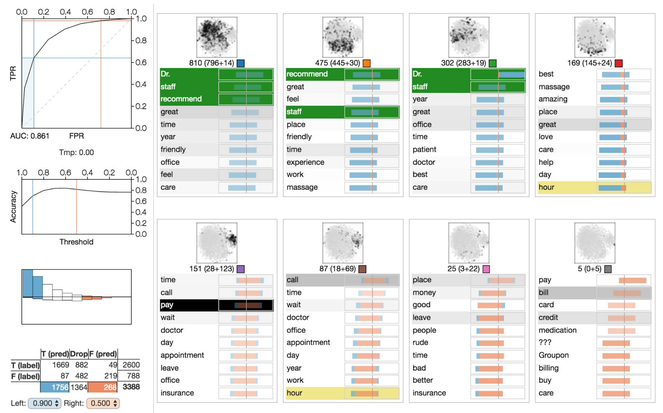
7/8 are predicted sick

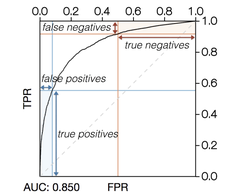
$$pdp_f(v) = \frac{1}{N}\sum_i^N pred(x_i) \text{ with } x_{if} = v$$

N is the number of rows in the input matrix x, pred is the prediction function that takes one input row, a feature vector, and returns a prediction score, and f is the feature used to compute the partial dependence plot (left).

Example 2: Class Signatures

Main goal: create visual signatures able to explain associations between input and output values as captured by a model.





Determining filter thresholds with receiver operating characteristics curve for. The curve shows how the true positive rate (TPR; y-axis) and the false positive rate (FPR; x-axis) changes with thresholds (threshold values are not shown in the graph).

Main Steps

1. **Model** predictive associations using a binary classifier
2. **Contrast** prediction scores with two thresholds to focus only on data items with a strong predictive signal.
3. **Cluster** both positive and negative examples *separately*.
4. **Rank** features in the computed clusters using discriminative analysis across *all clusters*.
5. **Interpret** the results with visual analytics

Each column represents one group (*cluster*-step) whereas each row shows the amount of patients in this group taking a particular medication (the bar from the middle towards the right shows the percentage of patients taking the medication; the bar towards the left shows not taking medication). The color of the bar shows the distribution of the true outcome labels as found in the input data. The background of the rows shows the discriminativeness of a medication (dark being more discriminative *wrt.* all other clusters; *rank*-step). Above each group a *t-SNE* projection of the items shows its relation to the other groups.

References

L. Breiman. Statistical modeling: The two cultures. *Statistical Science*,16(3):199–231, 08 2001.
Munzner, Tamara. Visualization Analysis and Design. A K Peters/CRC Press, Natick, MA, USA, 2014. ISBN 9781466508910.
Krause, Josua, Perer, Adam, and Ng, Kenney. Interacting with predictions: Visual inspection of black-box machine learning models. ACM CHI 2016, 2016.
Friedman, Jerome H. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5): 1189–1232, Oct 2001.