Causality-based Explanation of Classification Outcomes

Leopoldo Bertossi,Jordan Li, Maximilian Schleich,**Dan Suciu**, Zografoula Vagena

2020 Workshop on Data Management for End-to-End Machine Learning (DEEM)

Background

- Tackle interpretability of black-box models: We have seen LIME [1] and we also learned SHAP in the lectures
- This paper is about interpreting BBox models from the perspective of Causality

[1] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

Background : Causality

Let’s review some of the concepts we discussed in class and also used in this paper.

- Counterfactual cause : remove/change a ***feature*** would flip the prediction

A score of measuring the degree to which a feature is affecting “flipping the prediction”

$$\text{COUNTER}(\mathbf{e}^\star, F_i) \stackrel{\text{def}}{=} L(\mathbf{e}^\star) - \mathbf{E}\left[L(\mathbf{e}) | \mathbf{e}_{\mathcal{F}-\{F_i\}} = \mathbf{e}^\star_{\mathcal{F}-\{F_i\}}\right]$$

e★: an entity
L(e★): outcome prediction by a model L (assume to be 1 in this paper)
e: any entity that is not e★ that shares the same feature values except Fi

Background : Causality

Counterfactual example: Loan Approval

Name	Education l	Credit Score	Place of Origi	Approved	
Alice	Master's	High	China	1	T1
Bob	Master's	High	USA	1	T2
Carol	Master's	Low	USA	0	T3
David	Master's	Low	USA	0	T4
Eve	Master's	High	China	1	T5
Grace	Master's	Low	China	0	T6
Frank	Bachelor's	Middle	USA	0	T7

Why is Eve's loan approved by Model?
(1: approved, 0: denied)

Background : Causality

Name	Education l	Credit Score	Place of Origi	Approved	
Alice	Master's	High	China	1	T1
Bob	Master's	High	USA	1	T2
Carol	Master's	Low	USA	0	T3
David	Master's	Low	USA	0	T4
Eve	Master's	High	China	1	T5
Grace	Master's	Low	China	0	T6
Frank	Bachelor's	Middle	USA	0	T7

$$\text{COUNTER}(\mathbf{e}^\star, F_i) \stackrel{\text{def}}{=} L(\mathbf{e}^\star) - \mathbf{E}\left[L(\mathbf{e}) | \mathbf{e}_{\mathcal{F}-\{F_i\}} = \mathbf{e}^\star_{\mathcal{F}-\{F_i\}}\right]$$

Candidate exp 1: POO=china

COUNTER(t5, POO) = 1 - (1+1)/2 = 0

Background : Causality

Counterfactual example: Loan Approval

Name	Education l	Credit Score	Place of Origi	Approved	
Alice	Master's	High	China	1	T1
Bob	Master's	High	USA	1	T2
Carol	Master's	Low	USA	0	T3
David	Master's	Low	USA	0	T4
Eve	Master's	High	China	1	T5
Grace	Master's	Low	China	0	T6
Frank	Bachelor's	Middle	USA	0	T7

$$\text{COUNTER}(\mathbf{e}^\star, F_i) \stackrel{\text{def}}{=} L(\mathbf{e}^\star) - \mathbf{E}\left[L(\mathbf{e}) | \mathbf{e}_{\mathcal{F}-\{F_i\}} = \mathbf{e}^\star_{\mathcal{F}-\{F_i\}}\right]$$

Candidate exp 2: CS=High

COUNTER(t5, CS) = 1 - (1+0)/2 = 0.5

Background : Causality

- Actual Cause : *A pair (Fi, v) is called an actual cause with contingency (Γ, w), where Γ is a set of features and w is a set of values, if (Fi, v) is a counterfactual cause for e★[Γ := w]*

Alice: 30 years old, bachelors -> Denied
Alice: 35 years old, bachelors -> Denied
Alice: 35 years old, masters -> Approved

We say “masters” is actual cause with contingency set (age=35 years old), the responsibility is 1/2 since |Γ| = 1

Background : Causality

$$\text{COUNTER}(\mathbf{e}^\star, F_i) \stackrel{\text{def}}{=} L(\mathbf{e}^\star) - \mathbf{E}\left[L(\mathbf{e})|\mathbf{e}_{\mathcal{F}-\{F_i\}} = \mathbf{e}^\star_{\mathcal{F}-\{F_i\}}\right]$$

$\mathbf{e}' \stackrel{\text{def}}{=} \mathbf{e}^\star[\Gamma := \mathbf{w}]$

$\mathbf{e}'' \stackrel{\text{def}}{=} \mathbf{e}'[F_i := v]$

Extend with contingency set and actual cause

$$\text{RESP}(\mathbf{e}^\star, F_i, \Gamma, \mathbf{w}) \stackrel{\text{def}}{=} \frac{L(\mathbf{e}') - \mathbf{E}\left[L(\mathbf{e}'')|\mathbf{e}''_{\mathcal{F}-\{F_i\}} = \mathbf{e}'_{\mathcal{F}-\{F_i\}}\right]}{1 + |\Gamma|}$$

$$\varphi_i(v) = \frac{1}{\text{number of players}} \sum_{\text{coalitions excluding } i} \frac{\text{marginal contribution of } i \text{ to coalition}}{\text{number of coalitions excluding } i \text{ of this size}}$$

- **SHAP-score is based on the shapley value we discussed in the lecture**
- **A refresher:**

Fix an entity $\mathbf{e}^\star$ and a feature F_i. Let π be a permutation on the set of features $\mathcal{F}$; in other words, π fixes a total order on the set of features. Denote by $\pi^{<F_i}$ the set of features F_j that come before F_i in the order π; similarly, $\pi^{\leq F_i}$ denotes $\pi^{<F_i} \cup \{F_i\}$. The *contribution* of the feature F_i is defined as:

"Player" here is defined as feature

$$c(\mathbf{e}^\star, F_i, \pi) \stackrel{\text{def}}{=} \mathbf{E}\left[L(\mathbf{e}) | \mathbf{e}_{\pi^{\leq F_i}} = \mathbf{e}^\star_{\pi^{\leq F_i}}\right] - \mathbf{E}\left[L(\mathbf{e}) | \mathbf{e}_{\pi^{<F_i}} = \mathbf{e}^\star_{\pi^{<F_i}}\right]$$

$$\text{SHAP}(\mathbf{e}^\star, F_i) \stackrel{\text{def}}{=} \frac{1}{n!} \sum_{\pi} c(\mathbf{e}^\star, F_i, \pi)$$

Background: Connection between SHAP and COUNTER

S ⊆ F − {Fi}, and define the contribution of Fi w.r.t. S as

$$c'(\mathbf{e}^\star, F_i, S) \stackrel{\text{def}}{=} \mathbf{E}\left[L(\mathbf{e})|\mathbf{e}_{S\cup\{F_i\}} = \mathbf{e}^\star_{S\cup\{F_i\}}\right] - \mathbf{E}\left[L(\mathbf{e})|\mathbf{e}_S = \mathbf{e}^\star_S\right]$$

For a number 0 ≤ ℓ< n, We define the SHAP-score at level ℓ as

$$\text{SHAP}(\mathbf{e}^\star, F_i, \ell) \stackrel{\text{def}}{=} \frac{\ell!(n-\ell-1)!}{n!} \sum_{S\in\binom{\mathcal{F}-\{F_i\}}{\ell}} c'(\mathbf{e}^\star, F_i, S) \qquad (2)$$

$$\text{SHAP}(\mathbf{e}^\star, F_i) = \textstyle\sum_{\ell=0,n-1} \text{SHAP}(\mathbf{e}^\star, F_i, \ell)$$

$$\text{COUNTER}(\mathbf{e}^\star, F_i) \stackrel{\text{def}}{=} L(\mathbf{e}^\star) - \mathbf{E}\left[L(\mathbf{e})|\mathbf{e}_{\mathcal{F}-\{F_i\}} = \mathbf{e}^\star_{\mathcal{F}-\{F_i\}}\right]$$

SHAP value at level (n-1) is actually directly related to COUNTER value

$$\text{. SHAP}(\mathbf{e}^\star, F_i, n-1) = \frac{1}{n}\text{COUNTER}(\mathbf{e}^\star, F_i).$$

Enough formulas, let's start calculating!

- However, we need to address some other issues
  - We have a lot of conditional expectations in both RESP and SHAP, dataset at hand is just a sample.
  - **2 Probability spaces: approximate the actual distribution of data**
    - **Product space** : assuming independence (which is often not true in reality), prob for a tuple to appear in the distribution is

$$p(\langle x_1, \ldots, x_n \rangle) \stackrel{\text{def}}{=} \prod_i p(F_i = x_i)$$

As if we have more data by replacing data itself with probabilities

Actual Implementation

- Start by calculating COUNTER score
- If all the counterfactual scores are 0, then from bottom up enumerating contingency sets until find a solution

Enough formulas, let's start calculating!

- Turns out SHAP calculation in product space is #P-Hard:
  - A reduction from SHAP-scores to probability calculation problem
- **2nd probability space:**
  - **The Empirical Distribution:**
    - lets just focus on the data we have **and only** the data we have (we don't claim any probability of existence of any unseen tuples)
    - This is more restrictive, but with one advantage: we can do early stopping in SHAP-score

$$\mathrm{SHAP}(\mathbf{e}^\star, F_i, \ell) \stackrel{\text{def}}{=} \frac{\ell!(n-\ell-1)!}{n!} \sum_{S \in \binom{\mathcal{F}-\{F_i\}}{\ell}} c'(\mathbf{e}^\star, F_i, S) \qquad (2)$$

$$c(\mathbf{e}^\star, F_i, \pi) \stackrel{\text{def}}{=} \mathbf{E}\left[L(\mathbf{e}) | \mathbf{e}_{\pi \le F_i} = \mathbf{e}^\star_{\pi \le F_i}\right] - \mathbf{E}\left[L(\mathbf{e}) | \mathbf{e}_{\pi < F_i} = \mathbf{e}^\star_{\pi < F_i}\right]$$

As S size increases, it is more and more unlikely we will have any tuples left that fulfills

However, distribution model doesn't work well when computing RESP because of lack of data. Sigh

2 Probability spaces: a summary

Product Space

- Generalize data set to “generate” more data
- Good to use in computing COUNTER and RESP
- Doesn’t work well when dealing with SHAP

Empirical Distribution

- Focus on the data and only the data
- Makes the SHAP feasible because of early stopping opportunities
- Doesn’t work well when dealing with COUNTER and RESP because of limited data

Experiment evaluation

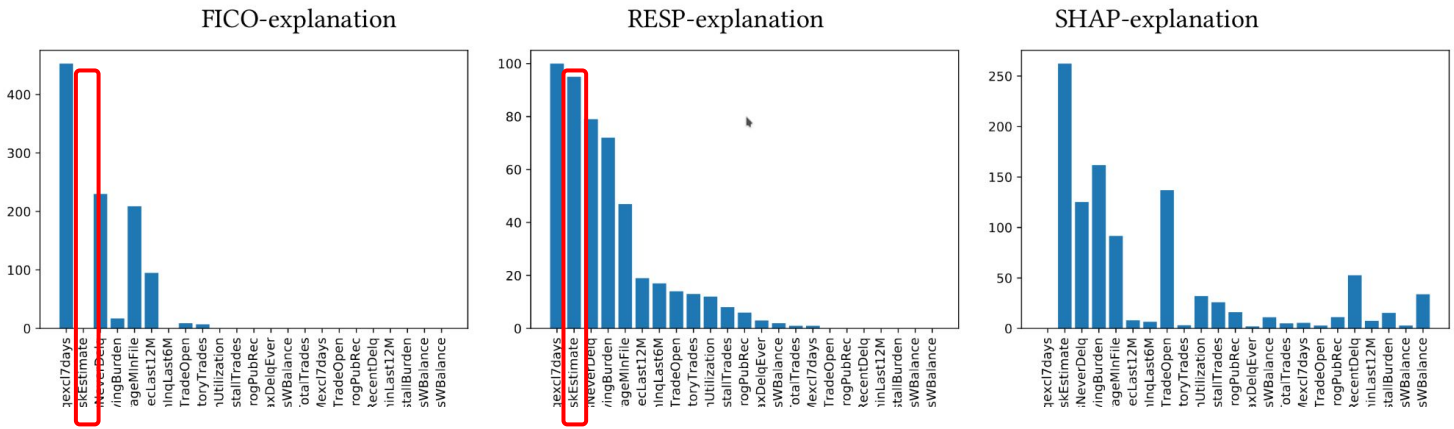
- Based on the results + tradeoffs from 2 probability spaces:
  - RESP score : calculated on product space
  - SHAP score : calculated on empirical space
- Datasets:
  - FICO data : loan grant risk
  - Kaggle Credit Card Fraud: fraud or not fraud

FICO

- Compare RESP, SHAP and FICO scores from a logistic regression based model from [1] (show demo briefly)
- FICO-score:
  - First find top M “subscales”
  - And find the top N features within each of the M subscales
  - Final sortorder is determined by sorting [subscale-score, feature-score-within-subscale]

[1] Chaofan Chen, et al. An interpretable model with globally consistent explanations for credit risk. CoRR, abs/1811.12615, 2018

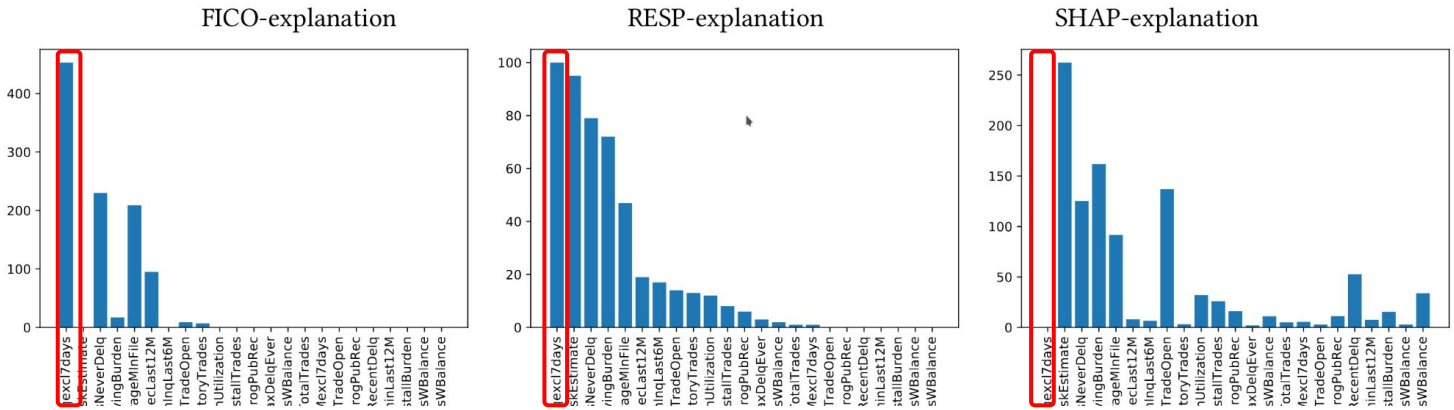
FICO



ExternalRiskEstimate

1. Due to the way FICO scores are sorted hierarchically
2. Model makes predictions on single cases and ignored the rest of the data
3. This also result in the fact that FICO-scores are less diverse (again, because of how the features are ranked)

FICO



MSinceMostRecentInqexcl7(MMR7)

Weights are head and tail heavy, MMR7 is actually evenly distributed among 2 classes in test data

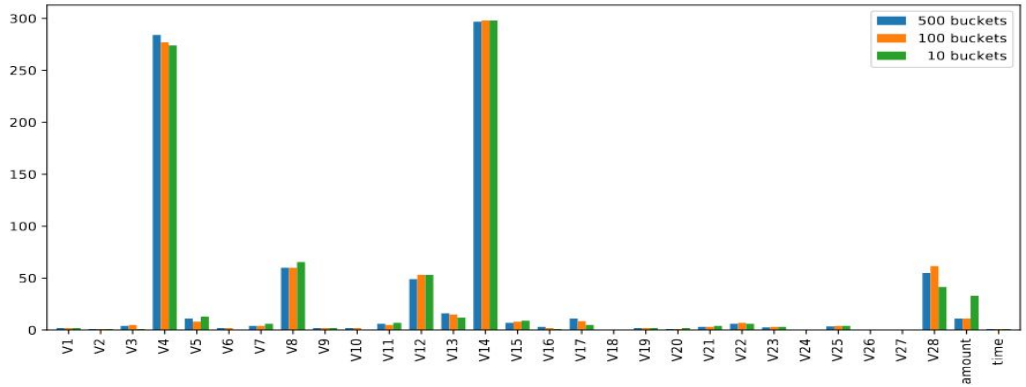
$$\mathrm{SHAP}(\mathbf{e}^{\star}, F_i, \ell) \stackrel{\text{def}}{=} \frac{\ell!(n-\ell-1)!}{n!} \sum_{S \in \binom{\mathcal{F}-\{F_i\}}{\ell}} c'(\mathbf{e}^{\star}, F_i, S) \qquad (2)$$

$$\mathrm{SHAP}(\mathbf{e}^{\star}, F_i) = \textstyle\sum_{\ell=0,n-1} \mathrm{SHAP}(\mathbf{e}^{\star}, F_i, \ell)$$

Kaggle Credit Card Fraud

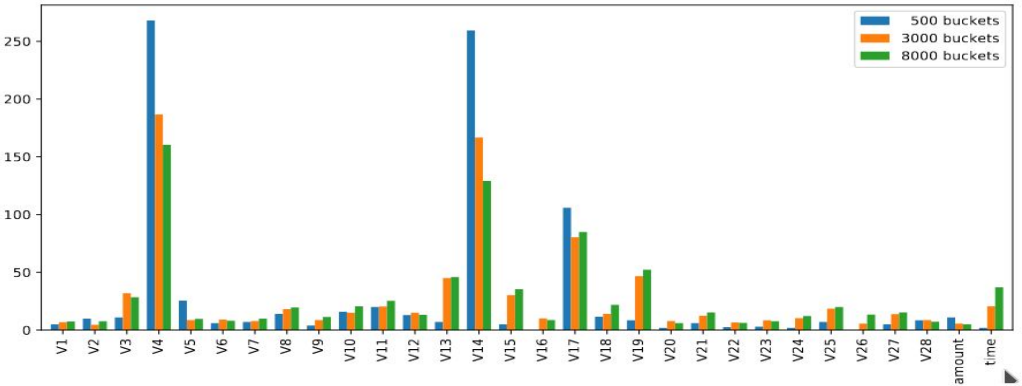
- Logistic regression as black-box-model
- Compare RESP and SHAP
- Feature values are also bucketized

Kaggle Credit Card Fraud



RESP:

More buckets- longer runtime (why?)
Insensitive to bucket size



SHAP:

More buckets- shorter runtime (why?)
sensitive to bucket size

Conclusion

- Investigated the feasibility and performance of using causality to approach explanation on black box models
- Because of the complexity of naively computing scores, some compromises have to be made : 2 probability spaces
- Experimentally show the results of RESP and SHAP and reasonably explained the reasons behind

- I am wondering practically speaking, which kind of explanation is more convincing for a real customer, thoughts?