Prospective Explanations: An Interactive Mechanism for Model Understanding

Rahul Nair RAHUL.NAIR@IE.IBM.COM and **Pierpaolo Tommasi** PTOMMASI@IE.IBM.COM
IBM Research Europe

Editor: Douwe Kiela, Marco Ciccone, Barbara Caputo

Abstract

We demonstrate a system for prospective explanations of black box models for regression and classification tasks with structured data. Prospective explanations are aimed at showing how models function by highlighting likely changes in model outcomes under changes in input This in contrast to most post-hoc explanability methods, that aim to provide a justification for a decision retrospectively. To do so, we employ a surrogate Bayesian network model and learn dependencies through a structure learning task. Our system is designed to provide fast estimates of changes in outcomes for any arbitrary exploratory query from users. Such queries are typical partial, i.e. involve only a selected number of features, the outcomes labels are shown therefore as likelihoods. Repeated queries can indicate which aspects of the feature space are more likely to influence the target variable. We demonstrate the system from a real-world application from the humanitarian sector and show the value of bayesian network surrogates.

Keywords: Explainability; Bayesian networks; Interactive AI

1. Setting

We demonstrate a system[1] for prospective explanations of black box models for regression and classification tasks with structured data. Prospective explanations are aimed at showing how models function by highlighting likely changes in model outcomes under changes in input (Shneiderman, 2020). A desired property of explanations is to help with creating the right mental model of an AI system. The right mental model leads to greater trust (Bansal et al., 2019). Interactive systems that allow for exploration have been shown to improve user comprehension (Cheng et al., 2019) albeit being more time consuming.

To achieve fast interactive exploration, we build a surrogate Bayesian network model. A surrogate here implies the use of model labels instead of ground truth labels to represent model behaviour. Bayesian network models are stored as directed acyclic graphs where links represent dependence between variables. This graph representation can be learnt directly from the data through a structure learning task, or be provided externally and only conditional probabilities estimated from the data. Bayesian networks are efficient in storing the joint distribution over feature sets and allow for fast inference over arbitrary queries.

Several previous works have sought present model-related metrics based on user inputs. In FairVis (Cabrera et al., 2019), take an intersectional fairness view, showing changes in accuracy and other metrics for changes in user specified cohorts. Gleicher et al. (2020) provide a mechanism for users to compare models by interactively looking at differing model performance for different models. Spinner et al. (2020) present a holistic framework that combines various explainability algorithms at different steps of the pipeline. Here we take a global surrogate view (Molnar, 2020), where the Bayesian network serves as the surrogate.

[1] https://prosp-exp.eu-gb.mybluemix.net/

2. System design

Specifically, we seek to explain a model f that maps an input vector $\mathbf{x}_i$ to an output y_i. Consider a validation dataset D, a set of $(\mathbf{x}_i, y_i)$ observations for $i = 1, \ldots, n$. For this validation data, we generate a set of labels $\hat{y}_i = f(\mathbf{x}_i)$. A structure learning algorithm is used, treating both features $\mathbf{x}_i$ and labels $\hat{y}_i$ as random variables to learn a graph $G(V, E)$ and associated probability tables. During inference, users provide arbitrary feature values, and the marginal distribution of the target class estimated using the formula

$$P(\underbrace{x_1 = a_1, x_2 = a_2, \ldots, x_k = a_k,}_{\text{user provided input}} y = y_1) = \prod_{v=1}^{n} P(x_v = a_v \mid x_j = a_j \forall j \in pa(v)), \qquad (1)$$

where $pa(v)$ denotes all parents of a node v in G. While inference in Bayesian networks is NP-Hard, assumptions on the structure of G admits fast inference in practice. We experimented with four strategies in particular. First, the network structure was defined *a priori* from expert opinion leading to a lower node degree. Second, we experimented with limiting the number of parents for each node. Third, we learnt the structure on a limited set of important features as determined by feature important scores. Lastly, the network is training on a subset of features used in f. In our experiments, these approaches did not degrade performance of the surrogate model substantively. See Table 1 for preliminary results.

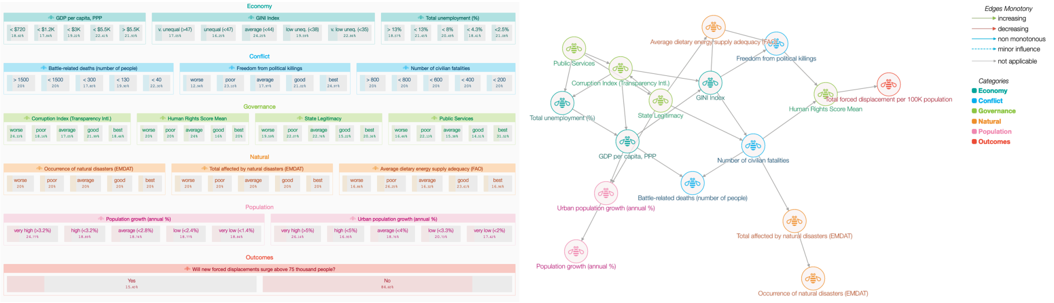
The main advantages of our approach are that (a) inference using Eq. 1 is very fast and supports real-time feedback allowing for interactivity, (b) inference can be done with partial information on features, and (c) any indirect effects are also considered in estimating target class distributions. Regression models involve an additional consideration. The target variable $\hat{y}_i$ is discretized before learning the structure. This is necessary to avoid the difficulty of perceiving changes in continuous probability distributions.

2.1. User oriented aspects

Targeting a wide and multidisciplinary audience for real-world machine learning application poses several challenges. Concepts like acyclic graphs and conditional probabilities tables can be difficult to grasp for non-experts. Even detailed visualizations are likely to add complexity. In our design we explicit focus on the data rather than the model, enabling deeper exploration of the latter only when needed. This visualization approach has been previously used to communicate Bayesian networks to domain experts (Tommasi et al., 2019).

This data-first view translates the model to a glossary that domain experts are accustomed to. The user experience includes several visual cues, such as feature clustering by color, and histograms to show marginal distributions. Specifically, for every user interaction, i.e. a click, the marginal distributions for all features are updated. This allows users to build mental models of how features are related, regardless of the model. This ‘Feature Board’ is shown in Figure 1. A similar histogram view is presented for the model-predicted outcomes, where additional baselines (typically based on ground truth) can be shown to highlight if the model deviates from observations.

The second view in Figure 1 (right) shows the graphical view of the Bayesisn networks with all the learnt dependent relationships between variables. This graph view is *not* meant to imply causality, but an information theoretic view on a structure that best explains the data.



Economy

GDP per capita, PPP: < $720 18.60% · < $1.2K 17.84% · < $3K 19.22% · < $5.5K 22.41% · > $5.5K 21.93%

GINI Index: v. unequal (>47) 17.05% · unequal (<47) 16.25% · average (<44) 24.25% · low uneq. (<38) 19.59% · v. low uneq. (<35) 22.86%

Total unemployment (%): > 13% 18.57% · < 13% 21.45% · < 8% 20.49% · < 4.3% 18.41% · <2.5% 21.08%

Conflict

Battle-related deaths (number of people): > 1500 20% · < 1500 20% · < 300 17.80% · < 130 19.90% · < 40 22.30%

Freedom from political killings: worse 12.94% · poor 23.10% · average 17.97% · good 21.02% · best 24.97%

Number of civilian fatalities: > 800 20% · < 800 20% · < 600 20% · < 400 20% · < 200 20%

Governance

Corruption Index (Transparency Intl.): worse 24.33% · poor 18.16% · average 17.05% · good 21.99% · best 18.46%

Human Rights Score Mean: worse 20% · poor 20% · average 24% · good 16% · best 20%

State Legitimacy: worse 19.59% · poor 22.07% · average 22.76% · good 15.22% · best 20.36%

Public Services: worse 16.46% · poor 22.13% · average 15.38% · good 14.51% · best 31.52%

Natural

Occurrence of natural disasters (EMDAT): worse 20% · poor 20% · average 20% · good 20% · best 20%

Total affected by natural disasters (EMDAT): worse 20% · poor 20% · average 20% · good 20% · best 20%

Average dietary energy supply adequacy (FAO): worse 16.86% · poor 26.25% · average 16.32% · good 23.61% · best 16.96%

Population

Population growth (annual %): very high (>3.2%) 24.77% · high (<3.2%) 18.89% · average (<2.8%) 18.74% · low (<2.4%) 18.77% · very low (<1.4%) 18.84%

Urban population growth (annual %): very high (>5%) 26.14% · high (<5%) 16.95% · average (<4%) 18.76% · low (<3.3%) 20.73% · very low (<2%) 17.42%

Outcomes

Will new forced displacements surge above 75 thousand people?: Yes 15.40% · No 84.60%

Average dietary energy supply adequacy (FAO)
Public Services
Freedom from political killings
Corruption Index (Transparency Intl.)
GINI Index
Total unemployment (%)
State Legitimacy
Human Rights Score Mean
Total forced displacement per 100K population
GDP per capita, PPP
Number of civilian fatalities
Battle-related deaths (number of people)
Urban population growth (annual %)
Total affected by natural disasters (EMDAT)
Population growth (annual %)
Occurrence of natural disasters (EMDAT)
Edges Monotony
increasing
decreasing
non monotonous
minor influence
not applicable
Categories
Economy
Conflict
Governance
Natural
Population
Outcomes

Figure 1: Main views of the system (a) Feature board view and a (b) Graph view

3. Demonstration use case

We have deployed the system for several health and social care domains. Here we demonstrate it for a humanitarian application that deals with forced displacement. We explain two model classes for two tasks, one classification and one regression. Forced displacement refers to the involuntary movement of people away from their homes. This sector is highly complex, and a wide range of factors can potentially influence the onset and severity of displacement crisis. Responding organizations have to plan under high uncertainty on the type of humanitarian need. Previous user-studies for this sector has uncovered the need to surface this uncertainty Andres et al. (2020) which is both aleatory (uncertainty due to things that cannot be predicted) and epistemic (uncertainty due to missing information).

While the regression task is to forecast volume of displaced persons, the classification task seeks to classify if the volume is likely to exceed a 'crisis' threshold (the UNHCR informally deems a crisis when displacement volumes are in excess of 75,000 persons). For each task, prospective explanations of two models each that are training on data from a period from 1980 through 2015. Data after 2015 is used for the validation set D. For classification, a logistic regression with validation accuracy (acc) 97.2% and a gradient boosted classifier (acc: 98.03%) are shown. Both models exhibit poor precision for the minority class. Compared to ground truth reference, the models are less likely to predict a crisis.

For regression we show a linear regression (MAE: 2086.03) and a gradient boosted regressor (MAE: 1155.69). For the latter, Figure 2 shows baseline performance (overestimates for 100-5000 ranges, and underestimates for > 5000). When uses toggle the human rights dimension to 'worse', predictions for > 5000 better match ground truth labels. While this estimator is highly non-linear, the interface the impact of influencing factors can be understood by interacting with features values and visual representation of outcome marginals.

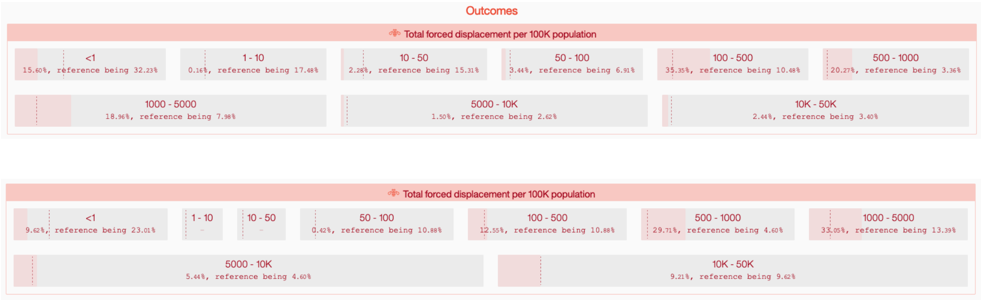

Outcomes
Total forced displacement per 100K population
<1
15.60%, reference being 32.23%
1 - 10
0.16%, reference being 17.48%
10 - 50
2.28%, reference being 15.31%
50 - 100
3.44%, reference being 6.91%
100 - 500
35.35%, reference being 10.48%
500 - 1000
20.27%, reference being 3.36%
1000 - 5000
18.96%, reference being 7.98%
5000 - 10K
1.50%, reference being 2.62%
10K - 50K
2.44%, reference being 3.40%
Total forced displacement per 100K population
<1
9.62%, reference being 23.01%
1 - 10
-
10 - 50
-
50 - 100
0.42%, reference being 10.88%
100 - 500
12.55%, reference being 10.88%
500 - 1000
29.71%, reference being 4.60%
1000 - 5000
33.05%, reference being 13.39%
5000 - 10K
5.44%, reference being 4.60%
10K - 50K
9.21%, reference being 9.62%

Figure 2: Example label histograms for regression tasks under an interactive scenario (a) Baseline prediction (top) and (b) scenario when human rights are 'worse' (bottom).

dataset	ML Model	ML acc.	BN acc.	training time	inference time
UCI Adult	Logistic	0.851913	0.913826	1.472805	0.0130
	XGBoost	0.871138	0.921688	1.637081	0.0126
	Random Forest	0.844666	0.865242	1.517839	0.0129
Bank marketing	Logistic	0.915271	0.948653	6.033707	0.0214
	XGBoost	0.923768	0.950352	5.858732	0.0220
	Random Forest	0.916849	0.941491	5.971814	0.0199

Table 1: Accuracy of a global surrogate Bayesian network model on two additional datasets, where the graph was limited to a maximum of two parent nodes. Training time (seconds) shows the time needed to learn the Bayesian network, and the inference time (seconds) is the time needed for per instance inference.

4. Discussion

We experimented with several choices in the design. Results on other public datasets show generally strong accuracy for the surrogate on unseen validation datasets. Table 1 summarizes sample results. Barring the one case of the Random Forest with the adult dataset, where accuracy was 86.5%, the surrogates are able to mimic underlying complex models well. Inference time, computed as the average of time taken for inference of the entire test set, shows the possibility of interactive applications.

Additional experiments and user studies are needed to show if mental models and decisions can be influenced through such interactive systems. In particular, such explorations could be used to highlight unsafe regions of operation for an ML model, particularly if the user has prior knowledge about the domain.

References

Josh Andres, Christine T Wolf, Sergio Cabrero Barros, Erick Oduor, Rahul Nair, Alexander Kjærum, Anders Bech Tharsgaard, and Bo Schwartz Madsen. Scenario-based XAI for humanitarian aid forecasting. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2020.

Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2429–2437, 2019.

Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. Fairvis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 46–56, 2019. doi: 10.1109/VAST47406.2019.8986948.

Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. *Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders*, page 1–12. Association for Computing Machinery, New York, NY, USA, 2019. ISBN 9781450359702. URL https://doi.org/10.1145/3290605.3300789.

Michael Gleicher, Aditya Barve, Xinyi Yu, and Florian Heimerl. Boxer: Interactive comparison of classifier results. In *Computer Graphics Forum*, volume 39, pages 181–193. Wiley Online Library, 2020.

Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

Ben Shneiderman. Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4):1–31, 2020.

Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady. explAIner: A visual analytics framework for interactive and explainable machine learning. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1064–1074, 2020. doi: 10.1109/TVCG.2019.2934629.

Pierpaolo Tommasi, Stephane Deparis, and Alessandra Pascale. HWProfile UI: facilitating the exploration of a patient centered risk model. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–2, 2019. doi: 10.1109/ICHI.2019.8904573.