Why Does My Model Fail? Contrastive Local Explanations for Retail Forecasting

Ana Lucic, Hinda Haned, Maarten de Rijke

University of Amsterdam

@__alucic

a.lucic@uva.nl

a-lucic

2 April 2020

Overview

1. Motivation
2. Problem Setting
3. Solution
4. Algorithmic Details
5. Evaluation
6. Conclusion

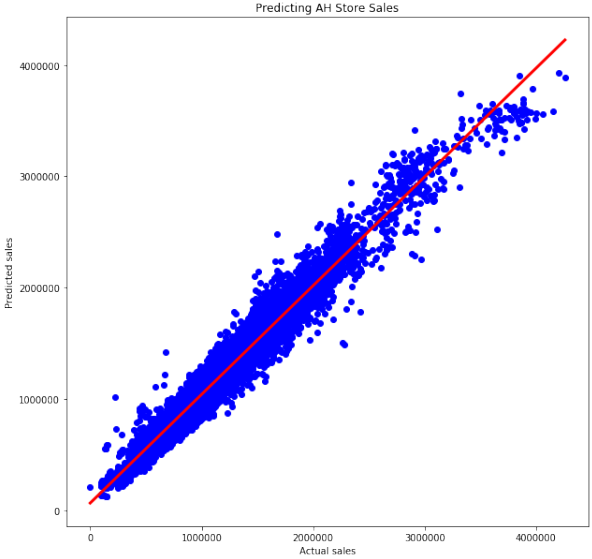
Motivation

- Explanations may be able improve trust and engagement with systems by providing insight into how decisions are made.
- GDPR requires "meaningful explanations of the logic involved" in automated decision making.
- We focus on prediction errors because users tend to be more interested in unexpected outcomes rather than ones that confirm prior beliefs [2].
- Prediction mistakes have been shown to significantly affect confidence in the model [1].
  - Users are more likely to choose a human forecaster instead, *even after seeing the algorithm outperform the human!*

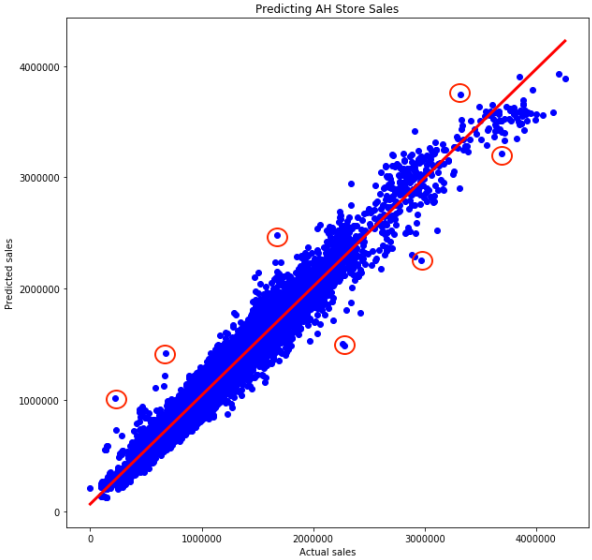
Problem Setting

- Our work was motivated by the needs of analysts at Ahold Delhaize, a large retailer in the Netherlands working in sales forecasting.
- Current production systems at the company are based on autoregressive models.
- There exists an interest in using more complex techniques, but not at the expense of interpretability.
- Large mistakes in predictions are particularly problematic in forecasting.

Problem Setting



Problem Setting



Solution: MC-BRP Explanations

- **MC-BRP:** Monte Carlo Bounds for Reasonable Predictions.

- Given an erroneous prediction, MC-BRP generates:
  1. Feature values that would result in a reasonable prediction, based on the n most important features.
  2. General trends between each feature and the target variable.

MC-BRP: Example explanation

Input	Trend	Value	Reasonable range
total_contract_hrs	As input increases, sales increase	9628	[4140,6565]
advertising_costs	As input increases, sales increase	18160	[8290,15322]
num_transactions	As input increases, sales increase	97332	[51219,75600]
total_headcount	As input increases, sales increase	226	[95,153]
floor_surface	As input increases, sales increase	2013	[972,1725]

MC-BRP: Large Errors

Definition 1: Large error

Let x be an observation in the test set X, let ϵ be the corresponding prediction error for x, and let E be the set of all errors of X. Then ϵ is a *large error* iff

$$\epsilon > Q_3(E) + 1.5(Q_3(E) - Q_1(E)),$$

where $Q_1(E), Q_3(E)$ are the first and third quartiles of the set of errors, respectively. We denote this threshold as ϵ_{large}.

- X can be decomposed into two sets:
  - R: set of observations resulting in reasonable predictions
  - L: set of observations results in large errors

MC-BRP: Algorithm

Objective: given a large error l, create a set of perturbed instances resulting in reasonable predictions, R'.

1. Identify ${\Phi_n}^{(l)}$: the top n most important features for l.
2. Randomly perturb each $\phi_j^{(l)} \in {\Phi_n}^{(l)}$ m times $\rightarrow$ mn perturbations of l.
3. Run each perturbed version, l', through original model f, obtain new prediction $f(l') = \hat{t}'$.
4. If $|\hat{t}' - t| < \epsilon_{large}$, add this l' to R'.

MC-BRP: Algorithm

Given a set of perturbed instances resulting in reasonable predictions, R', determine:

1. Bounds on feature values for reasonable predictions.
   - Based on mean and standard deviation of feature values in R'
2. Relationship between feature values and target
   - Based on Pearson correlation for observations in R'

Definition 2: Reasonable Bounds

The *reasonable bounds* for values of each feature ϕ_j in $\Phi^{(I)}$ are

$$\left[\mu(\phi_j^{(I)}) - \sigma(\phi_j^{(I)}), \mu(\phi_j^{(I)}) + \sigma(\phi_j^{(I)})\right],$$

where $\mu(\phi_j^{(I)})$ and $\sigma(\phi_j^{(I)})$ are the mean and standard deviation of each feature $\phi_j^{(I)}$, respectively, based on the set of perturbed instances resulting in reasonable predictions, R'.

Definition 2: Reasonable Bounds

The *reasonable bounds* for values of each feature ϕ_j in $\Phi^{(I)}$ are

$$\left[\mu(\phi_j^{(I)}) - \sigma(\phi_j^{(I)}), \mu(\phi_j^{(I)}) + \sigma(\phi_j^{(I)})\right],$$

where $\mu(\phi_j^{(I)})$ and $\sigma(\phi_j^{(I)})$ are the mean and standard deviation of each feature $\phi_j^{(I)}$, respectively, based on the set of perturbed instances resulting in reasonable predictions, R'.

Definition 3: Trend

The *trend*, $\rho_{\phi_j^{(I)}}$, of each feature is the Pearson coefficient between each feature $\phi_j^{(I)}$ and the predictions $\hat{t}'$ based on the observations in R'.

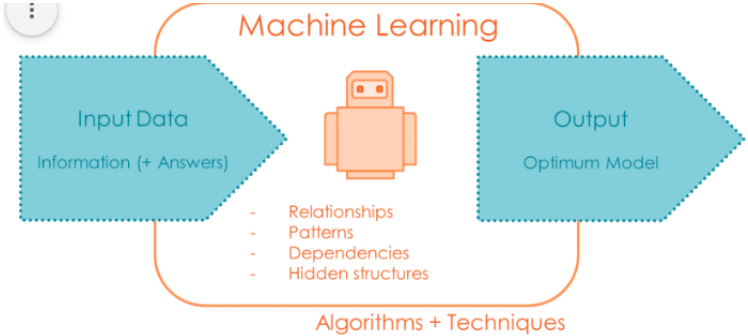
User Study

We evaluate our method through a user study with both objective and subjective components:

- Objective:
  - Forward simulations
  - Counterfactual simulations
- Subjective:
  - Split users into a treatment and control group and ask questions based on the user study by ter Hoeve et al. [3].

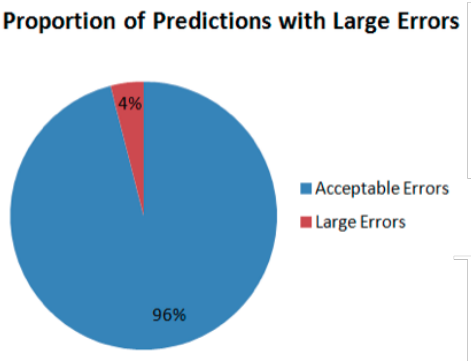
User Study

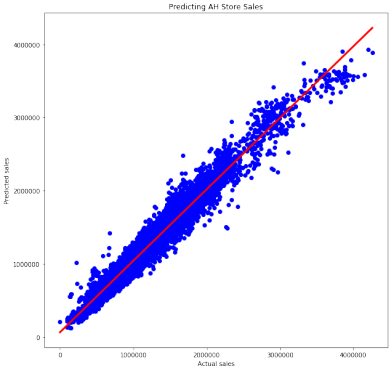
We first give the users a short primer about machine learning



User Study

Then we show a simple visual description of the model:





User Study: Objective

Forward simulation:

- *"Does this prediction result in a large error?"*

Input	**Trend**	**Value**	**Reasonable range**
total_contract_hrs	As input increases, sales increase	9628	[4140,6565]
advertising_costs	As input increases, sales increase	18160	[8290,15322]
num_transactions	As input increases, sales increase	97332	[51219,75600]
total_headcount	As input increases, sales increase	226	[95,153]
floor_surface	As input increases, sales increase	2013	[972,1725]

User Study: Objective

Counterfactual simulation:

- *"This prediction results in a large error. How can you change the input values in order to obtain a reasonable prediction?"*

Input	**Trend**	**Value**	**Reasonable range**
total_contract_hrs	As input increases, sales increase	9628	[4140,6565]
advertising_costs	As input increases, sales increase	18160	[8290,15322]
num_transactions	As input increases, sales increase	97332	[51219,75600]
total_headcount	As input increases, sales increase	226	[95,153]
floor_surface	As input increases, sales increase	2013	[972,1725]

User Study: Objective

We find that the majority of users are able to perform these simulations correctly, with an average accuracy of 81.7%.

Human accuracy	
Forward simulation 1	85.7%
Forward simulation 2	83.3%
Counterfactual simulation	76.2%
Average	81.7%

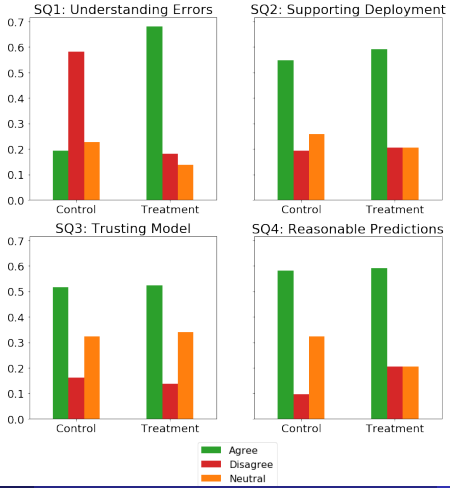
User Study: Subjective

We ask users in the treatment and control groups the following questions:

- **SQ1:** I understand why the model makes large errors in predictions.
- **SQ2:** I would support using this model as a forecasting tool.
- **SQ3:** I trust this model.
- **SQ4:** In my opinion this model produces mostly reasonable outputs.

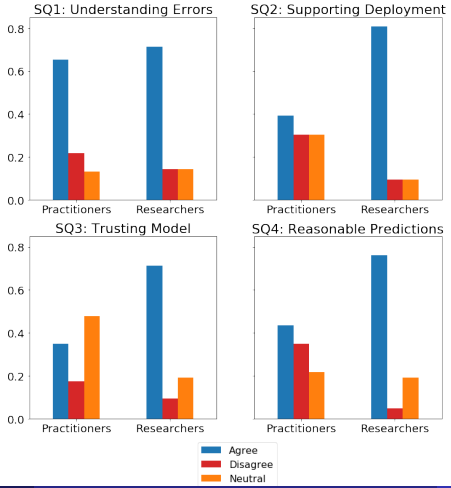
User Study Evaluation: Subjective

We find that users in the treatment group agree with **SQ1** significantly more than users in the control group. We do not find any other statistically significant differences.



User Study Evaluation: Practitioners vs. Researchers

In the treatment group, we find significant differences for **SQ2, SQ3, SQ4** when conditioning on users' backgrounds, *but not for SQ1*.



Conclusion

- We contribute MC-BRP: a method for explaining large errors in regression predictions.
- MC-BRP explanations are (i) interpretable and (ii) actionable with an average accuracy of 81.7%
- MC-BRP explanations have a significant effect on **helping users understand why the model makes mistakes**.
- MC-BRP explanations are **more beneficial to Researchers** in comparison to Practitioners in terms of **supporting deployment** of the model, **trust** in the model, and perceptions of the model's **performance**.

References

B. J. Dietvorst, J. P. Simmons, and C. Massey.
Algorithm aversion: People erroneously avoid algorithms after seeing them err.
Journal of Experimental Psychology, 144:114–126, 2015.

D. J. Hilton and B. R. Slugoski.
Knowledge-based causal attribution: The abnormal conditions focus model.
Psychological Review, 93:75–78, 1986.

M. ter Hoeve, M. Heruer, D. Odijk, A. Schuth, M. Spitters, and M. de Rijke.
Do news consumers want explanations for personalized news rankings?
In *FATREC Workshop on Responsible Recommendation*, 2017.

Thanks!

@__alucic

a.lucic@uva.nl

a-lucic

Paper: https://arxiv.org/abs/1908.00085

Code: https://github.com/a-lucic/mc-brp