Contrastive Explanations for Large Errors in Retail Forecasting Predictions through Monte Carlo Simulations

Ana Lucic[1*], **Hinda Haned**[1,2], **Maarten de Rijke**[1]
[1]University of Amsterdam
[2]Ahold Delhaize
{a.lucic, h.haned, derijke}@uva.nl

Abstract

At Ahold Delhaize, 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 aim to explore the impact of contrastive explanations about large errors on users' attitudes towards a 'black-box' model. In this work, we make two contributions. The first is an algorithm, Monte Carlo Bounds for Reasonable Predictions (MC-BRP). 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, based on Monte Carlo simulations. The second contribution is the evaluation of MC-BRP along with its outcomes, which has both objective and subjective components. We evaluate on a real dataset with real users from Ahold Delhaize by conducting a user study 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. The study shows that users are able to answer objective questions about the model's predictions with overall 81.7% accuracy when provided with these contrastive explanations. We also 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.

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 [Ribeiro *et al.*, 2016a]. As a result, algorithmic interpretability has gained significant interest and traction in the ML community over the past few years [Doshi-Velez and Kim, 2018]. 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 [Dietvorst *et al.*, 2015]. 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 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" [EU, 2016]. 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) [Khandani *et al.*, 2010]. At NIPS 2017, an Explainable ML Challenge was launched to combat this limitation, indicating the finance industry's interest in exploring algorithmic explanations [FICO, 2017].

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 [Hilton and Slugoski, 1986]. However, it has been shown that when users are confronted with errors in algorithmic predictions, they are less likely to use the model [Dietvorst *et al.*, 2015]. 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 [Dietvorst *et al.*, 2015]. 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.

In this work, we focus on explaining regression predictions since this is what most data scientists working in forecasting

*Contact author.

at Ahold Delhaize are confronted with, and this is what motivated our research. However, it should be noted that our method can be extended to classification tasks 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, and (2) the relationship between each of these features and the target. We examine the effectiveness of explaining large errors via MC-BRP through a user study aimed at determining users' understanding of the explanations as well as their trust in the model and attitudes towards deploying it based on the explanations.

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?*

- (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?*

- (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?

2 Related Work

Guidotti *et al.* [2018] compile a survey of current methods in interpretable machine learning and develop a taxonomy for classifying various methods using four criteria:

1. **Problem:**
    - (a) *Model explanations:* interpret black-box model as a whole (globally)
    - (b) *Outcome explanations:* interpret individual black-box predictions (locally)
    - (c) *Inspection:* interpret model behavior through visual representations (globally or locally)
    - (d) *Transparent design:* model is inherently interpretable (globally or locally)
2. **Model:** neural networks, tree ensembles, SVMs, model-agnostic
3. **Explanator:** decision trees/rules, feature importances, salient masks, sensitivity analysis, partial dependence plots, prototype selection, neuron activation
4. **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.

To the best of our knowledge, the only other work in a similar domain is that of Sharchilev *et al.* [2018]. Their methodology is based on finding influential training samples in order to automatically improve gradient boosted decision tree models. This differs from our work with respect to the objective: we are not trying to improve the model, but rather help humans understand where it makes mistakes and why it does so. Koh and Liang [2017] also use influence functions to show the effect of upweighting samples or perturbing feature values on a model's parameters, but their method only applies to smooth parametric models.

Ribeiro *et al.* [2016b] develop a method for explaining the predictions of a classifier by approximating it locally with a linear model. Their method, LIME, is model-agnostic and uses feature importances to explain the outcome of a given instance. 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.

In addition to prior work, we add (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.

Our work can also be viewed as a form of outlier detection. However, it differs from the standard literature outlined by Pimentel *et al.* [2014] 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.* [2017] perform a survey of the papers cited in the 'Related Works' section of the call for the IJCAI 2017 Explainable AI workshop [IJCAI, 2017] and find that the majority do not base their methods on the available research about explanations from other disciplines, or evaluate on real users. In contrast, our method is rooted in the philosophical literature and our evaluation is based on a user study.

3 Method

The intuition behind our method of generating contrastive explanations 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 in order to make our explanations simple and more accessible to those outside the field.

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.

We consider our task of identifying and explaining large errors 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 [Tukey, 1977]. A widely used version of this, called Tukey's fences, is defined as follows [Tukey, 1977]:

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

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

Definition 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 1 as a disjoint union of two sets:

- R: set of observations resulting in reasonable predictions
- L: set of observations resulting in large errors.

We determine the n most important features based on LIME, $\{\phi_j^{(x)}\}_{j=1}^n$, for all $x \in X$, and call this set $\Phi^{(x)}$. It should be noted that alternative methods can also be used for determining the most important features for a particular prediction.

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

1. $[a_{\phi_j^{(x)}}, b_{\phi_j^{(x)}}]$: the bounds for values of $\phi_j^{(x)}$ such that $x \in R, x \notin L$.
2. $\rho_{\phi_j^{(x)}}$: 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 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.

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

Input: instance l
Input: set of l's most important features $\Phi^{(l)}$
Input: 'black-box' model f
Input: large error threshold ϵ_{large}
Input: number of MC perturbations per feature m
1: $R' = \emptyset$
2: **for all** $\phi_j^{(l)}$ in $\Phi^{(l)}$ **do**
3: $\quad TF(\phi_j^{(l)}) \leftarrow$ Tukey's fences for $\phi_j^{(l)}$ ▷ Based on R
4: $\quad$ **for** i in range (0, m) **do**
5: $\quad\quad \phi_j'^{(l)} \leftarrow randomsample(TF(\phi_j^{(l)}))$
6: $\quad\quad l_i' \leftarrow l_i.replace(\phi_j^{(l)}, \phi_j'^{(l)})$
7: $\quad\quad \hat{t}_i' \leftarrow f(l_i')$ ▷ New prediction
8: $\quad\quad$ **if** $|\hat{t}_i' - t_i| < \epsilon_{large}$ **then**
9: $\quad\quad\quad R' \leftarrow R' \cup l_i'$
return R'

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 [Swinscow, 1997] of the top n features of $l \in L$, $\Phi^{(l)}$, based on this set.

Definition 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 [Swinscow, 1997].

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. 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.

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]

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.

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 industry users 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 with the default settings and obtain an R^2 of 0.96.

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 instances.

4 Experimental Setup

Current explanation methods mostly serve individuals with ML expertise [Guidotti *et al.*, 2018], but they should be extended to cater to users outside of the ML community [Miller, 2019]. Unlike previous work, our method generates contrastive explanations by framing the explanation around the prediction error, and aims to help users understand (1) what contributed to the large error, and (2) what would be needed to produce a prediction with an acceptable error. Presenting explanations in a contrastive manner helps frame the problem and narrows the user's focus regarding the possible outcomes [Hilton, 1990; Lipton, 1990].

Why LIME is insufficient. Hilton [2017] 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.

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

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

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.

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 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 42 participants in the treatment group and 31 participants 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 (1) make our description of the model as accessible as

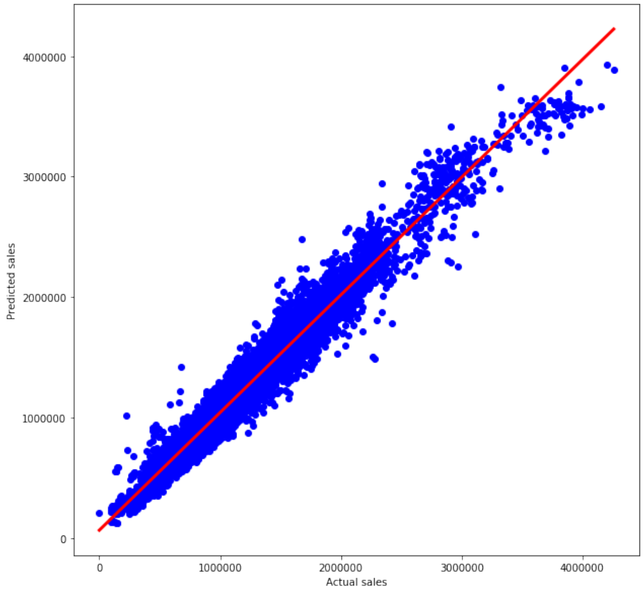

Figure 1: The visual description of the model shown to the users: a graph comparing the predicted sales and actual sales based on original model.

possible, and (2) allow them to form their own opinions about whether or not this is a good model. 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** from Section 1, 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, 2018], 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, 2018] and summarized in Table 3. The first is *forward simulation*, where we provide a participant with (1) the input features, (2) the explanation, and ask them to simulate the output. Here, the output is whether or not this prediction will result in a large error. The second is *counterfactual simulation*, where we provide participants with (1) the input features, (2) the explanation, and (3) the output, and ask them what they would have needed to change in the input in order to change the output. In other words, we asked participants to determine how the input features can be changed (according to the trend) in order to produce a reasonable prediction rather than one that results in large error. These objective questions are designed to test whether or not 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.

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 (1) the input features and (2) the output – whether or not the example resulted in a large error, and 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.

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

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

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 at the beginning of the study and once 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 are based on the user study by ter Hoeve *et al.* [2017] and are shown in Figure 2 (top).

5 Experimental Results

5.1 Objective Questions

We aim to give insight into the problem of algorithmic aversion [Dietvorst *et al.*, 2015] by explaining mistakes in a model's predictions. First, we evaluate users' comprehension of MC-BRP explanations through objective questions based on those in the treatment group. The results are summarized in Table 4. We see that explanations generated by MC-BRP are both (i) interpretable and (ii) actionable, with an average accuracy of 81.7%. This answers **RQ1**. When asked to perform forward simulations, the proportion of correct answers

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

Table 4: Results from the objective questions in the user study.

was 84.5%, averaged over the two 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 76.2%, but 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).

5.2 Subjective Questions

To ensure our populations had similar initial attitudes towards the model, we compared their answers on the subjective questions after only showing a visual description of the model (see Figure 1) and found no statistically significant difference (χ^2 test, $\alpha = 0.05$). This allows us to postulate that any difference discovered between the two groups is a result of the treatment they were given (explanation or no explanation). Figure 2 shows the subjective questions we asked along with the proportion of users that agree (or strongly agree) with these statements from both groups. Users in the treatment group agree with SQ1 significantly more than users in the control group ($\chi^2 = 16.88$, $\alpha = 0.001$). However, we find no statistically significant difference between the two groups for SQ2–SQ4 (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, unlike in [Dietvorst *et al.*, 2015].

6 Conclusion

We have proposed a method, MC-BRP, which provides users with contrastive explanations about predictions resulting in large errors based on (1) the set of bounds for which reasonable predictions would be expected for each of the most important features, and (2) the trend between each of these features and the target.

Given a large error, MC-BRP creates a set of perturbed versions of the original instance that result in acceptable errors. 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 real users from Ahold Delhaize and the University of Amsterdam. We answer **RQ1** by conducting two types of simulations to quantify

Number	Question
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.

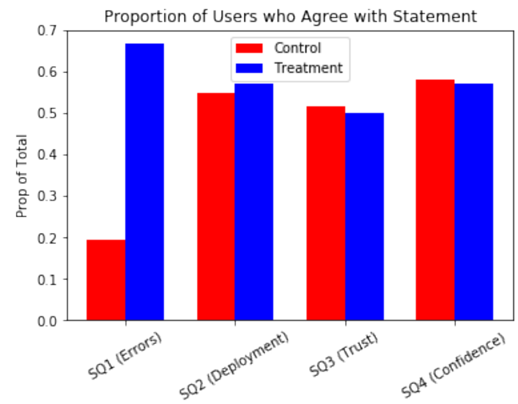

Figure 2: Results from within-subject study along with subjective question mapping.

how (i) interpretable, and (ii) actionable our explanations are. Through forward simulations, we show that users are able 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**).

For future work, we intend to explore allowing the 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.

Acknowledgements

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

[Dietvorst *et al.*, 2015] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology*, 144:114–126, 2015.

[Doshi-Velez and Kim, 2018] Finale Doshi-Velez and Been Kim. *Considerations for Evaluation and Generalization in Interpretable Machine Learning*. Springer, 2018.

[EU, 2016] EU. 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*, L119:1–88, 2016.

[FICO, 2017] FICO. Explainable machine learning challenge, 2017. https://community.fico.com/s/explainable-machine-learning-challenge.

[Guidotti *et al.*, 2018] Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51, 2018.

[Hilton and Slugoski, 1986] Dennis J. Hilton and Ben R. Slugoski. Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, 93:75–78, 1986.

[Hilton, 1990] Dennis J. Hilton. Conversational processes and causal explanation. *Psychological Bulletin*, 107:65–81, 1990.

[Hilton, 2017] Dennis J. Hilton. Social attribution and explanation. *The Oxford Handbook of Causal Reasoning*, 2017.

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

[Khandani *et al.*, 2010] Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. Consumer credit-risk models via machine learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

[Koh and Liang, 2017] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

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

[Miller *et al.*, 2017] Tim Miller, Piers Howe, and Liz Sonenberg. 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.

[Miller, 2019] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[Pimentel *et al.*, 2014] Marco A.F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

[Ribeiro *et al.*, 2016a] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *ICML Workshop on Human Interpretability in Machine Learning*, 2016.

[Ribeiro *et al.*, 2016b] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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*, pages 1135–1144. ACM, 2016.

[Sharchilev *et al.*, 2018] Boris Sharchilev, Yury Ustinovsky, Pavel Serdyukov, and Maarten de Rijke. Finding influential training samples for gradient boosted decision trees. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

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

[ter Hoeve *et al.*, 2017] Maartje ter Hoeve, Mathieu Heruer, Daan Odijk, Anne Schuth, Martijn Spitters, and Maarten de Rijke. Do news consumers want explanations for personalized news rankings? In *FATREC Workshop on Responsible Recommendation*, 2017.

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