Contextual Model Interpretation: A Model Agnostic Approach to Search For Explainable Data Sub-Sets

Karan Samel and Xu Chu
School of Computer Science, Georgia Tech

Background

Within model agnostic explanations, there are two categories of explanations:

1. Global interpretation over all the data [1].
2. Local interpretation on specific data instances [2].

Contextual Model Interpretation (CMI) provides a middle ground, where explanations are made on subsets, or contexts, of data. This aids users to find scenarios where their model behavior changes, and thus may be **incorrect or biased**.

Framework

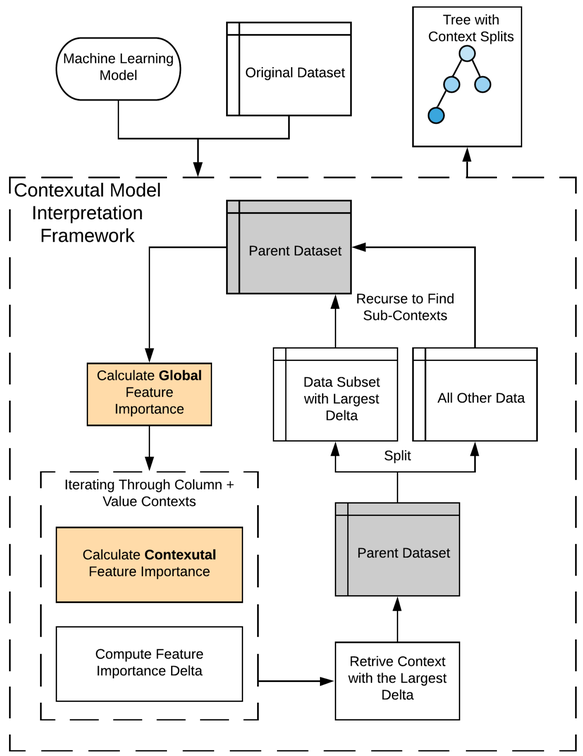

Fig. 1: Interesting contexts are found by recursively finding the best context with respect to the change in feature importance

Results

We applied CMI to the FICO Explainable Machine Learning Data. Here are the corresponding context splits for the first 3 levels.

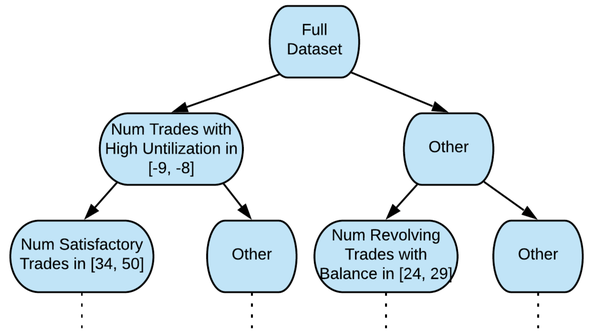

Fig. 2: Context tree splits using logistic regression.

Using the context splits above, the feature importance for those contexts are analyzed. Domain experts can determine if model explanations are reasonable. If not, the model can be modified or the predictions for the biased group can be disabled.

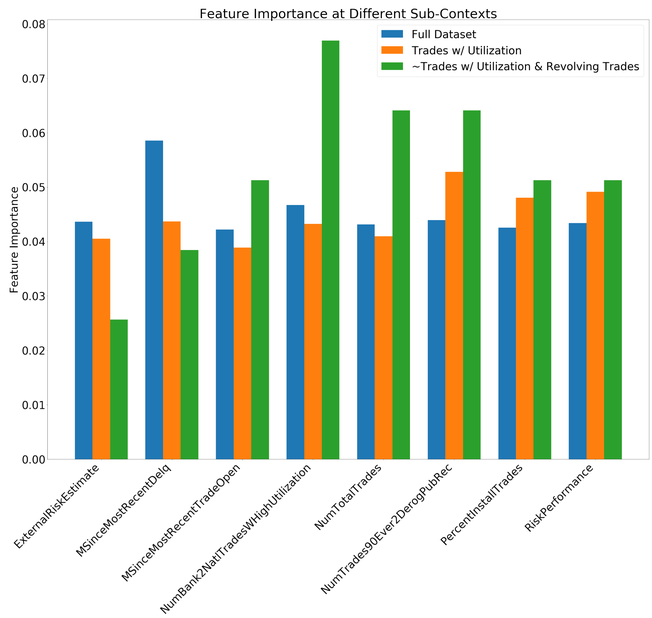

Fig. 3: Using the context splits above, the feature importance for those contexts are analyzed.

Context Size Testing

When computing the feature importance delta Δ_{FI} we also had to take into account the size of the contexts. The FI of very small contexts differ from the parent context FI_p and typically selected if only the un-boosted difference $\Delta_{UFI} = ||FI - FI_p||_1$ is used. To select larger contexts, the size of the contexts is added to the Δ_{FI} term, controlled by a context size booster λ.

$$\Delta_{FI} = ||FI - FI_p||_1 * log(|C|)^{\lambda}$$

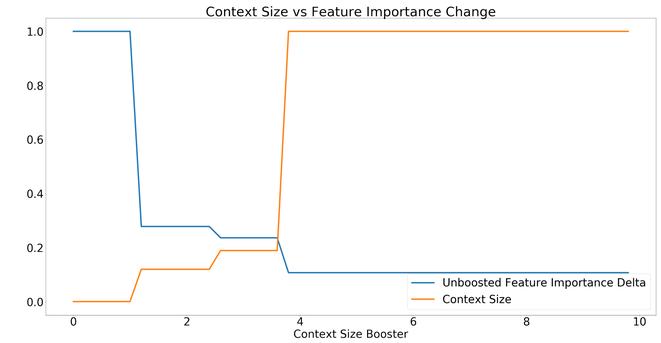

Fig. 4: We further explore the selection of λ by observing the trade-off between the true un-boosted Δ_{UFI} and the context size.

Choosing λ varies based on the user's objective:

- Smaller λ will provide more granular contexts which are useful to identify model prediction outliers for debugging.
- Larger λ will provide larger contexts and general trends in the data.

References

[1] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. "Model class reliance: Variable importance measures for any machine learning model class, from the "Rashomon" perspective". In: *arXiv preprint arXiv:1801.01489* (2018).

[2] Erik Štrumbelj and Igor Kononenko. "Explaining prediction models and individual predictions with feature contributions". In: *Knowledge and information systems* 41.3 (2014), pp. 647–665.