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

Ana Lucic
University of Amsterdam
Amsterdam, Netherlands
a.lucic@uva.nl

Hinda Haned
Ahold Delhaize
Zaandam, Netherlands
hinda.haned@aholddelhaize.com

Maarten de Rijke
University of Amsterdam
Amsterdam, Netherlands
derijke@uva.nl

ABSTRACT

In various business settings, there is an interest in using more complex machine learning techniques for sales forecasting. It is difficult to convince analysts, along with their superiors, to adopt these techniques since the models are considered to be "black boxes," even if they perform better than current models in use. We examine the impact of contrastive explanations about large errors on users' attitudes towards a "black-box" model. We propose an algorithm, Monte Carlo Bounds for Reasonable Predictions. Given a large error, MC-BRP determines (1) feature values that would result in a reasonable prediction, and (2) general trends between each feature and the target, both based on Monte Carlo simulations. We evaluate on a real dataset with real users by conducting a user study with 75 participants to determine if explanations generated by MC-BRP help users understand why a prediction results in a large error, and if this promotes trust in an automatically-learned model. Our study shows that users are able to answer objective questions about the model's predictions with overall 81.1% accuracy when provided with these contrastive explanations. We show that users who saw MC-BRP explanations understand why the model makes large errors in predictions significantly more than users in the control group. We also conduct an in-depth analysis of the difference in attitudes between Practitioners and Researchers, and confirm that our results hold when conditioning on the users' background.

CCS CONCEPTS

• **Computing methodologies** → *Artificial intelligence*; *Machine learning*; *Supervised learning by regression*; *Ensemble methods*.

KEYWORDS

Explainability, Interpretability, Erroneous predictions

ACM Reference Format:
Ana Lucic, Hinda Haned, and Maarten de Rijke. 2020. Why Does My Model Fail? Contrastive Local Explanations for Retail Forecasting. In *Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3351095.3372824

FAT '20, January 27–30, 2020, Barcelona, Spain*

ACM ISBN 978-1-4503-6936-7/20/02…$15.00
https://doi.org/10.1145/3351095.3372824

1 INTRODUCTION

As more and more decisions about humans are made by machines, it becomes imperative to understand how these outputs are produced and what drives a model to a particular prediction [19]. As a result, algorithmic interpretability has gained significant interest and traction in the machine learning (ML) community over the past few years [3]. However, there exists considerable skepticism outside of the ML community due to a perceived lack of transparency behind algorithmic predictions, especially when errors are produced [2]. We aim to evaluate the effect of explaining model outputs, specifically large errors, on users' attitudes towards trusting and deploying complex, automatically learned models.

Further motivation for interpretable ML is provided by significant societal developments. Important examples include the recently enacted European General Data Protection Regulation (GDPR), which specifies that individuals will have the right to "the logic involved in any automatic personal data processing" [4]. In Canada and the United States, this right to an explanation is an integral part of financial regulations, which is why banks have not been able to use high-performing "black-box" models to evaluate the credit-worthiness of their customers. Instead, they have been confined to easily interpretable algorithms such as decision trees (for segmenting populations) and logistic regression (for building risk scorecards) [12]. At NeurIPS 2017, an Explainable ML Challenge was launched to combat this limitation, indicating the finance industry's interest in exploring algorithmic explanations [5].

We use explanations as a mechanism for supporting innovation and technological development while keeping the human "in the loop" by focusing on predictive modeling as a tool that aids individuals with a given task. Specifically, our interest lies with interpretability in a scenario where users with varying degrees of ML expertise are confronted with large errors in the outcome of predictive models. We focus on explaining large errors because people tend to be more curious about unexpected outcomes rather than ones that confirm their prior beliefs [10].

However, Dietvorst et al. [2] showed that when users are confronted with errors in algorithmic predictions, they are less likely to use the model. Seeing an algorithm make mistakes significantly decreases confidence in the model, and users are more likely to choose a human forecaster instead, even after seeing the algorithm outperform the human [2]. This indicates that prediction mistakes have a significant impact on users' perception of the model. By focusing on explaining mistakes, we hope to give insight into this phenomenon of algorithm aversion while also giving users the types of explanations they are interested in seeing.

Our work was motivated by the needs of analysts at Ahold Delhaize, a large Dutch retailer, working on sales forecasting. Current

models in production are based on simple autoregressive methods, but there is an interest in exploring more complex techniques. However, the added complexity comes at the expense of interpretability, which is problematic for Ahold Delhaize, especially when a complex model produces a forecast that is very different from the actual target value. This leads us to focus on explaining errors in regression predictions in this work. However, it should be noted that our method can be extended to classification predictions by defining "distances" between classes or by simply defining all errors as large errors.

We focus on two aspects of explainability in this scenario: the *generation* of explanations of large errors and the corresponding *effectiveness* of these explanations. Prior methods for generating explanations fail at generating explanations for large errors because they produce similar explanations for predictions resulting in large errors and those resulting in reasonable predictions (see Table 2 in Section 4 for an example). We propose a method for explaining large prediction errors, called *Monte Carlo Bounds for Reasonable Predictions* (MC-BRP), that shows users:

(1) The required bounds of the most important features in order to have a prediction resulting in a reasonable prediction.
(2) The relationship between each of these features and the target.

It should be noted that in our work, we focus on explaining errors *in hindsight*, that is, we examine large errors once they have occurred and are not predicting them in advance without having access to the ground truth. We are also not using these explanations to improve the model, but rather examine the effectiveness of explaining large errors via MC-BRP on users' trust in the model and attitudes towards deploying it, as well as their understanding of the explanations. We test on a wide range of users, including both Practitioners and Researchers, and analyze the differences in attitudes between these users. We also reflect on the process of conducting a user study by outlining limitations of our study and make recommendations for future work.

We address the following research questions:

RQ1: *Are the contrastive explanations generated by MC-BRP about large errors in predictions (i) interpretable, or (ii) actionable?* More specifically,

(i) Can contrastive explanations about large errors give users enough information to simulate the model's output (forward simulation)?
(ii) Can such explanations help users understand the model such that they can manipulate an observation's input values in order to change the output (counterfactual simulation)?

RQ2: *How does providing contrastive explanations generated by MC-BRP for large errors impact users' perception of the model?* Specifically, we investigate the following:

(i) Does being provided with contrastive explanations generated by MC-BRP impact users' understanding of why the model produces errors?
(ii) Does it impact their willingness to deploy the model?
(iii) Does it impact their level of trust in the model?
(iv) Does it impact their confidence in the model's performance?

Consequently, we make the following contributions:

- We contribute a method, MC-BRP, for generating contrastive explanations specifically for large errors in regression tasks.
- We evaluate our explanations through a user study with 75 participants in both objective and subjective terms.
- We conduct an analysis on the differences in attitudes between Practitioners and Researchers.

In Section 2 we discuss related work and identify how our problem relates to the current literature. In Section 3 we formally describe the methodology of explanations based on MC-BRP and in Section 4 we motivate our choice of dataset and describe the user study setup. In Section 5 we detail the results of the user study; we conduct further analyses in Section 6. In Section 7 we conclude and make recommendations for future work.

2 RELATED WORK

Guidotti et al. [6] compile a survey of current methods in interpretable machine learning and develop a taxonomy for classifying methods using four criteria:

- **Problem:**
  (i) *Model explanations:* interpret black-box model as a whole (globally)
  (ii) *Outcome explanations:* interpret individual black-box predictions (locally)
  (iii) *Inspection:* interpret model behavior through visual representations (globally or locally)
  (iv) *Transparent design:* model is inherently interpretable (globally or locally)
- **Model:** neural networks, tree ensembles, SVMs, model-agnostic
- **Explanator:** decision trees/rules, feature importances, salient masks, sensitivity analysis, partial dependence plots, prototype selection, neuron activation
- **Data:** tabular, image or text

Based on this schema, our setting is an *outcome explanation* problem for *tree ensembles*. We use *sensitivity analysis*, specifically Monte Carlo simulations, on *tabular* data to generate our explanations.

Existing work on generating outcome explanations specifically for tree ensembles involves finding counterfactual examples [25], identifying influential training samples [22], or identifying important features [14]. Importantly, none of these publications are specifically about (i) explaining errors, or (ii) explaining regressions. On the contrary, these publications are all based on binary classification tasks and the explanations do not necessarily provide insight into prediction mistakes.

Tolomei et al. [25] propose a method for generating counterfactual examples by identifying decision paths of interest that would result in a different prediction, then traversing down each of these paths and perturbing the instance x such that it satisfies the path in question. If this perturbation, x', (i) satisfies the decision path, and (ii) changes the prediction in the overall ensemble, then it is a candidate transformation of x. After computing all possible candidate transformations by traversing over all paths of interest (i.e., those leading to a different prediction), the candidate transformation with the smallest distance from x is selected as the counterfactual example. The explanation, then, is the difference between x and x'. Although Tolomei et al. [25]'s method also produces contrastive

explanations, our method differs from theirs since we are not aiming to identify one counterfactual example, but rather a range of feature values for which the prediction would be different. Another difference is that we do not assume full access to the original model.

Sharchilev et al. [22] also generate outcome explanations for tree ensembles. Their methodology is based on finding influential training samples in order to automatically improve the model, which differs from our work since their explanations are not of a contrastive nature. These influential training samples help us understand why a certain class was predicted for a given instance, but they make no reference to the alternative class(es). It should be noted that they include a use case on identifying harmful training examples — ones that contributed to incorrect predictions — which can be seen as a way to explain errors.

Lundberg et al. [14] propose a method for determining how much each feature contributes to a prediction and present a ranked list of the most important features as the explanation. The approach is based on the computationally intensive Shapley values [15], for which the authors develop a tree-specific approximation. This differs from our method since identifying the most important features is only a preliminary step in our pipeline — our work extends beyond this by including (1) feature bounds that result in reasonable predictions, and (2) the relationship between the features and the target as a tool to help users inspect what goes wrong when the prediction error is large.

Ribeiro et al. [20] also propose a method for identifying local feature importances and this is the one we use in our pipeline. Their method, LIME, is model-agnostic and is based on approximating the original model locally with a linear model. We share their objective of evaluating users' attitudes towards a model through local explanations but we further specify our task as explaining instances where there are large errors in predictions. Based on preliminary experiments, we find that LIME is insufficient for our task setting for two reasons:

(i) For regression tasks, LIME's approximation of the original model is not exact. This "added" error can be quite large given that our target is typically of order 10^6, and this convolutes our definition of a large error.
(ii) The features LIME deems most important are similar regardless of whether the prediction results in a large error or not, which does not provide any specific insight into why a large error occurs. These experiments are detailed in Section 4.

Other work on contrastive explanations includes identifying features that should be present or absent in order to justify a classification [1, 7] or model-agnostic counterfactuals [21, 27]. These all differ from our method since they are not specifically about explaining errors. Furthermore, the work by Dhurandhar et al. [1] and Hendricks et al. [7] is based on the binary presence/absence of input features, whereas our method perturbs inputs instead of removing them altogether.

Our work can also be viewed as a form of outlier detection. However, it differs from the standard literature outlined by Pimentel et al. [18] with respect to the objective: we are not necessarily trying to identify outliers in terms of the training data but rather explain instances in the test set whose errors are so large that they are considered to be anomalies.

Miller et al. [17] perform a survey of the papers cited in the "Related Works" section of the call for the IJCAI 2017 Explainable AI workshop [11] and find that the majority do not base their methods on the available research about explanations from other disciplines such as philosophy, psychology or cognitive sciences, or evaluate on real users. In contrast, our method is rooted in the corresponding philosophical literature [8, 10, 13] and our evaluation is based on a user study.

3 METHOD

The intuition behind MC-BRP is based on identifying the unusual properties of a particular observation. We make the assumption that large errors occur due to unusual feature values in the test set that were not common in the training set.

Given an observation that results in a large error, MC-BRP generates a set of bounds for each feature that would result in a reasonable prediction as opposed to a large error. We also include the trend as part of the explanation in order to help users understand the relationship between each feature and the target, and how the input should be changed in order to change the output.

As pointed out previously, we consider our task of identifying and explaining large errors somewhat similar to that of an outlier detection problem. A standard definition of a statistical outlier is an instance that falls outside of a threshold based on the interquartile range. A widely used version of this, called Tukey's fences, is defined as follows [26]:

$$[Q_1 - 1.5(Q_3 - Q_1), Q_3 + 1.5(Q_3 - Q_1)],$$

where Q_1 and Q_3 are the first and third quartiles, respectively.

Definition 3.1. Let x be an observation in the test set X and let t, $\hat{t}$ be the actual and predicted target values of x, respectively. 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}.

We can view X in Definition 3.1 as a disjoint union of two sets:

(i) R: the set of observations resulting in reasonable predictions, and
(ii) L: the set of observations resulting in large errors.

We determine the n most important features based on LIME $\Phi^{(x)} = \{\phi_j^{(x)}\}_{j=1}^n$, for all $x \in X$. It should be noted there exist alternative methods for determining the most important features for a particular prediction [15], which would also be appropriate.

Given $x \in X$, for each $\phi_j^{(x)} \in \Phi^{(x)}$, we determine two sets of characteristics through Monte Carlo simulations:

(i) $[a_{\phi_j^{(x)}}, b_{\phi_j^{(x)}}]$: the bounds for values of $\phi_j^{(x)}$ such that $x \in R$, $x \notin L$.
(ii) $\rho_{\phi_j^{(x)}}$: the relationship between $\phi_j^{(x)}$ and the target we are trying to predict, t.

We perturb the feature values for $l \in L$ using Monte Carlo simulations in order to determine what feature values are required

Table 1: An example of an explanation generated by MC-BRP. Here, each of the input values is outside of the range required for a reasonable prediction, which explains why this particular prediction results in a large error.

Input	Definition	Trend	Value	Reasonable range
A	total_contract_hrs	As input increases, sales increase	9628.00	[4140,6565]
B	advertising_costs	As input increases, sales increase	18160.67	[8290,15322]
C	num_transactions	As input increases, sales increase	97332.00	[51219,75600]
D	total_headcount	As input increases, sales increase	226.00	[95,153]
E	floor_surface	As input increases, sales increase	2013.60	[972,1725]

to produce a reasonable prediction. The algorithm for determining R', the set of Monte Carlo simulations resulting in reasonable predictions, is detailed in Algorithm 1.

In line 3, given $l \in L$, we determine Tukey's fences for each feature in $\Phi^{(l)}$ based on the feature values from R. This gives us the bounds from which we sample for our feature perturbations.

In line 5, we randomly sample from these bounds for each $\phi_j^{(l)} \in \Phi^{(l)}$ m-times to generate mn versions of our original observation, l. We call the i-th perturbed version l'_i, where $i \in \{1, \ldots, mn\}$.

In lines 7 and 8, we test the original model f on each l'_i, obtain a new prediction, $\hat{t'_i}$, and construct R', the set of perturbations resulting in reasonable predictions.

Once R' is generated, we compute the mean, standard deviation and Pearson coefficient [23] of the top n features of $l \in L$, $\Phi^{(l)}$, based on this set.

Algorithm 1 Monte Carlo simulation: creates a set of perturbed instances resulting in reasonable predictions R' for each large error $l \in L$

Require: instance l
Require: set of l's most important features $\Phi^{(l)}$
Require: 'black-box' model f
Require: large error threshold ϵ_{large}
Require: number of MC perturbations per feature m

1: $R' = \emptyset$
2: **for all** $\phi_j^{(l)}$ in $\Phi^{(l)}$ **do**
3: $TF(\phi_j^{(l)}) \leftarrow$ Tukey's fences for $\phi_j^{(l)}$ ▷ Based on R
4: **for** i in range $(0, m)$ **do**
5: $\phi_j'^{(l)} \leftarrow randomsample(TF(\phi_j^{(l)}))$
6: $l'_i \leftarrow l_i.replace(\phi_j^{(l)}, \phi_j'^{(l)})$
7: $\hat{t'_i} \leftarrow f(l'_i)$ ▷ New prediction
8: **if** $|\hat{t'_i} - t_i| < \epsilon_{large}$ **then**
9: $R' \leftarrow R' \cup l'_i$
return R'

Definition 3.2. The *trend*, $\rho_{\phi_j^{(x)}}$, of each feature is the Pearson coefficient between each feature $\phi_j^{(x)}$ and the predictions $\hat{t'_i}$ based on the observations in R'. It is a measure of linear correlation between two variables [23].

The set of bounds for each feature in $\Phi^{(x)}$ such that $\hat{t}$ results in a reasonable prediction are based on the mean and standard deviation of each $\phi_j^{(x)} \in \Phi^{(x)}$.

Definition 3.3. The *reasonable bounds* for values of each feature ϕ_j in $\Phi^{(x)}$, $[a_{\phi_j^{(x)}}, b_{\phi_j^{(x)}}]$, are

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

where $\mu(\phi_j^{(x)})$ and $\sigma(\phi_j^{(x)})$ are the mean and standard deviation of each feature, respectively, based on R'.

We compute the trend and the reasonable bounds for each of the n most important features and present them to the user in a table. Table 1 shows an example of an explanation generated by MC-BRP; the dataset used for this example is detailed in Section 4.1.

4 EXPERIMENTAL SETUP

Current explanation methods mostly serve individuals with ML expertise [6], but they should be extended to cater to users outside of the ML community [16]. Unlike previous work, our method, MC-BRP, generates contrastive explanations by framing the explanation around the prediction error, and aims to help users understand (i) what contributed to the large error, and (ii) what would need to change in order to produce a reasonable prediction. Presenting explanations in a contrastive manner helps frame the problem and narrows the user's focus regarding the possible outcomes [8, 13].

Our explanations are contrastive because they display to the user what would have needed to change in the input order to obtain an alternative outcome from the model — in other words, why this prediction results in a large error as opposed to a reasonable prediction.

4.1 Dataset and model

Our task is predicting monthly sales of Ahold Delhaize's stores with 45 features including financial, workforce and physical store aspects. Since not all of our Practitioners have experience with ML, using an internal dataset with familiar features allows them to leverage some of their domain expertise. The dataset includes 45,628 observations from 563 stores, collected at four-week intervals spanning from 2010–2015. We split the data by year (training: 2010–2013, test: 2014–2015) to simulate a production environment, and we treat every unique combination of store, interval and year as an independent observation. After preprocessing, we have 21,415 and 12,239 observations in our training and test sets, respectively. We train the gradient boosting regressor from scikit-learn[1] with the default settings and obtain an R^2 of 0.96.

[1] https://scikit-learn.org/

Table 2: The top $n = 5$ features according to LIME for observations resulting in large errors vs. reasonable predictions.

Large errors		Reasonable Predictions	
advertising_costs	0.188	advertising_costs	0.187
total_contract_hrs	0.175	total_contract_hrs	0.179
num_transactions	0.151	num_transactions	0.156
floor_surface	0.124	total_headcount	0.134
total_headcount	0.123	floor_surface	0.122
month	0.109	month	0.094
mean_tenure	0.046	mean_tenure	0.046
earnings_index	0.033	earnings_index	0.031

We verify our assumption that large errors are a result of unusual features values by generating MC-BRP explanations for all instances in our test set using $n = 5$ features and $m = 10{,}000$ Monte Carlo simulations. In our dataset, we find that 48% of instances resulting in large errors have feature values outside the reasonable range for all of the $n = 5$ most important features, compared to only 24% of instances resulting in reasonable predictions. Although this is not perfect, it is clear that MC-BRP produces explanations that are at least somewhat able to distinguish between these two types of predictions.

4.2 Why existing solutions are insufficient

Hilton [9] states that explanations are selective — it is not necessary or even useful to state all the possible causes that contributed to an outcome. The significant part of an explanation is what distinguishes it from the alternative outcome. If LIME explanations were suitable for our problem, then we would expect to see different features deemed important for instances resulting in large errors compared to those resulting in acceptable errors. This would help the user understand why a particular prediction resulted in a large error.

However, when generating LIME explanations for our test set using $n = 5$ features, we do not see much of a distinction in the most important features between predictions that result in large errors and those that do not. For example, advertising_costs is one of the top 5 most important features in 18.8% of instances with large errors and 18.7% of instances with reasonable predictions. These results are summarized in Table 2.

Furthermore, we originally tried to design our control group user study using explanations from LIME, but found that test users from Ahold Delhaize could not make sense of the objective questions about prediction errors because LIME does not provide any insight about errors specifically. Given that we could not even ask questions about errors using LIME explanations to users without confusing them, it is clear that LIME is inappropriate for our task.

4.3 User study design

We test our method on a real dataset with real users, both from Ahold Delhaize. We include a short tutorial about predictive modeling along with some questions to check users' understanding as a preliminary component of the study. This is because our users are a diverse set of individuals with a wide range of capabilities, including data scientists, human resource strategists, and senior

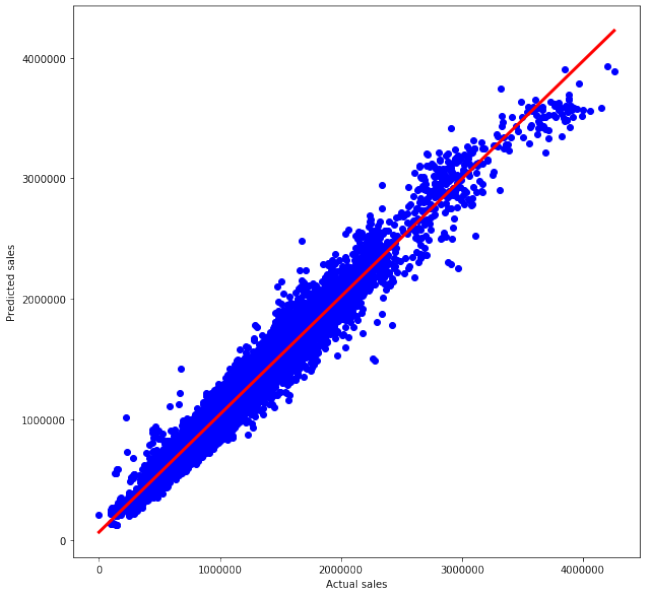

Figure 1: The visual description of the model shown to the users: a graph comparing the predicted sales and actual sales based on the original model. The red line depicts perfect predictions.

members of the executive team. We also include participants from the University of Amsterdam to simulate users who could one day work in this environment. In total, we have 75 participants: 44 in the treatment group and 31 in the control group.

All users are first provided with a visual description of the model: a simple scatter plot comparing the predicted and actual sales (as shown in Figure 1). We also show a pie chart depicting the proportion of predictions that result in large errors to give users a sense of how frequently these mistakes occur. In our case, this is 4%. Since our users are diverse, we want to make our description of the model as accessible as possible while allowing them to form their own opinions about how well the model performs. Participants in the treatment group are shown MC-BRP explanations, while those in the control group are not given any explanation.

The study contains two components, objective and subjective, corresponding to **RQ1** and **RQ2**, respectively. The objective component is meant to quantitatively evaluate whether or not users understand explanations generated by MC-BRP, while the subjective component assesses the effect of seeing the explanation on users' attitudes towards, and perceptions of, the model.

We base the objective component on *human-grounded metrics*, a framework proposed by Doshi-Velez and Kim [3], where the tasks conducted by users are simplified versions of the original task. We modify the original sales prediction task into a binary classification one: we ask users to determine whether or not a prediction will result in a large error, as it seems unreasonable to expect humans to correctly predict retail sales values of order 10^6.

To answer **RQ1**, we ask users in the treatment group to perform two types of simulations, both suggested by Doshi-Velez and Kim [3] and summarized in Table 3. The first is *forward simulation*,

Table 3: Summary of simulations performed in objective portion of the user study.

Type	Provide user with	User's task
Forward	(1) Input values (2) Explanation	Simulate output
Counterfactual	(1) Input values (2) Explanation (3) Output	Manipulate input to change output

where we provide participants with the (i) input values, and (ii) explanation. We then ask them to simulate the output — whether or not this prediction will result in a large error. The second is *counterfactual simulation*, where we provide participants with the (i) input values, (ii) explanation, and (iii) output. We then ask them what they would have needed to change in the input in order to change the output. That is, we want participants to determine how the input features can be changed (according to the trend) in order to produce a reasonable prediction as opposed to one that results in large error. These objective questions are designed to test whether a participant understands the explanations enough to predict or manipulate the model's output. We ask every participant in the treatment group to perform two forward simulations and one counterfactual simulation, and we show the same examples to all users.

For the control group, we found that we could not ask the objective questions in the same way we did for the treatment group. This is because the objective component involves simulating the model based on the explanations (see Table 3), which is not possible if the explanations are not provided. In fact, we initially left the objective questions in the control group study, but preliminary testing on some users from Ahold Delhaize showed that this was confusing and unclear, similar to when we tried using LIME explanations. We were concerned this confusion would skew users' perceptions of the model and therefore convolute the results of RQ2. Instead, we show participants in the control group the (i) input values, and (ii) output — whether or not the example resulted in a large error, In this case, we ask them *if they have enough information* to determine why the example does (or does not) result in a large error. This serves as a dummy question to engage users with the task without confusing them. We cannot ask users in the control group to simulate the model since they do not see the explanations, but we want to mimic the conditions of the treatment group as closely as possible. Therefore, **RQ1**, is solely evaluated on users from the treatment group.

To answer **RQ2**, we contrast results from the treatment and control groups. We ask both groups of users the same four subjective questions twice, once towards the beginning of the study and once again at the end. We ask the questions at the beginning of the study to evaluate the distribution of preliminary attitudes towards the model, based solely on the visual description. We ask the questions at the end of the study to evaluate the effectiveness of MC-BRP explanations, by comparing the results from the treatment and control groups. The questions we devised are based on the user study by ter Hoeve et al. [24]. Table 4 summarizes the experimental setup for the treatment and control groups. Again, the treatment and control groups are treated exactly the same with the exception of the objective questions – we only ask these to the treatment group since we cannot ask users to simulate the model without giving them the explanation.

Table 4: Summary of tasks performed in user study for the treatment and control groups. The subjective questions are asked twice.

Treatment	Control
Short modeling tutorial	Short modeling tutorial
Visual model description	Visual model description
Subjective questions	Subjective questions
Objective questions	Dummy questions
Subjective questions	Subjective questions

5 EXPERIMENTAL RESULTS

In this section, we evaluate the explanations generated by MC-BRP in terms of (i) objective questions, and (ii) subjective questions.

5.1 Objective questions

The results for users' objective comprehension of MC-BRP explanations are summarized in Table 5. We see that explanations generated by MC-BRP are both: (i) interpretable and (ii) actionable, with an average accuracy of 81.1%. This answers **RQ1**. When asked to perform forward simulations, the proportion of correct answers was 84.1% for both questions. This indicates that the majority of users were able to interpret the explanations in order to simulate the model's output (**RQ1:** interpretable). When asked to perform counterfactual simulations, the proportion of correct answers was slightly lower at 75.0%, but still indicates that the majority of users were able to determine how to manipulate the model's input in order to change the output (**RQ1:** actionable).

Table 5: Results from the objective questions in the user study.

Human accuracy	
Forward simulation 1	84.1%
Forward simulation 2	84.1%
Counterfactual simulation	75.0%
Average	81.1%

5.2 Subjective questions

In order to understand the impact of MC-BRP explanations on users' attitudes towards the model, we ask them the following subjective 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.

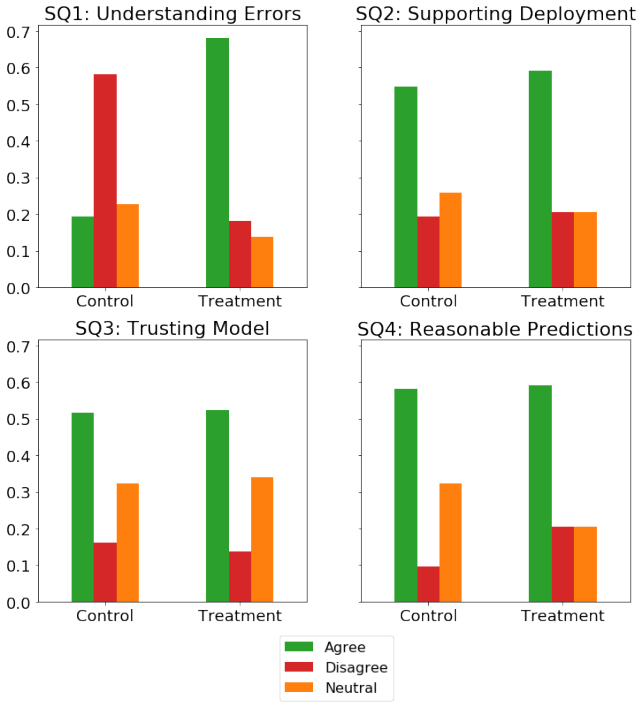

Figure 2: Results from a within-subject study comparing answers between the Treatment (MC-BRP explanation) and Control (no explanation) groups.

To ensure our populations did not have different initial attitudes towards the model, we compared their answers on the subjective questions after only showing a visual description of the model. The visual description is a graph comparing the predicted sales to the actual sales, which allows users to see the distribution of errors made by the model (see Figure 1). We found no statistically significant difference (χ^2 test, $\alpha = 0.05$) in initial attitudes towards the model, which allows us to postulate that any difference discovered between the two groups is a result of the treatment they were given (i.e., MC-BRP explanation vs. no explanation).

Figure 2 shows the distributions of answers to the four subjective questions in the treatment and control groups. The difference in distributions is significant for SQ1 ($\chi^2 = 18.2$, $\alpha = 0.0001$): users in the treatment group agree with the statement more than users in the control group. However, we find no statistically significant difference between the two groups for the remaining questions (χ^2 test, $\alpha = 0.05$). That is, MC-BRP explanations help users understand why the model makes large errors in predictions, but do not have an impact on users' trust or confidence in the model, or on their willingness to support its deployment.

6 DISCUSSION

Since our original motivation was to provide an explanation system that can be used by analysts at Ahold Delhaize, we conducted a more in-depth analysis of the results to determine if there was a difference in attitudes between users depending on their background (e.g., Practitioners from Ahold Delhaize or Researchers from the University of Amsterdam).

6.1 Comparing attitudes conditioned on background

Table 6 shows the distribution of Practitioners and Researchers in the treatment and control groups. Since we have a slight imbalance in background between the treatment and control groups, we test whether or not our results still hold when conditioning on background and confirm that they do.

Again, we do not find statistically significant differences in initial attitudes towards the model (χ^2 test, $\alpha = 0.05$). For Researchers, the distribution of answers between treatment and control groups is significantly different for SQ1 ($\chi^2 = 14.2$, $\alpha = 0.001$), but does not differ for SQ2–SQ4 (χ^2 test, $\alpha = 0.05$). The same holds for Practitioners: the distributions are significantly different only for SQ1 ($\chi^2 = 6.94$, $\alpha = 0.05$). This is consistent with our results in Section 5. In both cases, users in the treatment group agree with SQ1 more than users in the control group, indicating that MC-BRP explanations help users understand why the model makes large errors in predictions, regardless of whether they are Practitioners or Researchers. Although the results are statistically significant for both groups, the results hold more strongly for Researchers compared to those for Practitioners, given the χ^2 values.

Table 6: Distribution of Practitioners and Researchers in the treatment and control groups.

Background	Practitioners	Researchers
Treatment	52%	48%
Control	58%	42%

6.2 Comparing attitudes in the treatment group

Based on the users who saw the explanations, we compare the distributions of answers between Practitioners and Researchers in Figure 3 in order to understand the needs of different types of users. We find that there is a significant difference between Practitioners and Researchers for SQ2 ($\chi^2 = 7.94$, $\alpha = 0.05$), indicating that more Resesearchers are in favor of using the model as a forecasting tool, and less are against it or have a neutral attitude, in comparison to the Practitioners. We also find a significant difference for SQ3 ($\chi^2 = 5.98$, $\alpha = 0.05$): a larger proportion of Researchers trust the model, while the majority of Practitioners have neutral feelings. The results for SQ4 are significant as well ($\chi^2 = 6.86$, $\alpha = 0.05$): although the majority of users in both groups believe the model produces reasonable predictions, a larger proportion of the Practitioners disagree with this statement in comparison to the Researchers.

We see no significant difference between groups for SQ1 (χ^2 test, $\alpha = 0.05$), which makes sense given that we showed that MC-BRP explanations have a similar effect on both Practitioners and Researchers when comparing users in the treatment and control groups in Section 6.1.

Overall, these results suggest that our user study population is fairly heterogeneous, and that users from different backgrounds have different criteria for deploying or trusting a model, and varying levels of confidence regarding the accuracy of its outcomes.

Figure 3: Results from a within-subject study comparing answers between participants who are Practitioners or Researchers (in the treatment group).

6.3 User study limitations

Like any user study, ours has some limitations. It would have been preferable to distribute users more evenly in terms of the proportion of users in the treatment and control groups, as well as the proportion of Practitioners and Researchers in each of these groups. Unfortunately, this was not possible in our case because we recruited participants in two rounds: first for the treatment group, and then afterwards for the control group. One option could be to discard some Practitioners in the control group in order to have a better balance in terms of background, but we felt it was more important to have as many users as possible, and it would not be clear how to choose which users to discard. Fortunately, we found that our results still hold when conditioning on background as mentioned in Section 6.1. In future work, we plan to recruit for both groups at the same time to avoid issues like these.

We also acknowledge that not having a baseline method to compare to is a limitation of our study. In our case, the main issue is that there simply does not exist a method that is specifically for explaining errors in regression predictions, which would make asking questions about errors (i) unfair, and (ii) confusing, as mentioned in Sections 4.2 and 4.3. However, now that MC-BRP exists, it can serve as a baseline for future work on erroneous predictions, which is another contribution of this paper.

7 CONCLUSION

We have proposed a method, Monte Carlo Bounds for Reasonable Predictions (MC-BRP), that provides users with contrastive explanations about predictions resulting in large errors based on: (i) the set of bounds for which reasonable predictions would be expected for each of the most important features. (ii) the trend between each of these features and the target.

Given a large error, MC-BRP generates a set of perturbed versions of the original instance that result in reasonable predictions. This is done by performing Monte Carlo simulations on each of the features deemed most important for the original prediction. For each of these features, we determine the bounds needed for a reasonable prediction based on the mean and standard deviation of this new set of reasonable predictions. We also determine the relationship between each feature and the target through the Pearson correlation, and present these to the user as the explanation.

We evaluate MC-BRP both objectively (**RQ1**) and subjectively (**RQ2**) by conducting a user study with 75 real users from Ahold Delhaize and the University of Amsterdam. We answer **RQ1** by conducting two types of simulations to quantify how (i) interpretable, and (ii) actionable our explanations are. Through forward simulations, we show that users are able to interpret MC-BRP explanations by simulating the model's output with an average accuracy of 84.5%. Through counterfactual simulations, we show that MC-BRP explanations are actionable with an accuracy of 76.2%.

We answer **RQ2** by conducting a between-subject experiment with subjective questions. The treatment group sees MC-BRP explanations, while the control group does not see any explanation. We find that explanations generated by MC-BRP help users understand why models make large errors in predictions (**SQ1**), but do not have a significant impact on support in deploying the model (**SQ2**), trust in the model (**SQ3**), or perceptions of the model's performance (**SQ4**). These results still hold when conditioning on users' background (Practitioners vs. Researchers).

We also conduct an analysis on the treatment group to compare results between Practitioners and Researchers. We find significant differences for **SQ2, SQ3** and **SQ4**, but do not find a significant difference in attitudes for **SQ1**.

For future work, we intend to explore allowing a predictive model to abstain from prediction when a particular instance has unusual feature values and determine the impact this has on users' trust, deployment support and perception of the model's performance. We also plan to compile a more comprehensive set of subjective questions by using multiple questions to evaluate users' impressions on the same topic.

REPRODUCIBILITY

To facilitate the reproducibility of the results reported in this work, our code for the experimental implementation of MC-BRP is available at `http://github.com/a-lucic/mc-brp`.

ACKNOWLEDGMENTS

This research was supported by Ahold Delhaize, the Association of Universities in the Netherlands (VSNU), the Innovation Center for Artificial Intelligence (ICAI), and the Netherlands Organisation for Scientific Research (NWO) under project nr. 652.001.003. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. In *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 592–603.

[2] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing them Err. *Journal of Experimental Psychology* 144 (2015), 114–126.

[3] Finale Doshi-Velez and Been Kim. 2018. Considerations for Evaluation and Generalization in Interpretable Machine Learning. *Explainable and Interpretable Models in Computer Vision and Machine Learning* (2018).

[4] EU. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union* L119 (2016), 1–88.

[5] FICO. 2017. Explainable Machine Learning Challenge. `https://community.fico.com/s/explainable-machine-learning-challenge`

[6] Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys* 51 (2018).

[7] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Grounding Visual Explanations. In *Computer Vision – ECCV 2018*. Vol. 11206. Springer International Publishing, Cham, 269–286.

[8] Dennis J. Hilton. 1990. Conversational Processes and Causal Explanation. *Psychological Bulletin* 107 (1990), 65–81.

[9] Dennis J. Hilton. 2017. Social Attribution and Explanation. *The Oxford Handbook of Causal Reasoning* (2017).

[10] Dennis J. Hilton and Ben R. Slugoski. 1986. Knowledge-based Causal Attribution: The Abnormal Conditions Focus Model. *Psychological Review* 93 (1986), 75–78.

[11] IJCAI. 2017. Explainable AI Workshop. `http://home.earthlink.net/~dwaha/research/meetings/ijcai17-xai/`

[12] Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. 2010. Consumer Credit-risk Models via Machine Learning Algorithms. *Journal of Banking & Finance* 34, 11 (2010), 2767–2787.

[13] Peter Lipton. 1990. Contrastive Explanation. *Royal Institute of Philosophy Supplement* 27, 247-266 (1990).

[14] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2019. Explainable AI for Trees: From Local Explanations to Global Understanding. *arXiv preprint arXiv:1905.04610* (2019).

[15] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 4765–4774.

[16] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267 (2019), 1–38.

[17] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *IJCAI Workshop on Explainable Artificial Intelligence* (2017).

[18] Marco A.F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. 2014. A Review of Novelty Detection. *Signal Processing* 99 (2014), 215–249.

[19] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-Agnostic Interpretability of Machine Learning. *ICML Workshop on Human Interpretability in Machine Learning* (2016).

[20] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. 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*. ACM, 1135–1144.

[21] Chris Russell. 2019. Efficient Search for Diverse Coherent Explanations. *arXiv preprint arXiv:1901.04909* (Jan. 2019).

[22] Boris Sharchilev, Yury Ustinovsky, Pavel Serdyukov, and Maarten de Rijke. 2018. Finding Influential Training Samples for Gradient Boosted Decision Trees. In *Proceedings of the 35th International Conference on Machine Learning*.

[23] Thomas Douglas Victor Swinscow. 1997. *Statistics at Square One*. BMJ Publishing Group.

[24] Maartje ter Hoeve, Mathieu Heruer, Daan Odijk, Anne Schuth, Martijn Spitters, and Maarten de Rijke. 2017. Do News Consumers Want Explanations for Personalized News Rankings?. In *FATREC Workshop on Responsible Recommendation*.

[25] Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. 2017. Interpretable Predictions of Tree-based Ensembles via Actionable Feature Tweaking. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17*. ACM, 465–474.

[26] John W. Tukey. 1977. *Exploratory Data Analysis*. Addison-Wesley.

[27] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal* (2017).