On the Importance of Application-Grounded Experimental Design for Evaluating Explainable ML Methods

Kasun Amarasinghe[1], Kit T. Rodolfa[2], Sérgio Jesus[3], Valerie Chen[1], Vladimir Balayan[3], Pedro Saleiro[3], Pedro Bizarro[3], Ameet Talwalkar[1], Rayid Ghani[1]

[1]Carnegie Mellon University, Pittsburgh, PA
[2]Stanford University, Palo Alto, CA
[3]Feedzai, Lisboa, Portugal

amarasinghek@cmu.edu, krodolfa@law.stanford.edu, sergio.jesus@feedzai.com, valeriechen@cmu.edu, {vladimir.balayan, pedro.saleiro, pedro.bizarro}@feedzai.com, {talwalkar, rayid}@cmu.edu

Abstract

Most existing evaluations of explainable machine learning (ML) methods rely on simplifying assumptions or proxies that do not reflect real-world use cases; the handful of more robust evaluations on real-world settings have shortcomings in their design, generally leading to overestimation of methods' real-world utility. In this work, we seek to address this by conducting a study that evaluates post-hoc explainable ML methods in a setting consistent with the application context and provide a template for future evaluation studies. We modify and improve a prior study on e-commerce fraud detection by relaxing the original work's simplifying assumptions that departed from the deployment context. Our study finds no evidence for the utility of the tested explainable ML methods in the context, which is a drastically different conclusion from the earlier work. This highlights how seemingly trivial experimental design choices can yield misleading conclusions about method utility. In addition, our work carries lessons about the necessity of not only evaluating explainable ML methods using tasks, data, users, and metrics grounded in the intended application context but also developing methods tailored to specific applications, moving beyond general-purpose explainable ML methods.

Introduction

Despite the rapid expansion in explainable machine learning (ML) method development, rigorous approaches for evaluating the utility of explainable ML methods in real-world contexts have remained elusive (Doshi-Velez and Kim 2017; Chen et al. 2022; Amarasinghe et al. 2020). According to Amarasinghe et al., four key elements are required to evaluate the real-world utility of an explainable ML method: (1) the *real task* performed in the deployment context, including performance metrics that represent operational goals of the setting, (2) *real data* collected from the application setting that capture the nuances and complexities of the use case, (3) *real users* with domain expertise who perform the task in the real-world, and (4) *a robust inference strategy* to evaluate the incremental impact of explainable ML methods, including sufficient sample sizes for statistical power, appropriate hypotheses and experimental variants, and a valid analytical methodology to capture uncertainty in the data and support conclusions (Amarasinghe et al. 2020). We argue that existing evaluation studies violate at least one of these requirements, which we elaborate on in the next section.

In this work, we seek to bridge this critical gap by conducting a study that evaluates explainable ML methods in a setting *consistent with the application context*. Our study builds on the e-commerce fraud detection setting used in a previous study (Jesus et al. 2021) consisting of professional fraud analysts reviewing real-world e-commerce transactions to detect fraud. We identify several simplifying assumptions made by the previous study that deviated from the deployment context and modify the setup to relax those assumptions making the experimental setup faithful to the application context. Our setup results in *dramatically different conclusions* of the relative utility of ML predictions and explanations compared to the earlier work of (Jesus et al. 2021). For instance, while the authors of (Jesus et al. 2021) find that the ML predictions and explanations impact the decision correctness of fraud analysts (based on confusion matrix-based metrics), we do not find any utility of the ML models or explanations in improving the decision correctness metric that captures the operational objectives. In light of these results, we see our main contributions as:

1. Conducting a robust application-grounded evaluation study of explainable ML methods that includes domain expert users, the intended task, real-world data, and an inference strategy grounded in the operational goals.

2. Highlighting the critical importance of using evaluation metrics that capture operational goals, going beyond confusion matrix-based metrics and self-reported assessments of explanation quality.

3. Illustrating the importance of designing experiments that reflect the application context and providing wide-reaching lessons for future evaluations of explainable ML methods in other domains by highlighting the contrasting conclusions between the earlier work and ours.

4. Highlighting the inefficacy of popular general-purpose explainable ML methods in our real-world setting, suggesting a need for developing use-case-specific methods.

Evaluation of Explainable ML Methods

Three types of explainable ML method evaluations exist: (1) functionally-grounded, where intrinsic qualities (e.g., fidelity to the underlying model) of the explanation are evaluated through algorithmic means, (2) human-grounded, where methods are evaluated with user studies but with simplified tasks, and (3) application-grounded, where user studies involve domain experts in the intended deployment context (Doshi-Velez and Kim 2017).

Application-grounded evaluations entail logistical challenges and complexities in executing trials directly on the use case (Bhatt et al. 2020; Chen et al. 2022; Doshi-Velez and Kim 2017; Amarasinghe et al. 2020), and have remained elusive. Existing user studies mostly rely on proxies and simplifying assumptions:

1. relying on highly simplified or unrealistic tasks (e.g., forward simulation, predicting an individual's income or housing prices, answering LSAT questions, identifying the profession of a biography)) (Hase and Bansal 2020; Bansal et al. 2021; Zhang, Liao, and Bellamy 2020; Liu, Lai, and Tan 2021; Poursabzi-Sangdeh et al. 2021)
2. using readily available proxy users who lack domain expertise (e.g., workers on Amazon Mechanical Turk or users in academic settings) (Buçinca et al. 2020; Shen and Huang 2020; Bansal et al. 2021; Kim et al. 2022; Bell et al. 2022)
3. or using subjective measures of explanation quality to assess explanation utility (Islam, Eberle, and Ghafoor 2020; Yalcin, Fan, and Liu 2021; Ribeiro, Singh, and Guestrin 2016, 2018; Lundberg, Erion, and Lee 2018; Lundberg and Lee 2017; Lakkaraju, Bach, and Leskovec 2016; Singla et al. 2023).

We argue that method performance on these simplified settings do not generalize to real-world performance and violates the requirements of (Amarasinghe et al. 2020)

As mentioned, application-grounded evaluations are rare in the existing literature. While researchers are increasingly engaging with potential end users of explainable ML systems (e.g., data scientists and engineers (Kaur et al. 2020; Hong, Hullman, and Bertini 2020), child welfare workers (Kawakami et al. 2022), healthcare professionals (Tonekaboni et al. 2019; Cai et al. 2019), birders (Kim et al. 2023)), many of these evaluations are limited to qualitative investigations of user needs and fall short of evaluating the actual utility of existing methods. However, the existing handful that do perform more rigorous quantitative evaluations often suffers from shortcomings in their experimental design. We describe two examples in detail:

Lundberg et al.(Lundberg et al. 2018) evaluated the utility of an ML system with SHAP (Lundberg and Lee 2017) explanations informing anesthesiologists during surgery. Their trial consisted of five anesthesiologists tasked with detecting hypoxemia risk using historical data collected during surgeries. The study showed compelling evidence that the ML system could improve decision-making in this high-stakes context. However, the experiment only compared the performance of the physicians aided by the explainable ML system (i.e., with both ML prediction and explanation) to their performance based on their domain knowledge. By failing to compare the explainable ML system to the ML model prediction alone, the authors failed to provide evidence of the incremental utility of the explanations.

Jesus and colleagues (Jesus et al. 2021) evaluated post-hoc explanations in an e-commerce fraud detection context. Their experiment consisted of three fraud analysts tasked with detecting fraud in credit card transactions. While the experiment consisted of the variants necessary to isolate the incremental effects of explanations, there were two major simplifying assumptions that diverged their experiment from the deployment setting: (1) resampling the data to a 50/50 distribution between fraudulent and non-fraudulent transactions deviating from the deployment setting, (2) using performance metrics that did not align with the operational business goal to measure task performance, i.e., they used decision accuracy without accounting for the transaction value and the different costs of false negatives and false positives. While the authors attributed a statistically significant accuracy increase to the explainable ML methods, the simplifying assumptions make it unclear how well their results would generalize to a deployed system.

Designing an Application-Grounded Evaluation Study

We focused on designing an experiment that replicates the objectives of the application context by relaxing the simplifying assumptions typically made by evaluation studies. In particular, our work expands and improves upon a previous study (Jesus et al. 2021) that studied the utility of ML explanations in an e-commerce fraud detection setting.

Application Context: E-Commerce Fraud Detection

We partnered with the same large e-commerce merchant[1] as Jesus et al. (Jesus et al. 2021), making use of the fraud risk model (a Random Forest model) currently deployed in their production system. In this context, the ML model scores each transaction for its risk of being fraudulent and scales to a range of 0–1000. Transactions with low risk ($<$500) are automatically approved (82.1% instances), those with high risk ($>$617) are automatically declined (3.2% instances), but human analysts review those with intermediate risk scores (500–617) [2]. The data set contains 231,362 historical transactions that took place between 26 Sep 2019 and 20 Nov 2019. While the same fraud analysts who participated in the study of (Jesus et al. 2021) participated here, they were not shown any transactions they had seen before.

An analyst sees detailed information about each transaction including billing and shipping addresses, purchased items, client-side information (e.g., geolocation of the IP address, type of browser and device, etc), and the history

[1] company name withheld to discuss the details of the results

[2] A small proportion of transactions outside of this score range are also flagged for review based on client-specific heuristics, representing 6.2% of instances in our dataset.

	Jesus et al. (Jesus et al. 2021)	This Study
Label Positive Base Rate	50% (resampled)	15% (actual rate)
Evaluation Metric	confusion matrix-based metrics	Percent Dollar Regret (PDR) reflecting the domain-specific goal/metric
Interface	Accept or Reject Transactions	Accept, Reject, or Escalate Transactions
Experimental Arms	Data, Model, SHAP, TreeInterpreter, LIME	Data, Model, SHAP, TreeInterpreter, LIME, Random Explanations, Irrelevant Explanations
Sample Sizes (n)	200 (controls), 300 (explainers)	500 (all conditions)
Post-Decision Questions	Specific to explainers	Comparable across all arms
Follow-Up Interviews	Not Performed	Conducted with all analysts

Table 1: A summary of our modifications to the setup of Jesus et al. (Jesus et al. 2021) to bring the experimental setup closer to the deployment context.

of the credit card or user account associated with the purchase. Based on the information, the analyst must choose to either: *allow* the transaction to proceed, generating revenue if the transaction is legitimate but risking approving fraudulent ones; *reject* the transaction to avoid fraud, but risking blocking a legitimate transaction; or, *escalate* the transaction for further review by a more senior analyst signaling uncertainty. This additional review typically entails further research or contacting the customer directly to confirm that they are attempting to make the purchase. While this escalation is much more likely to arrive at the correct decision, it incurs a high cost in terms of time and resources.

Making the Experiment Setup Consistent with the Deployment Context

Table 1 summarizes the crucial modifications we made to improve the study by (Jesus et al. 2021). We below discuss each improvement in more detail:

Performance metrics that reflect operational goals: In reality, fraud analysts' decisions are measured on two dimensions: (1) decision correctness and (2) decision speed. One crucial simplification of the previous study is in how decision correctness was measured: they used traditional confusion matrix-based metrics to represent task performance and evaluate whether explainable ML methods meet the needs of the business. However, such metrics fail to capture the business goals by ignoring the dollar value of the transaction (i.e., assumes all transactions are equally important to the merchant) and the relative importance of the two types of errors (i.e., assumes that both types of errors have the same impact on the merchant). The concept of "correctness" needs to capture the impacts of the decisions on the business objectives, which is a more nuanced quantity than plain "accuracy".

Here, we formalize correctness through a utility metric we refer to as *Percent Dollar Regret* (PDR), which captures the revenue lost due to incorrect decisions relative to what would be realized if all the transactions were correctly classified. PDR captures the relative cost/benefit of each decision based on the monetary value of the transaction and its position in the confusion matrix. To capture these nuances, we weigh each decision/transaction by a coefficient that captures both above factors, and we apply it to the dollar value (in USD) of the transaction, v_i.

The coefficient is based on where a decision falls in the confusion matrix and its downstream impacts on the business goals:

True Negative (TN): A legitimate transaction is correctly identified and approved, generating revenue from the sale, and any future worth the customer creates. We weigh TNs with the coefficient $(1 + \lambda)$, where λ is the expected value for the long-term worth the customer would generate as a multiple of the current transaction value.

True Positive (TP): A fraudulent transaction is correctly identified and blocked. The merchant will neither lose nor gain any revenue. Hence, the weight of a TP is 0.

False Negative (FN): A fraudulent transaction is incorrectly classified as legitimate and approved. In the short term, the merchant will return the money to the credit card's real owner, lose the item, and pay the surcharge to credit card processors. There are two longer-term risks: first, perpetrators of fraud can discover the lax practices and draw more fraud attempts to the site, and second, high fraud rates can lead to penalties from credit card processors. Together, these long-term effects increase the relative costs of false negatives and introduce a risk aversion to letting fraud go undetected. We capture these combined short- and long-term costs with a weight parameter, α.

False Positive (FP): A legitimate transaction is marked as fraud and is blocked. The merchant risks losing the sale and the customer with their long-term revenue. We weigh a FP with the weight $(1-\beta)+(1-\delta)*\lambda$. In the short term, a customer can use a different payment method to complete the purchase, and β is the probability of losing the current sale; we capture the probability of longer-term customer loss with δ. We calculate the PDR metric over a set of transactions in the following form:

$$PDR = 1 - \frac{\text{Realized Revenue}}{\text{Possible Revenue}}$$

$$= \frac{\sum_i (\mathbb{1}(y_i = 0, \hat{y}_i = 1)(\beta + \delta\lambda) + \mathbb{1}(y_i = 1, \hat{y}_i = 0)\alpha) v_i}{\sum_i \mathbb{1}(y_i = 0)(1 + \lambda) v_i}$$

where $\mathbb{1}(\cdot)$ is an indicator function that takes the value 1 if the argument is satisfied and 0 otherwise, y_i is the actual label of transaction i (1 indicating fraud and 0 indicating legitimate), $\hat{y}_i$ is the label assigned by the analyst for transaction i, v_i is the value (in US Dollars) of transaction i, and each sum is taken over all transactions.

Ability to escalate "suspicious" transactions In the previous study, analysts could only decline or approve each transaction, but in the live system, they can *escalate* transactions to a more senior analyst. To better reflect the deployment context — and because this escalation rate may be an outcome of interest in its own right — we provided analysts with all three options. This impacts the metric calculation; when an analyst marks a transaction "suspicious" and elevates it to a senior analyst for a deeper dive, it incurs a time cost. Therefore, for each escalated decision, we assume that the senior analyst arrives at the correct decision but with a time penalty (τ).

Additional hypotheses and experiment variants While we evaluate the same set of post-hoc explanation methods as the previous work to enable comparison, we added two additional control/placebo arms — one where we show random explanations to the analysts (i.e., randomly picked features) and another with completely irrelevant explanations (e.g., seconds component of the transaction timestamp, last two digits of the IP address, etc.) — to evaluate whether any observed impact of explanations was merely due to the presence of an explanation.

Post-decision Questionnaire In the previous work, analysts were asked post-decision questions about the relevance and usefulness of ML explanations only after the explanation arms. To expand this analysis and provide a better baseline, our study asked a uniform set of questions in every experiment arm. The analysts were asked to indicate their confidence in their decision and the perceived quality of the available information.

Increasing sample sizes to improve statistical power We increased the power of the analysis by using 500 transactions per experimental arm (compared to 200 in control arms and 300 in explainer arms in the previous study).

Conducting Post-experiment Analyst Interviews Following the completion of data collection, we conducted qualitative interviews with each of the analysts to understand how they perceived the explanations, the balance they tried to strike between speed and correctness, and the extent to which the experimental setting may have departed from the day-to-day evaluation of live transactions.

Experimental Hypotheses and Arms

To understand the incremental impact of additional information (raw transaction details, ML model risk scores, and explanations of those scores) on analyst decisions, we designed our experiments to evaluate five hypotheses:

H1 *The model score improves analysts' decision-making performance compared to only the transaction data*. While our primary goal is to evaluate the incremental impact of post-hoc explanation methods, we posit that an important initial step is to evaluate the impact of just the ML model predictions on analyst performance.

H2 *Explanations of the model score improve analysts' decision-making performance compared to the model score and data*. This is the primary hypothesis of our experiment, and we are interested in evaluating any incremental impact of showing analysts the explanations. We hypothesize that the explanations can help the analysts focus their attention on relevant pieces of data, help them confirm or override the model predictions, and improve their performance.

H3 *Explanations make the analysts more confident in their decisions, resulting in fewer transactions being escalated to a senior analyst*. We are additionally interested in learning whether the analysts become more confident/decisive when they are presented with model explanations.

H4 *The impact of explanations on decisions is different based on which post-hoc explainer is used*. In general, different post-hoc explanation methods do not yield the same explanation for the same prediction and same model. Therefore, we hypothesize that explanations generated from distinct methods would have varying impacts on the analysts' decisions.

H5 *Explanations generated from an ad-hoc method would be worse compared to those generated by explanation methods*. We assess whether the observed effects associated with the explanations result from the mere presence of an explanation (which could influence how analysts engage with the task) or if they are attributable to the content of the explanation and the performance of the generating method.

We study the same post-hoc, feature attribution type explanation methods as the previous experiment(Jesus et al. 2021) — LIME (Ribeiro, Singh, and Guestrin 2016), TreeSHAP (Lundberg, Erion, and Lee 2018), and TreeInterpreter (Saabas 2015). As in the previous study, we present the top 6 features (i.e., the 6 features with the largest absolute importance) to the analysts. The polarity of each feature's importance is represented using green and red colors; green indicates that the feature is influencing the score toward "non-fraudulent" and red indicates an influence toward "fraudulent". Our experiment consists of seven experiment variants, as summarized in Table 2. The same three professional fraud analysts from the previous study participated in our experiments. Since we were limited to three analysts, we were unable to randomize at the user level to create the treatment and control/placebo arms. Instead, we randomized at the transaction level and organized the experiment into three stages: (1) we only showed the transaction data, (2) we introduced the ML-based fraud score, and (3) we introduced explanations and conducted all the explanation experiment arms in sequence.

Variant	Information to the analysts	Related Hypotheses
Data	Raw data and history for the transaction	H1 Control
ML Model	Data and the ML fraud score	H1 Treatment, H2 & H3 Control
TreeSHAP	Data, ML score, and TreeSHAP explanations	H2, H3, H4 Treatment
TreeInterpreter	Data, ML score, and TreeInterpreter explanations	H2, H3, H4 Treatment
LIME	Data, ML score, and LIME explanations	H2, H3, H4 Treatment
Random	Data, ML score, and explanations with random features	H5 Control
Irrelevant	Data, ML score, and irrelevant explanations	H5 Control

Table 2: Experimental variants/arms used to evaluate our hypotheses

Variant	PDR	Time(s)	Acc	FPR	TPR	Prec	Approve	Decline	Esclate
Data Only	9.5	79.2	76.6	18.7	48.6	30.7	71.7	22.4	5.9
Model	8.9	51.3	82.2	12	49.3	42	80.8	17.0	2.2
TreeSHAP	9.7	67.3	81.9	10.6	38.4	38.4	83.1	13.7	3.2
TreeInterpreter	10	65.6	80.9	12.1	40.5	37	81.7	15.5	2.8
LIME	11.6	57.7	83.2	8	38.7	43.3	85.2	12.2	2.6
Random	9	56.3	82.7	10	38.7	42	85.5	13.3	1.2
Irrelevant	9.7	62.2	81.2	9.5	29.9	36.5	85.8	12.6	1.6

Table 3: Performance summary across the experiment arms

Key Findings

We calculated PDR and the mean decision time using the following parameter values[3]: $\alpha = -3$, $\beta = 0.5$, $\delta = 0.1$, $\lambda = 3$ and $\tau = 600s$. We conducted a sensitivity analysis of our PDR metric to the parameter values and conclude that our findings are stable. We provide more details about hypothesis tests and parameter initialization in the supplementary material.

Figure 1(a) shows the PDR metric against the average decision time for all the experiment arms, while Figure 1(b) shows the same with all explainers combined into one group. We also report decision accuracy metrics for completeness and additional context, as shown in Table 3. Since our study involved only three analysts, to account for the variance across analysts, we use generalized linear models (GLMs) alongside group mean comparisons to account for the "analyst effect".

ML Models Made Analysts Faster

An important baseline for evaluating the impact of ML explanation methods is understanding the effect of the ML model score. We explore this by comparing *Data* and *ML Model* variants. Analysts became significantly faster with the *ML Model* compared to *Data*, reducing the average decision time by almost 30 seconds ($p = .0001$) (see Figure 1).

Mismatch between accuracy and PDR: We found appreciable improvements on several confusion matrix metrics: accuracy improved by 6pp ($p = .03$), FPR decreased by 6pp, and Precision increased by 9pp (See Table 3). However, we did not find any appreciable improvements in the PDR metric. The merchant is more risk-averse towards FNs than FPs, meaning that the PDR metric will be more sensitive to reductions in FNs rather than FPs, and in both variants, we observed similar FN values. This contrasting result reinforces the importance of choosing evaluation metrics that appropriately reflect operational objectives. We elaborate on this further in the supplementary material.

Analysts became more decisive: Further, the *ML model* made the analysts more confident, with a decrease in the escalated transactions by over half ($p = .003$). This reduction in the escalation rate is a significant contributor to the improvement in decision speed. The analysts also showed significantly higher levels of self-reported confidence with the *ML model* compared to *Data* (with the analysts being sure of their decisions 24% of the time compared to 12%, respectively) ($p < 10^{-5}$). Similarly, they reported a significant increase in the perceived quality of information with the *ML model*, with analysts reporting they had high-quality information (ratings of 4 or 5 out of 5) 54% of the time compared to 43% of the time with *Data* alone ($p = .002$).

Explainable ML Methods Did Not Help Analysts

We compared each *Explanation* arm and a *Combined Explainer*—an arm with all the explanation methods grouped together—to the *ML model* arm to evaluate our main hypothesis (H2). We found no significant differences in PDR, any confusion-matrix-based metrics, or escalation rates between any individual *Explanation* arm and *ML Model*. However, explainers made the analysts slower, with slowdowns of $\sim$14 seconds with *TreeInterpreter* of (($p = .03$)), $\sim$16 seconds with *TreeSHAP* ($p = .01$). As the speed of decisions directly impacts business utility, the *explanation methods worsened the outcomes of interest*. Further, we compared *Explanation* arms against each other to evaluate H4 and, again, did not observe any statistically significant differences in any metric. This suggests that all the tested explanations negatively impact analyst decision speed without improving decision-

[3]These values were based on data analysis and consulting the analysts post-experiment. Elaborated in the Supplementary materials

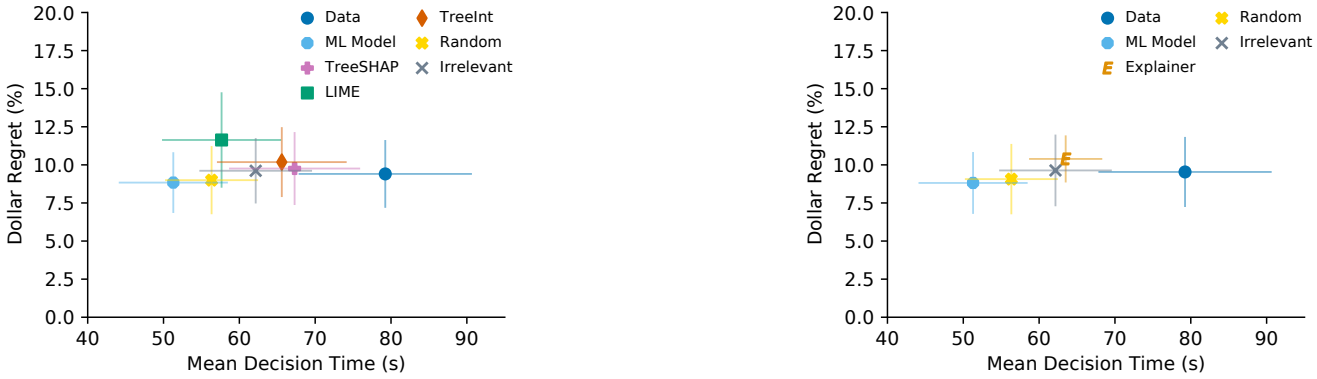

Figure 1: PDR versus Mean Decision Time for: (a) individual experiment arms, (b) explanation methods grouped together. While PDR remains similar across all variants, introducing the ML score makes the analysts faster, whereas introducing explanations slows them down without improving performance. The error bars show the 90% CIs.

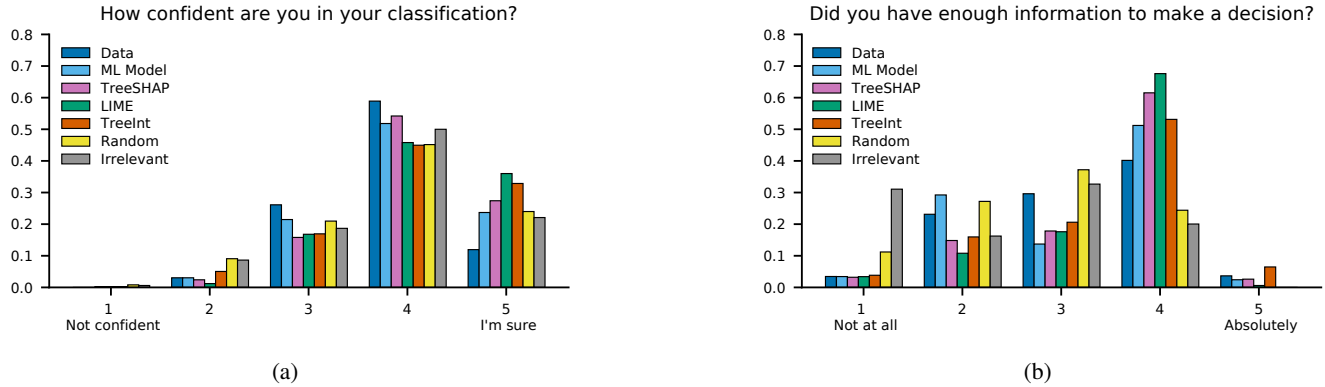

Figure 2: Analyst responses to post-decision questions for rating their confidence and perceived quality of available information. Both ratings increased as the experiment progressed with no correlation to their performance. (a) Analysts' confidence in their decision. (b) Perceived quality of the available information.

making.

Mismatch between objective and subjective metrics: Despite the absence of improvement in decision-making, the analysts reported *higher* confidence in their decisions, with over 10pp point higher rate of "I'm sure" ($(p < .001)$), and higher perceived quality in information, with 10pp increase in the instances with a rating of 4 or higher ($p < 0.06$), with explanation arms compared to the ML model. This result highlights the mismatch between self-reported metrics and outcomes of interest. It is worth noting that this comparative analysis was not possible in the previous study as it only included a questionnaire to study the explainers (Jesus et al. 2021).

Random and Irrelevant Explanations Showed Similar Outcomes to Explainable ML Methods

We ran the *Random* and *Irrelevant* experiment arms to evaluate H5. We found that both variants resulted in similar PDR and confusion matrix metrics to the *ML model* and *Explanations*. We further noticed that irrelevant explanations slowed down the analysts compared to the ML model score ($p = .08$) to a similar degree as the *real* explanation methods. Interestingly, *TreeSHAP* made the analysts slower than Random explanations ($p = .08$) *and* showed significantly higher escalation rate ($p = .03$)[4].

Analysts reported significantly lower confidence on both *Random* ($p < .002$) and *Irrelevant* explanations ($p \leq .05$) compared to the post-hoc explanation methods, with decision confidence ratings comparable to the *ML Model* arm. (Figure 2). The analysts also reported drastically lower perceived quality of information, even when compared with the *Data* and *ML Model* arms ($p = 0$), despite having access

[4]These differences could be attributed to the fact that TreeSHAP was the first variant shown to the analysts while Random and Irrelevant were the last, leaving analysts with time to learn to ignore unhelpful explanations

to the same raw data and model score. Surprisingly, these decreases in self-reported confidence and information quality did not lead to a higher escalation rate (Table 3), further emphasizing the mismatch between objective and subjective metrics of explanation quality.

Linear Models Controlling for the "Analyst Effect" Corroborated Our Results

As we were limited to three professional fraud analysts, to ensure that our findings are not a result of aberrations in analyst performance, we use GLMs to control for analyst contributions. Given that decision time was the metric where we observed significant differences across experiment variants, we focus on that metric here. We used multiple linear regression models to test if the treatment (e.g., showing the ML score and showing the explanations) impacted the decision time while controlling for the analyst and the transaction value. Even when accounting for the variance across analysts, the linear regression model confirmed our findings presented in the prior sections: Showed that the ML model score reduced the decision time significantly ($\beta = -27.9$, $p = .001$) and that explanations from TreeSHAP ($\beta = 15.85$, $p = .02$) and TreeInterpreter ($\beta = 14.24$, $p = .03$) slowed down the analysts. These findings confirm that randomizing at the transaction level with three analysts does not invalidate the drawn conclusions on the explainable ML methods' utility in the application context.

Analysts Highlighted the Mismatch between Their Needs and Explanations

The primary goal of the post-completion analyst interviews was to understand how ML explanations could be improved to better assist them in their tasks. Analysts proposed two main improvements: (1) providing more context around the feature attribution type explanations to build trust over time (e.g., how the highlighted features align with past decisions), (2) using the explanations to modify the interface to help them focus on the pertinent information and save time. This feedback further emphasizes the need for designing explainable ML methods and human-computer interfaces that address the specific requirements of use cases.

Limitations and Future Work

Although we sought to carefully design an experiment that closely reflects the true deployment setting, there are several potential limitations of the present study, providing viable avenues for future work.

Evaluating user-level hypotheses with a large user base The ideal experiment would not only randomize at the *transaction level*, but also would include a large number of users to allow for randomization at the *user level*. However, identifying such settings poses practical challenges because, typically there is limited availability of users in most real-world contexts with sufficient domain expertise who can serve as participants. It is important to note that: (1) our experiment reflects the deployment context as it used *all the analysts available for the deployment context*, and (2) randomizing at the transaction level still enabled us to test hypotheses related to explainable ML method usefulness in the context of fraud detection. Identifying real-world contexts that make larger-scale studies possible would be a promising avenue for future work to explore potential interactions between the utility of explanation methods and individual characteristics (e.g., the effect of level of experience).

Evaluating the impact of continual feedback to the analysts Due to the relatively short time the analysts spent with each arm, there was no opportunity for intermediate feedback on their performance. In our interviews, the analysts highlighted that the ongoing process of learning and improvement by receiving feedback was an important difference between the experiment and the real-world setting. This poses an interesting avenue for future work. Such an experiment might involve giving the analysts a period of initial training with each arm after which they receive a performance review, where experimental differences are measured only on subsequent decisions.

Conclusions

In this work, we illustrate the importance of application-grounded evaluation of explainable ML methods by designing and conducting a robust evaluation study that closely reflects the deployment context. Our findings highlighted several points: (1) *A need to design methods that tackle specific real-world use-cases*, rather than trying to develop "general purpose" explainable ML methods that lack grounding in the requirements of practical applications. We observed that the tested post-hoc explanation methods did not improve analysts' performance. In fact, explanations made the analysts slower compared to the ML score, reducing business utility; (2) *The critical importance of designing experiments that reflect the deployment context*. The modifications we made to the experiment setup resulted in vastly different conclusions to the ones made in the previous study applying the same post-hoc explanation methods to the same fraud detection context; (3) *The importance of going beyond confusion matrix-based metrics and choosing metrics that reflect operational objectives.* We observed that improvements in the accuracy metrics did not necessarily translate to improvements in the PDR metric that captured business utility; And, (4) *the disconnect between objective performance metrics and subjective self-reported measures of explanation utility*. Fraud analysts reported high levels of confidence when they were shown explanations compared to the ML model. However, we did not observe any significant differences in their behavior in terms of decision rates, or actual performance in terms of operational metrics. Beyond the context of fraud detection, we hope that this work 1) provides a template for experimental design and lessons for future evaluations of explainable ML methods in other contexts for explainable ML researchers, and 2) serves as a case study for practitioners exploring the use of explainable ML methods in real-world settings.

Acknowledgments

We would like to thank the Block Center for Technology and Society at Carnegie Mellon University for partially funding this work. We would also like to thank Wenbo Cui for the helpful discussions on the early iterations of this work.

References

Amarasinghe, K.; Rodolfa, K.; Lamba, H.; and Ghani, R. 2020. Explainable Machine Learning for Public Policy: Use Cases, Gaps, and Research Directions. *Data & Policy*.

Bansal, G.; Wu, T.; Zhou, J.; Fok, R.; Nushi, B.; Kamar, E.; Ribeiro, M. T.; and Weld, D. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.

Bell, A. B.; Solano-Kamaiko, I.; Nov, O.; and Stoyanovich, J. 2022. It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Inc.

Bhatt, U.; Xiang, A.; Sharma, S.; Weller, A.; Taly, A.; Jia, Y.; Ghosh, J.; Puri, R.; Moura, J. M.; and Eckersley, P. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 648–657.

Buçinca, Z.; Lin, P.; Gajos, K. Z.; and Glassman, E. L. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, 454–464. New York, NY, USA: Association for Computing Machinery. ISBN 9781450371186.

Cai, C. J.; Winter, S.; Steiner, D.; Wilcox, L.; and Terry, M. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Chen, V.; Li, J.; Kim, J. S.; Plumb, G.; and Talwalkar, A. 2022. Interpretable Machine Learning: Moving from Mythos to Diagnostics. *Queue*, 19(6): 28–56.

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Hase, P.; and Bansal, M. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? In *ACL*.

Hong, S. R.; Hullman, J.; and Bertini, E. 2020. Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1): 1–26.

Islam, S. R.; Eberle, W.; and Ghafoor, S. K. 2020. Towards quantification of explainability in explainable artificial intelligence methods. In *The Thirty-Third International Flairs Conference*.

Jesus, S.; Belém, C.; Balayan, V.; Bento, J. a.; Saleiro, P.; Bizarro, P.; and Gama, J. a. 2021. How Can I Choose an Explainer? An Application-Grounded Evaluation of Post-Hoc Explanations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 805–815. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Kaur, H.; Nori, H.; Jenkins, S.; Caruana, R.; Wallach, H.; and Wortman Vaughan, J. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 1–14. New York, NY, USA: Association for Computing Machinery. ISBN 9781450367080.

Kawakami, A.; Sivaraman, V.; Stapleton, L.; Cheng, H.-F.; Perer, A.; Wu, Z. S.; Zhu, H.; and Holstein, K. 2022. "Why Do I Care What's Similar?" Probing Challenges in AI-Assisted Child Welfare Decision-Making through Worker-AI Interface Design Concepts. In *Designing Interactive Systems Conference*, 454–470.

Kim, S. S.; Meister, N.; Ramaswamy, V. V.; Fong, R.; and Russakovsky, O. 2022. HIVE: Evaluating the human interpretability of visual explanations. In *European Conference on Computer Vision*, 280–298. Springer.

Kim, S. S.; Watkins, E. A.; Russakovsky, O.; Fong, R.; and Monroy-Hernández, A. 2023. " Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–17.

Lakkaraju, H.; Bach, S. H.; and Leskovec, J. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 1675–1684. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342322.

Liu, H.; Lai, V.; and Tan, C. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–45.

Lundberg, S. M.; Erion, G. G.; and Lee, S.-I. 2018. Consistent Individualized Feature Attribution for Tree Ensembles. arXiv:1802.03888.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, 4765–4774.

Lundberg, S. M.; Nair, B.; Vavilala, M. S.; Horibe, M.; Eisses, M. J.; Adams, T.; Liston, D. E.; Low, D. K.-W.; Newman, S.-F.; Kim, J.; et al. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10): 749–760.

Poursabzi-Sangdeh, F.; Goldstein, D. G.; Hofman, J. M.; Vaughan, J. W.; and Wallach, H. 2021. Manipulating and Measuring Model Interpretability. *arXiv:1802.07810 [cs]*. ArXiv: 1802.07810.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. KDD '16, 1135–1144. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342322.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *AAAI*, volume 18, 1527–1535.

Saabas, A. 2015. Interpreting random forests.

Shen, H.; and Huang, T.-H. 2020. How useful are the machine-generated interpretations to general users? a human evaluation on guessing the incorrectly predicted labels. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, 168–172.

Singla, S.; Eslami, M.; Pollack, B.; Wallace, S.; and Batmanghelich, K. 2023. Explaining the black-box smoothly—A counterfactual approach. *Medical Image Analysis*, 84: 102721.

Tonekaboni, S.; Joshi, S.; McCradden, M. D.; and Goldenberg, A. 2019. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, 359–380. PMLR.

Yalcin, O.; Fan, X.; and Liu, S. 2021. Evaluating the Correctness of Explainable AI Algorithms for Classification. arXiv:2105.09740.

Zhang, Y.; Liao, Q. V.; and Bellamy, R. K. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 295–305.