Visualizations For Justifying Machine Learning Predictions

David Johnson

Motivation

- Strengths of ML allowed expansion to diverse fields
- Fields and contexts far removed from traditional ML
- Users not trained in ML
  - Eg. Medical field: Doctors use ML to predict disease given symptoms
  - The ML is a black box to them: Input → ? → Output

$$\begin{aligned}\text{maximize } f(c_1 \ldots c_n) &= \sum_{i=1}^{n} c_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i c_i (\varphi(\vec{x}_i)\cdot\varphi(\vec{x}_j)) y_j c_j \\ &= \sum_{i=1}^{n} c_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i c_i k(\vec{x}_i, \vec{x}_j) y_j c_j\end{aligned}$$

$$\text{subject to } \sum_{i=1}^{n} c_i y_i = 0, \text{ and } 0 \le c_i \le \frac{1}{2n\lambda} \text{ for all } i.$$

Previous Work

The prediction, given by Linear Regeression, is **Y**

The most important evidence for the prediction is in SLOPE and Y_PRIOR. This is normal, as these features are often important for predictions of this class.

Normally, we would see powerful counter-evidence in DIAMETER, but it is missing in this case.

Significant counter-evidence exists in VENUE. This is exceptional, as it is not usually a strong feature.

Key feature list:
- SLOPE (Normal evidence)
- Y_PRIOR (Normal evidence)
- DIAMETER (Missing counter-evidence)
- VENUE (Exceptional counter-evidence)

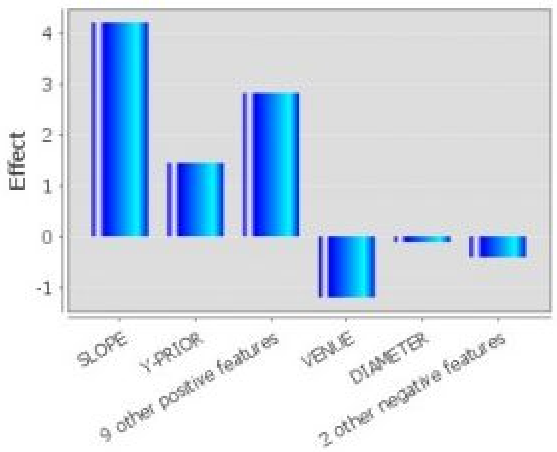


Figure: Biran, O., MckKeown, K. (2014). Justification Narratives for Individual Classifications. *AutoML workshop at ICML 2014*.

Previous Work

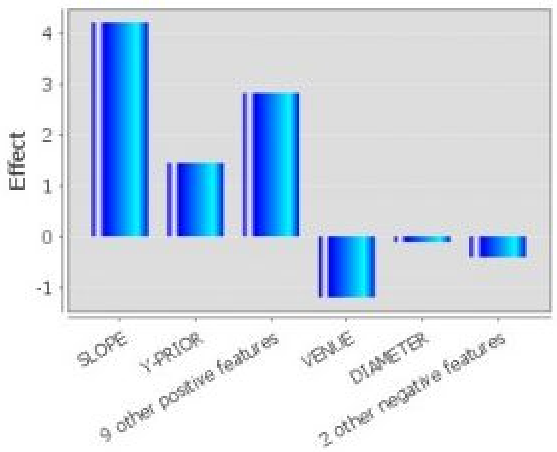
The prediction, given by Linear Regeression, is Y

The most important evidence for the prediction is in SLOPE and Y_PRIOR. This is normal, as these features are often important for predictions of this class.

Normally, we would see powerful counter-evidence in DIAMETER, but it is missing in this case.

Significant counter-evidence exists in VENUE. This is exceptional, as it is not usually a strong feature.

Key feature list:
- SLOPE (Normal evidence)
- Y_PRIOR (Normal evidence)
- DIAMETER (Missing counter-evidence)
- VENUE (Exceptional counter-evidence)



Some issues:

- The vis relies on NLG quite a bit
- Vis isn't very clear for non-experts (what is Y-Prior? What is Slope?)

Figure: Biran, O., MckKeown, K. (2014). Justification Narratives for Individual Classifications. *AutoML workshop at ICML 2014*.

Goals

- Justify a ML prediction to a non-expert user
- Show features providing evidence for/against the prediction
- Select and visualize key features
- Focus on interpretable models
- Simplicity not complexity...

Figure: Munzner, T. (2014). Visualization Analysis and Design. CRC Press.

Goals

- Justify a ML prediction to a non-expert user
- Show features providing evidence for/against the prediction
- Select and visualize key features
- Focus on interpretable models
- Simplicity not complexity...

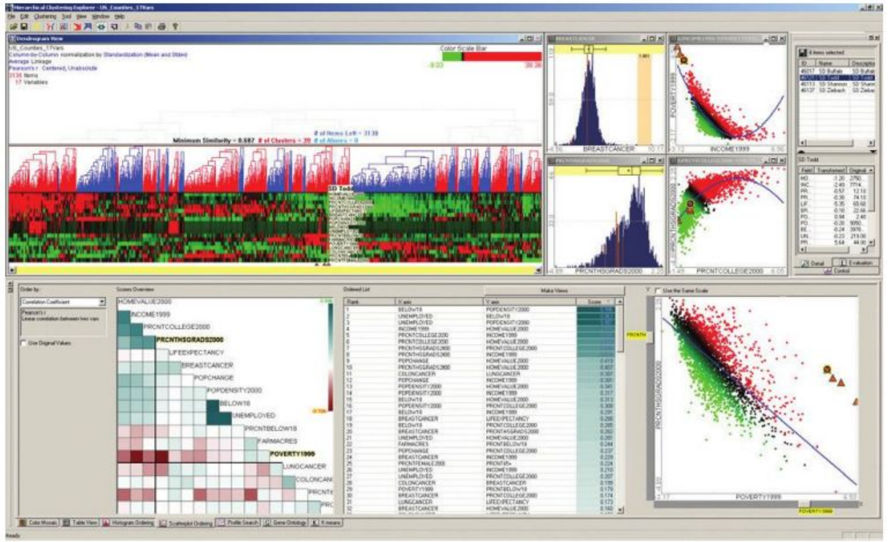


Figure: Munzner, T. (2014). Visualization Analysis and Design. CRC Press.

Feature Visualizing

Vis can show effect and importance[1]

- Effect: extent to which a feature contributes toward or against prediction

$$\text{Effect}_{ji} = \theta_{ji} x_i$$

- Importance: Expected effect of the feature for a particular class (mean feature value for the class)

$$\text{Importance}_{ji} = \theta_{ji} \frac{\sum_{x \in X^j} x_i}{|X^j|}$$

[1]Biran, O., MckKeown, K. (2014). Justification Narratives for Individual Classifications. *AutoML workshop at ICML 2014*.

Abstraction

- Some raw data: arbitrary data with training/test sets
- Task abstraction:
  - Analyze: discover, enjoy, derive
- Data abstraction:
  - Items, attributes, values in a table
- Two quantitative variables: effect, importance -- scatterplot effective

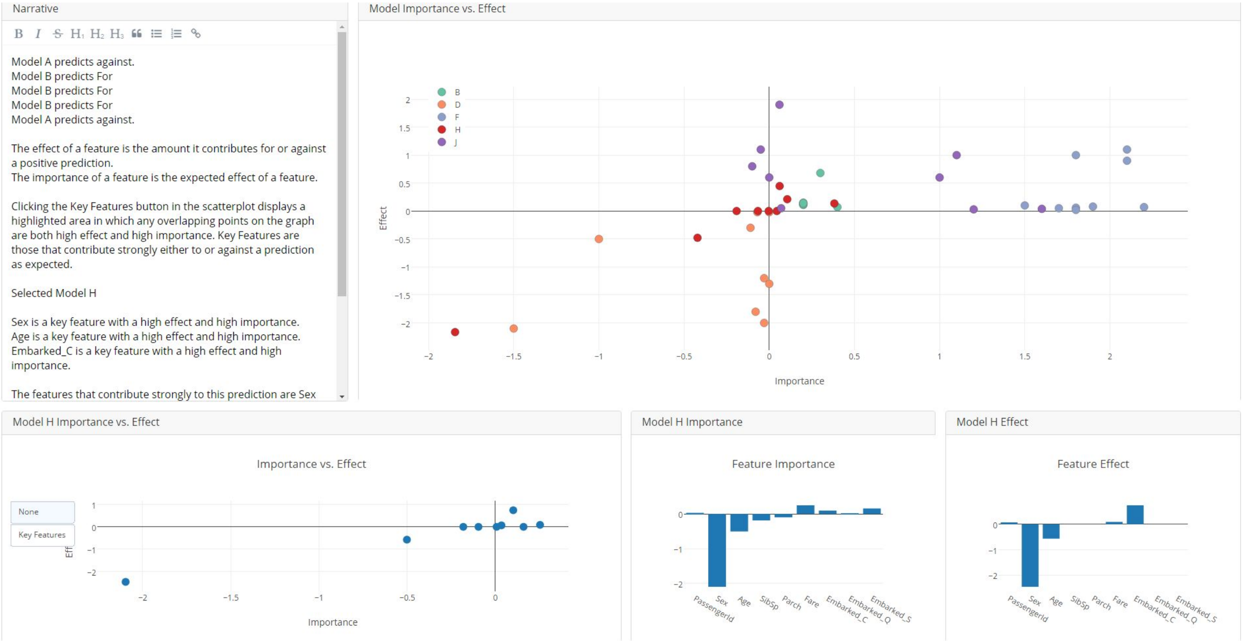
Demo

Future Direction

NLG implemented

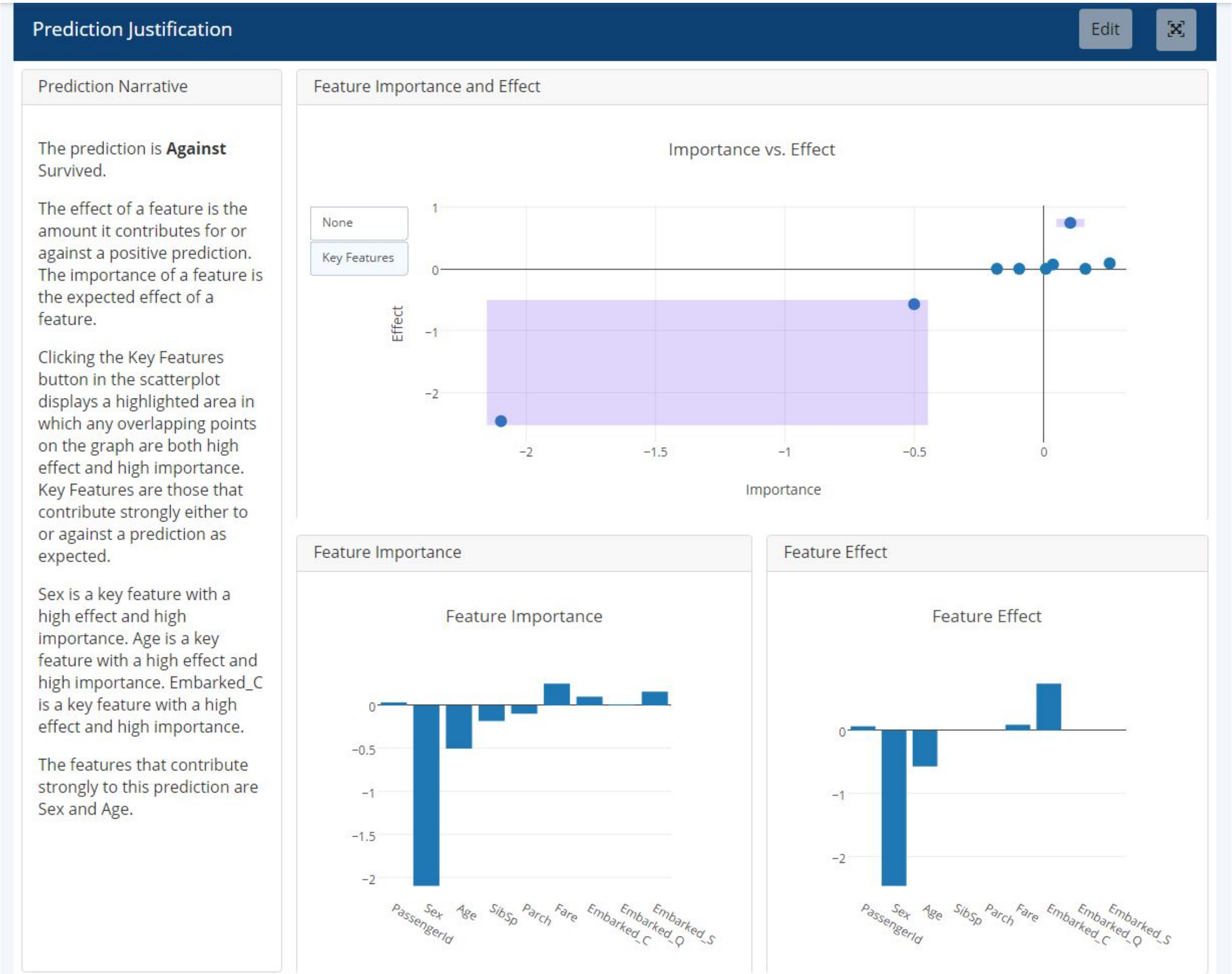
Full web app implementation

Expanded scope:



Thanks!

Questions?


Prediction Justification
Edit
Prediction Narrative
The prediction is **Against** Survived.
The effect of a feature is the amount it contributes for or against a positive prediction. The importance of a feature is the expected effect of a feature.
Clicking the Key Features button in the scatterplot displays a highlighted area in which any overlapping points on the graph are both high effect and high importance. Key Features are those that contribute strongly either to or against a prediction as expected.
Sex is a key feature with a high effect and high importance. Age is a key feature with a high effect and high importance. Embarked_C is a key feature with a high effect and high importance.
The features that contribute strongly to this prediction are Sex and Age.
Feature Importance and Effect
Importance vs. Effect
None
Key Features
Effect
Importance
Feature Importance
Feature Importance
Feature Effect
Feature Effect
PassengerId
Sex
Age
SibSp
Parch
Fare
Embarked_C
Embarked_Q
Embarked_S