Explaining Decision Forests

Julian Hatwell

Birmingham City University

julian.hatwell@bcu.ac.uk

August 7, 2019

About Me

- Over 15 years in enterprise systems, management information systems and BI in the for-profit education sector, UK and Singapore
- MSc (Distinction) Business Intelligence, Birmingham City University. Master's Dissertation: *An Association Rules Based Method for Imputing Missing Likert Scale Data*
- Research to PhD (in progress): *Designing Explanation Systems for Decision Forests*, Data Analytics and Artifical Intelligence research group, Birmingham City University

Introduction

Considerations

CHIRPS

Conclusion

Machine Learning as a Black Box

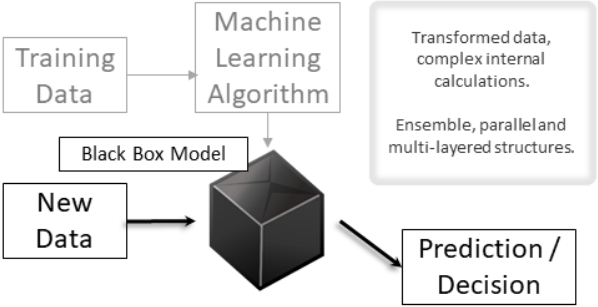
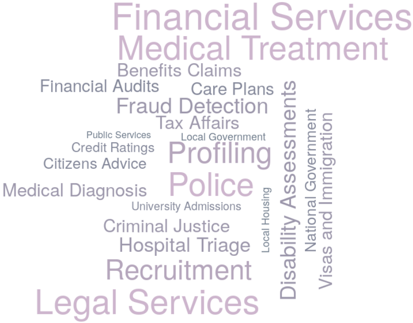

Training Data
Machine Learning Algorithm
Transformed data, complex internal calculations.
Ensemble, parallel and multi-layered structures.
Black Box Model
New Data
Prediction / Decision

Figure 1: Modern ML models such as neural nets and decision forests. Explaining how they make each decision is very difficult.

The Unrelenting Automation Knowledge Work



In the next decade, AI and ML technologies are set to transform "the professions" Susskind and Susskind 2015.

eXplainable Artificial Intelligence (XAI)

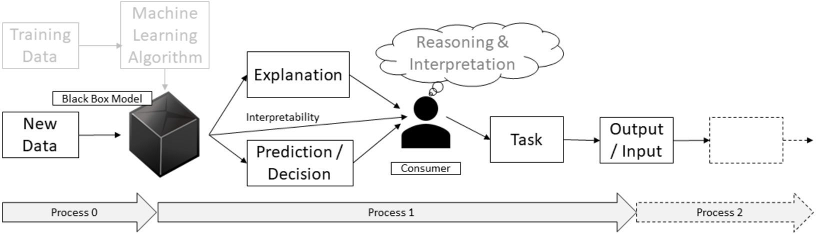


Figure 2: Process 0 submit new data to the trained model, Process 1 is any process where the consumer must interpret the model output, and Process 2 is some downstream task requiring the consumer's judgement.

Introduction

Considerations

CHIRPS

Conclusion

Quantifying Interpretability and Modelling Explanations

Interpretability and explanations are very poorly defined in the ML literature Lipton 2016.

Formalising and then meeting clear success factors has to be a requirement of XAI research.

Quantifying Interpretability and Modelling Explanations

Miller 2017 proposes modeling *the way humans explain their decisions and behaviour to one another* with some simple guiding principles. Explanations should:

- Refer to observations, not abstractions.
- Generalise well and be relevant to many examples.
- Be minimally complete.
- Contrast factual and counter-factual cases.
- Consider context;
  - Why is an explanation required for a given application?
  - What is the downstream process?

Global vs. Local

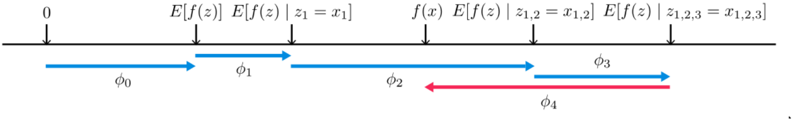
Global Interpretability

- RuleFit Friedman and Popescu 2008
- Bayesian Rule Lists Letham et al. 2015
- defragTrees Hara and Hayashi 2016
- inTrees, Deng 2014
- Soft decision trees Frosst and Hinton 2017

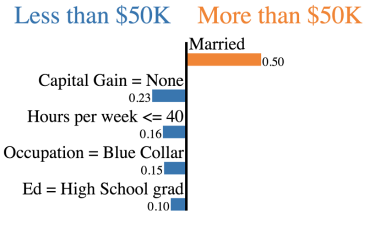
Local Explanations

- treeinterpreter, Sabaas 2015
- LIME Ribeiro et al. 2016
- SHAP Lundberg and Lee 2017
- Anchors Ribeiro et al. 2018
- LORE Guidotti et al. 2018

Local Linear Models vs Classification Rules

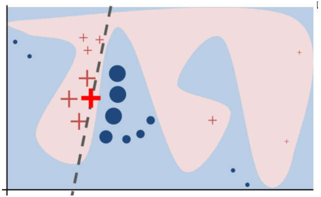


(a) Image source: Lundberg and Lee 2017

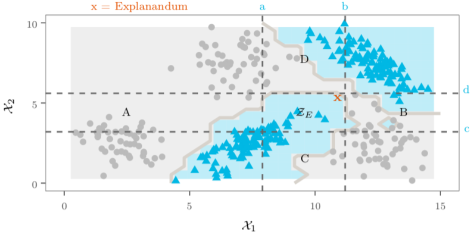


(b) Image source: Ribeiro et al. 2018

Local Linear Models vs Classification Rules



(a) Image source: Ribeiro et al. 2016



(b) Rule terms are axis aligned cuts

Model Agnostic vs Model Specific

Anchors, LORE, SHAP and LIME are all model agnostic.

- Works with any model?
- Requires access to model inputs and outputs only?

but...

- These assumptions are violated for tabular data Ribeiro et al. 2018; Michal 2019.
- Such explanations have been shown to be unstable Fen et al. 2019.
- The scenarios that require a model agnostic method may be quite rare, e.g. third party audit

Introduction

Considerations

CHIRPS

Conclusion

Collection of High Importance Random Path Segments

Relax the model-agnostic assumption

Why do we need to be model-agnostic if we own the model and the training data?

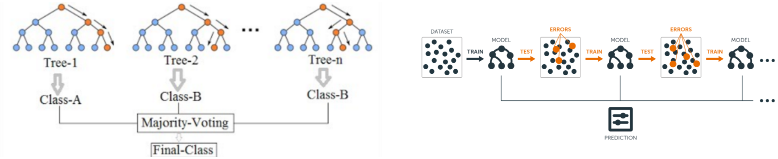
Implement Classification Rules as Explanations

Explanations from additive methods (LIME, SHAP) are not easily transferable to other unseen instances. Contrastingly, rule based explanations are unambiguous based on coverage.

Develop methods for Decision Forests

Decision Forests are very easy to deploy with high accuracy out of the box, therefore popular among ML practitioners. They also serve as a pool of potentially useful classification rules.

Decision Forests



(a) Source: Wikipedia

(b) Source: BigML blog

Figure 5: Decision Forests: Random Forests and Boosted Models

Ensembles of many (10's-1000's) of decision trees, each generated from random starting conditions. It is a sample of a tree-structured random variable, *parameterised by the training data*. The statistical properties of the sample reveal useful information about the model (Paluszynska 2017; Louppe 2014; Biau 2012)

Theoretical Basis

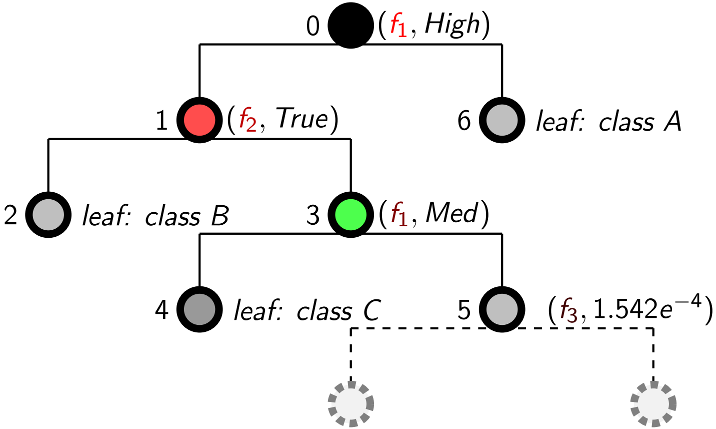


Figure 6: The most discriminative attributes tend to appear at shallower nodes. The probability of visiting a node diminishes exponentially with node depth.

Theoretical Basis

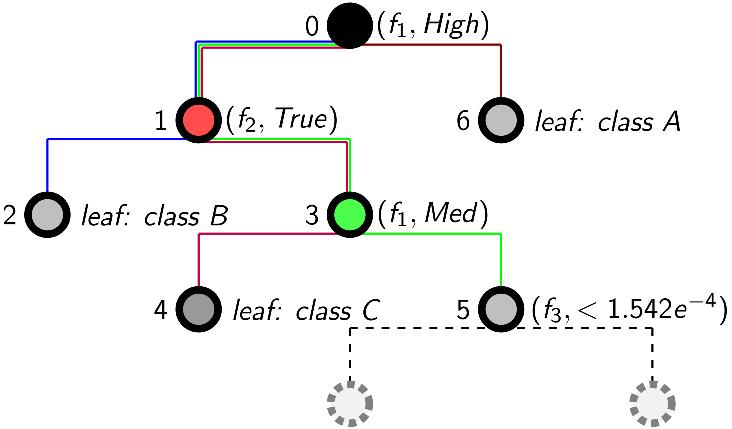
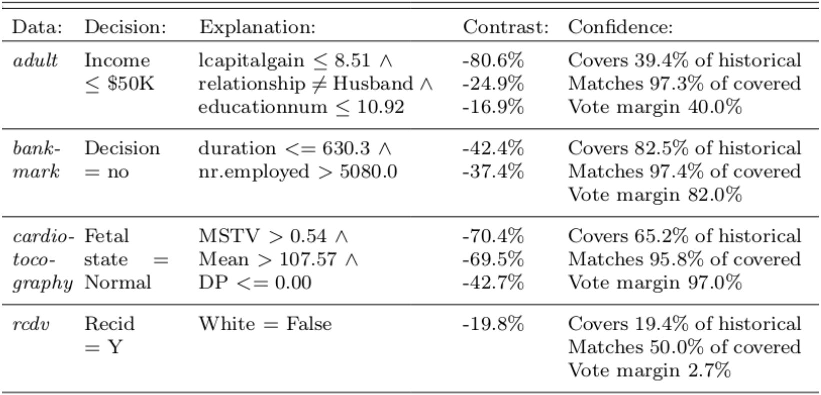


Figure 7: Attributes that discriminate well on interaction will tend to appear in the same path. Similar instances pass through the same nodes with high probability, leading to good generalisation.

Collection of High Importance Random Path Segments

1. Traverse instance through all the trees
2. Extract the path taken in each tree
3. Discard any that didn't vote with the majority
4. Perform Frequent Pattern Mining - paths as transactions, decision nodes as items
5. Greedily merge highest scoring (support and KL-divergence) into a set X
6. Explanation is the logical rule $X \implies Y$ where Y is the predicted class.

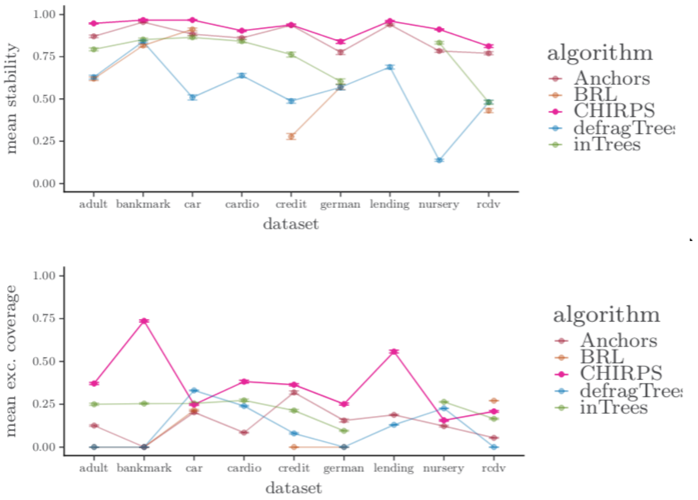
CHIRPS Output



Data:	Decision:	Explanation:	Contrast:	Confidence:
adult	Income $\leq$ \$50K	lcapitalgain $\leq 8.51 \wedge$ relationship $\neq$ Husband $\wedge$ educationnum ≤ 10.92	-80.6% -24.9% -16.9%	Covers 39.4% of historical Matches 97.3% of covered Vote margin 40.0%
bank-mark	Decision = no	duration $<= 630.3 \wedge$ nr.employed > 5080.0	-42.4% -37.4%	Covers 82.5% of historical Matches 97.4% of covered Vote margin 82.0%
cardio-toco-graphy	Fetal state = Normal	MSTV $> 0.54 \wedge$ Mean $> 107.57 \wedge$ DP $<= 0.00$	-70.4% -69.5% -42.7%	Covers 65.2% of historical Matches 95.8% of covered Vote margin 97.0%
rcdv	Recid = Y	White = False	-19.8%	Covers 19.4% of historical Matches 50.0% of covered Vote margin 2.7%

These outputs meet Miller's guidelines and well-defined success criteria for explanations.

CHIRPS Benchmarking



Next Steps

- Apply the method to AdaBoost (2 variants) and Gradient Tree Boosting
- Develop and enhance the information output for the end user, including visualisations

End of Presentation

Thank You

Questions & Discussion

References I

- Biau, Gerard (2012). "Analysis of a Random Forests Model". In: *Journal of Machine Learning Research* 13, pp. 1063–1095.
- Deng, Houtao (2014). "Interpreting tree ensembles with intrees". In: *arXiv preprint arXiv:1408.5456*.
- Fen, Hui et al. (2019). "Why should you trust my interpretation? Understanding uncertainty in LIME predictions". In: *arXiv:1904.12991*.
- Friedman, Jerome and Bogdan E Popescu (2008). "Predictive Learning via Rule Ensembles". In: *The Annals of Applied Statistics* 2.3, pp. 916–954.
- Frosst, Nicholas and Geoffrey Hinton (2017). "Distilling a Neural Network Into a Soft Decision Tree". In: *arXiv preprint arXiv:1711.09784*.
- Guidotti, Riccardo et al. (2018). "Local Rule-Based Explanations of Black Box Decision Systems". In: *arXiv:1805.10820*.

References II

- Hara, Satoshi and Kohei Hayashi (2016). "Making Tree Ensembles Interpretable: A Bayesian Model Selection Approach". In: *arXiv:1606.09066 [stat]*.
- Letham, Benjamin et al. (2015). "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model". In: *The Annals of Applied Statistics* 9.3, pp. 1350–1371.
- Lipton, Zachary Chase (2016). "The mythos of model interpretability". In: *arXiv preprint arXiv:1606.03490*.
- Louppe, Gilles (2014). "Understanding Random Forests: From Theory to Practice". PhD thesis. Liège, Belgium: Université de Liège.
- Lundberg, Scott M and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*, pp. 4768–4777.

References III

- Michal, Fracek (2019). *"Please, explain." Interpretability of black-box machine learning models*. URL: https://tinyurl.com/y5qruqgf (visited on 04/19/2019).
- Miller, Tim (2017). "Explanation in Artificial Intelligence: Insights from the Social Sciences". In: *arXiv preprint arXiv:1706.07269*.
- Paluszynska, Aleksandra (2017). "Structure mining and knowledge extraction from random forest with applications to The Cancer Genome Atlas project". PhD thesis. University of Warsaw.
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). "Why Should I Trust You?: Explaining the Predictions of Any Classifier". In: ACM Press, pp. 1135–1144.
- – (2018). "Anchors: High-Precision Model-Agnostic Explanations". In: *Conference on Artificial Intelligence, 2018*. New Orleans.

References IV

- Sabaas, Ando (2015). *Interpreting Machine Learning Models (Slideshare)*. URL: https://www.slideshare.net/andosa/interpreting-machine-learning-models (visited on 10/11/2017).
- Susskind, Richard E. and Daniel Susskind (2015). *Susskind, Richard E., and Daniel Susskind. The future of the professions: How technology will transform the work of human experts.*. USA: Oxford University Press.