Proxy Model Explanations for Time Series RNNs

Zach Wood-Doughty
Northwestern University
Evanston, IL 60208
zach@northwestern.edu

Isabel Cachola
Johns Hopkins University
Baltimore, MD 21218
icachola@cs.jhu.edu

Mark Dredze
Johns Hopkins University
Baltimore, MD 21218
mdredze@cs.jhu.edu

***Abstract*—While machine learning models can produce accurate predictions of complex real-world phenomena, domain experts may be unwilling to trust such a prediction without an explanation of the model's behavior. This concern has motivated widespread research and produced many methods for interpreting black-box models. Many such methods explain predictions one-by-one, which can be slow and inconsistent across a large dataset, and ill-suited for time series applications. We introduce a proxy model approach that is fast to train, faithful to the original model, and globally consistent in its explanations. We compare our approach to several previous methods and find both that methods disagree with one another and that our approach improves over existing methods in an application to political event forecasting.**

***Index Terms*—interpretability, explainability, causality, time series, recurrent neural network**

I. Introduction

Machine learning (ML) methods have demonstrated their ability to make accurate predictions across many domains. ML methods can automate decision-making based on consideration of high-dimensional and heterogeneous data that may be overwhelming for human domain experts. In high-stakes domains such as clinical care or political decision-making, high accuracy alone is not necessarily enough to motivate expert adoption of automated systems. If a neural network for medical image analysis arrives at a diagnosis and recommended treatment that contradict a physician's judgment, that doctor may be unwilling to accept the recommendation without an explanation as to the system's reasoning [29]. Yet modern neural networks, with millions or billions of parameters, may be impossible for a non-expert to understand [10]. Trust in ML models is a function of both accuracy and explainability; model predictions need to be accompanied by an explanation that can be interpreted by the people who will decide whether to act on that prediction.

This need to understand model predictions has motivated a large body of ML research into interpretability and explainability. The need for explanations introduces a number of competing goals. The local versus global explanation trade-off means that more detailed explanations for a single example (local) can provide greater insights but may also make it more difficult to understand overall trends (global) in the model's behavior if each explanation is overly specific to a single example. Another trade-off comes in the division of explainability methods into two types: explainable-by-design models and post hoc explanations for an existing trained model. Explainable-by-design requires training a model from scratch, which is unattractive when an application-specific model has already been developed and validated. However, many post hoc methods explain a single example in isolation, which can be slow and inconsistent across examples [30].

We balance trade-offs between local versus global explanations, and explainable-by-design versus post hoc through the introduction of a *proxy model*. A proxy model is a new, inherently-interpretable model that globally approximates the behavior of a complex ML system. The proxy model itself is explainable-by-design, but is trained to closely mirror the predictive behavior of an existing trained model. Whereas much previous work was primarily developed for computer vision applications and convolutional neural network (CNN) models, our method draws inspiration from causal inference research to control for the confounding effect of past predictions for our time series application.

To validate our approach we consider a challenging task that relies on a deep learning model: forecasting global disruptive political events (i.e. irregular government leadership changes) using a recurrent neural network (RNN). We compare a proxy model for generating explanations against several methods from popular explainability toolkits. We find that existing methods are inconsistent, making aggregated understanding of feature importance values difficult. Our proxy model is several orders of magnitude faster than these methods for generating explanations and produces globally-consistent estimates of feature importance across examples, providing a more holistic summary of the trained model.

II. Explainable AI

Recognition of the importance of explanations has driven a wave of research in Explainable Artificial Intelligence (XAI). We summarize several relevant research trends that our work builds upon. See [2], [8] for a comprehensive review.

"Post hoc" methods of explanation assume the existence of a trained ML model that requires explanation. This popular approach is representative of the common setting where an effective model exists but is difficult to interpret; for example, the model may be a state-of-the-art neural network with billions of parameters [10], [11], [24]. This approach excludes a variety of methods that are designed to train a new model in such a way that it is interpretable by design [25], but this may require substantial changes to an existing effective solution.

LIME [22] is one of the most commonly used post hoc methods for explaining black-box ML models. LIME explains a model's predictions one at a time by perturbing the input to the model and observing how those perturbations change the output. It then fits a linear model to the (perturbed input, model output) examples to produce a local explanation of which features have the greatest impact on the model's predictions. LIME has been widely used to explain ML methods for tasks ranging from predicting strokes [20] to labeling offensive language [23] Despite widespread adoption, it has known issues. First, sampling hundreds or thousands of perturbed inputs is expensive. Second, LIME focuses exclusively on *local* explanations which do not lend themselves to aggregation that can explain the model's general behavior [3], [30]. Finally, its sampling approach may also be vulnerable to adversarial attacks that can produce misleading explanations which mask malicious behavior [28].

There are many methods other than LIME that provide explanations of black-box models. While a comprehensive survey is outside the scope of this work [12], we consider three widely-used salience methods, which use the gradient of a neural network to determine which features are most influential for a given prediction. These methods take in the trained model and its input and compute gradients of the model's output with respect to its feature values. We use implementations for these three methods found in the `iNNvestigate` package [1]. The first of these is referred to as "Simple Gradient" or "Gradient" and is just the raw gradient of the prediction with respect to the inputs [27]. Input × Gradient ($\mathrm{I} \times \mathrm{G}$) takes that same gradient and multiplies it by the values of the input features [26]. Layer-Wise Relevance Propagation computes gradients in a modified backpropagation approach that maintains a local conservation property, which can be viewed as a sequence of Taylor decompositions [4], [17]. `iNNvestigate` provides many additional methods based on other published work, but we found that all other methods either could not run on our non-CNN model or duplicated another method's explanations. Future work could compare against additional methods.

III. Politically Disruptive Event Forecasting

While our proposed methods are generally applicable to ML systems, we focus on a challenging task with a complex model: the forecasting of politically disruptive events. This task highlights the ability of our proxy method to explain a complex model with high-dimensional time series data.

Our data comes from Global Data on Events, Location, and Tone (GDELT), which includes millions of geolocated events [15] in the form of monthly aggregated counts of news articles that reference events in 164 countries. These events cover different types of political actions, e.g. "Threaten to halt negotiations." As the amount of news coverage ingested into the GDELT dataset grows year-to-year from 2002 to 2019, each month's event counts are normalized by the total number of events of that type worldwide. Our data supplements GDELT events with country-specific features drawn from the World Development Indicators (WDI) [7].

TABLE I
Faithfulness evaluation using Spearman ρ, Kendall τ, Average Precision Score (APS), and ROC Area Under the Curve (AUC).

Comparison	Crystal Cube				Ground Truth	
Model	ρ	τ	APS	AUC	APS	AUC
Crystal Cube	1	1	1	1	0.037	0.776
Baseline	0.024	0.020	0.033	0.504	0.009	0.506
Proxy	0.907	0.799	0.904	0.995	0.029	0.770
L1 Proxy	0.923	0.816	0.869	0.989	0.031	0.778

Our goal is to utilize these features to forecast Irregular Leadership Change (ILC) events drawn from [21], defined as a political leadership change that does not follow the country's normal laws or conventions [5]. These events are rare, with only 180 occurring in the data between 2002 and 2019. This complicates training and evaluation, as a model can achieve 99.5% accuracy by never predicting an ILC event.

We begin with "Crystal Cube," [6], [18] a pre-existing model developed to predict ILC events from the GDELT and WDI data. Crystal Cube uses an LSTM [13] with 1454 input features derived from GDELT article counts and WDI statistics, with an instance representing each month for each country. The model maintains a hidden state that is updated each month before making a prediction, which allows the model to capture ongoing changes in a country's political environment that may foreshadow an ILC event.

Crystal Cube achieves a ROC AUC score of 0.776 and an average precision score (APS) of 0.037 for the task of predicting ILC events. The gap between these metrics is due to the imbalanced nature of the labels. [18] introduced the model and demonstrated that it outperformed several forecasting baselines. For comparison, our logistic regression baseline in Section V-A (Table I) achieves an ROC AUC of 0.506 and APS of 0.009. Crystal Cube is a complex neural network with 50 million parameters. However, it is also a carefully-trained, domain-specific model that achieves good performance, and thus we seek to create explanations for it, rather than replace it by training an explainable-by-design model from scratch [25]. This makes it a perfect application for our post hoc proxy model method. We refer readers to [18] for full details on the dataset and model training and evaluation.

IV. Proxy Model Explanations

We utilize a linear regression as a proxy model to explain Crystal Cube's predictions. Our proxy model takes as input the same data as Crystal Cube but is trained to predict its probabilistic outputs, rather than the binary ground-truth labels. We model the output probabilities to allow the proxy to learn from the detailed behavior of Crystal Cube, rather just the binary decisions. The proxy provides explanations based on its few parameters that can be individually understood as measuring the contribution of a feature to Crystal Cube's predicted probabilities. Our work is closely related to that of [31], though our focus is unique to time series data. In the context of other explainable AI research, our approach is a post

TABLE II
MOST IMPORTANT FEATURES ACROSS METHODS

	Gradient		I × G		LRP		LIME		Proxy		L1 Proxy	
Feature (week offset)	Rank	Value	Rank	Value	Rank	Value	Rank	Value	Rank	Value	Rank	Value
Accede to demands for change in leadership (wk-1)	20	1.2e-3	4	4.5e-5	22	5.6e-4	85	4.1e-3	174	0.112	2	0.178
Accede to demands for change in leadership (wk-2)	32	1.1e-3	2	4.9e-5	33	5.2e-4	28	4.9e-3	592	0.030	5	0.096
Accede to demands for political reform (wk-1)	31	1.1e-3	5	4.4e-5	31	5.2e-4	154	3.3e-3	388	0.048	14	0.029
Accede to demands for political reform (wk-2)	58	9.8e-4	3	4.7e-5	61	4.7e-4	132	3.5e-3	211	0.092	21	0.021
Accede to demands for political reform (wk-5)	60	9.8e-4	62	1.4e-5	62	4.7e-4	301	2.6e-3	120	0.165		
Appeal for change in leadership (wk-1)	30	1.1e-3	71	1.2e-5	30	5.3e-4	59	4.4e-3	509	0.036		
Appeal for change in leadership (wk-2)	3	1.4e-3	36	2.0e-5	3	6.5e-4	7	6.3e-3	163	0.118		
Appeal for change in leadership (wk-4)	41	1.0e-3	22	2.8e-5	42	4.9e-4	13	5.9e-3	310	0.062	9	0.042
Appeal for change in leadership (wk-5)	26	1.1e-3	52	1.6e-5	28	5.3e-4	112	3.7e-3	145	0.131	6	0.056
Demand release of persons or property (wk-2)	124	7.9e-4	170	6.0e-6	125	3.8e-4	1	8.1e-3	794	0.021		
Previous Model Prediction									72	0.814	1	0.862
Threaten with administrative sanctions (wk-2)	1	1.8e-3	912	8.4e-7	1	8.6e-4	1318	3.2e-4	472	0.039		
World Development Index Feature 23	1239	1.1e-4	1	5.4e-5	1246	5.4e-5	66	4.3e-3	21	44.661		
World Development Index Feature 3	1363	6.9e-5	29	2.6e-5	1369	3.2e-5	638	1.7e-3	1	5.1e+2		

hoc method [9] that is simulatible, algorithmically transparent, and decomposible [16].

Unlike the sequential structure of Crystal Cube, our proxy predicts the probability of a country's ILC using only the features of a single month. To address the time series nature of the application, we also give the proxy model access to the previous month's prediction probability from Crystal Cube. This draws inspiration from causal inference methods, in that it allows our model to *adjust for confounding* of the Crystal Cube's prior state [19]. Because Crystal Cube is a recurrent model, its month-to-month predictions are highly correlated. Our proxy model seeks to *explain* the recurrent model by pointing to which features in a given month caused the model to change its prediction.

While a linear proxy model is easily interpretable, for it to be effective it must be *faithful* to Crystal Cube's predictions [9], [14]. In contrast to previous work (see Section II), we can explicitly measure the faithfulness of our explanations by evaluating how closely our proxy outputs match those of Crystal Cube using a held-out test set of predictions.

Our training set consists of the GDELT and WDI data from 2002 to 2011, and our test set covers 2012 to 2019. Our loss function is the mean squared error (MSE) between proxy outputs and the real-valued probability outputs of Crystal Cube. To evaluate faithfulness, we compute the held-out Spearman ρ correlation and Kendall τ correlation between Crystal Cube's predictions and the proxy model's outputs. In addition to these regression-based metrics, we threshold (at 0.22, following [6]) Crystal Cube's predictions to create binary labels, and compute ROC AUC score and APS comparing our real-valued proxy outputs against these binary labels. Finally, we compute ROC AUC and APS with true ILC event labels.

We use `sklearn` models to enable reproducibility. A simple linear regression model without regularization is denoted as 'Proxy' in our tables. We then introduce an L_1 (Lasso) regularization penalty with α of 5e-5, which induces sparsity into models coefficients. We denote this model as 'L1 Proxy.' We compare both proxies against a baseline logistic regression ('Baseline') that is trained to directly predict the ILC event labels, independently of Crystal Cube. We expect the baseline to correlate with Crystal Cube as they model the same data, but our proxy models should be much more faithful.

TABLE III
JACCARD SIMILARITY BETWEEN METHODS

	Gradient	I × G	LRP	LIME	Proxy
I × G	0.076				
LRP	1.000	0.076			
LIME	0.058	0.058	0.058		
Proxy	0.000	0.506	0.000	0.024	
L1 Proxy	0.072	0.169	0.072	0.030	0.000

V. EVALUATION

A. Proxy Model Evaluation

Our first evaluation explores whether our proposed proxy model approach is *faithful* to Crystal Cube. Table I shows our evaluation of Crystal Cube, the logistic regression baseline, Proxy, and L1 Proxy models on how well they predict the held-out test set Crystal Cube predictions and the ground truth labels for ILC events, using the faithfulness metrics described in Section IV. All of these metrics have a maximum of 1.0; Crystal Cube perfectly correlates with and predicts itself.

The logistic regression Baseline has almost no predictive power for either Crystal Cube's predictions or the true ILC event labels. The unregularized Proxy model scores well on all four faithfulness metrics, while losing 0.008 and 0.006 points on APS and AUC scores respectively for predicting ILC events. Introducing L1 regularization further reduces the APS and AUC scores against Crystal Cube, but increases the correlation scores and the ground truth predictive performance. While these results do not suggest we should simply discard Crystal Cube and use our proxy models to predict events, the proxies perform extremely similar to Crystal Cube across the entire test set. Although L1 Proxy scores slightly lower than Proxy on predictive metrics of faithfulness, with only 40 nonzero features (compared to Proxy's 1455) it provides an even simpler representation of Crystal Cube's behavior.

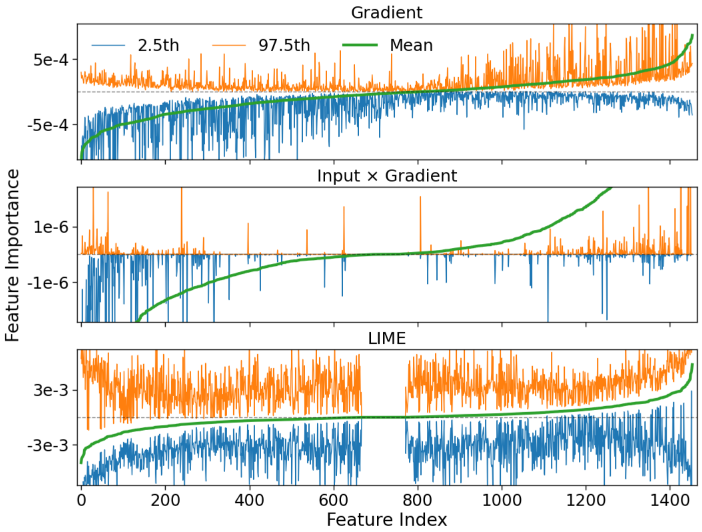

Fig. 1. Distribution of feature values for local methods. For almost every feature, the empirical distribution of importance scores contains zero between its 2.5th and 97.5th percentiles. Local methods are not consistent in determining whether features make ILC predictions more or less likely. Whitespace appears in I × G and LIME plots because there are 713 and 105 features, respectively, where both the 2.5th and 97.5th percentiles are 0.

B. Differences Between Explanations

We now compare our proxy model explanations against four methods for producing post hoc local explanations described in Section II. Each local method uses the trained Crystal Cube model and its input features to individually explain the model's prediction for each country and month. From the `iNNvestigate` library released by [1], we use the Gradient, I × G , and LRP methods for local explanations. LIME is our fourth local method, using the `LimeTabularExplainer` class in the published package. Due to runtime constraints, we reduced LIME's number of sampled perturbations from the default of 1000 to 500, but otherwise left all defaults unchanged. We run these methods across all countries and months in our train and test sets. Table II shows the top features across methods when aggregating by taking the mean absolute value of each feature importance across all examples, regardless of the ground-truth ILC event label and whether Crystal Cube correctly predicted it. For each feature, we show its rank and its mean absolute importance according to each method. The table shows several top-ranked features, revealing several trends:

1) The scale of importance scores is different between methods; I × G has a maximum value of 5.4e-5 compared to Proxy at 5.1e+2.
2) Many of the top features are either "accede to demands" or "appeal for change" events. However, even if we focus only on the "appeal for change in leadership" events within the same method, we see large variance in the feature importance and ranking between weeks.
3) Gradient and LRP are the most correlated methods, whereas Proxy and L1 Proxy are surprisingly unrelated.

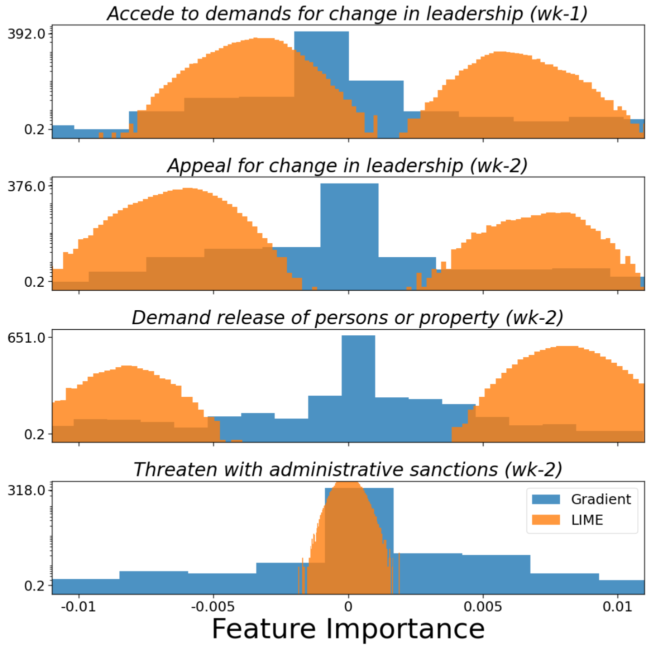

Fig. 2. **Log-scale** histograms for the distribution of importance scores for a few top features from Table II. For the Gradient method (and I × G and LRP not shown), almost all importance scores are close to zero, and are otherwise roughly symmetric. For three of the LIME features, the distribution is asymmetric but bimodal.

We now more closely examine correlations between methods. Table III shows Jaccard similarities between the top 64 most important features for each method. We rank features by taking the mean of their absolute importance scores across all countries and months to also account for negative values (features that reduce the predicted probability of an ILC event). We compute Jaccard similarities involving the L1 Proxy using only its 40 nonzero features. We see that Gradient and LRP methods are equivalent in terms of Jaccard similarity, even though the scale of their feature importances are quite different in Table II. Except for that pair of methods, no two methods are particularly closely related. I × G most closely overlaps with the unregularized Proxy, followed by the L1 Proxy. Other than these examples, no pair exceeds a modest similarity score of 0.1. Because there is no ground truth for the *correct* way to explain Crystal Cube's predictions, we cannot directly evaluate whether one method provides better explanations than the rest. We can either choose to aggregate across methods (such as in Table II) or use domain knowledge to choose a specific method.

The methods we consider make different assumptions about the nature of how a trained model should be explained, and thus it is not entirely surprising that they produce different explanations. For a domain expert, a greater acceptance of one method's assumptions may give more weight to its explanations. While we do not have expert testimony for this data, we can provide some analyses of the domain-independent pros and cons of the various methods.

C. Runtime

A major benefit of our proposed proxy model approach, in addition to its overall simplicity, is its speed. Whereas the other methods we consider make *local* explanations and thus must consider each combination of month and country independently, our proxy model considers all such examples simultaneously. This results in an improvement in runtime of several orders of magnitude. Training a proxy model regression takes only a few minutes on the 20k training examples; once trained, its parameters do not change between explanations. In contrast, the local explanation methods we compare against require at least a few seconds for each prediction they seek to explain. Because our train and test set together contain 164 countries and 215 months, this results in over 40 hours of computation for the Gradient method and over 140 hours for LIME. Although parallelization can decrease the wall-clock time dramatically, these runtime numbers reflect the total time to run on a single CPU. However, the enormous speed advantage of our proxy model approach can enable analyses that might otherwise be too expensive.

D. Consistency and Plausibility

Another differentiating factor between the proxy model and local explanations methods is the variance within the explanations of each method. The proxy model produces only a single importance score for each feature across the entire dataset. Each local method produces an importance score per feature per example, which means we must aggregate those explanations to provide a summary of the overall behavior of the model. In doing so, we must handle variance: what happens if a feature is sometimes highly predictive of an ILC event and sometimes predictive of no event?

Figure 1 shows plots of the mean and percentiles for the feature importance values of the Gradient, $I \times G$, and LIME methods. In these plots, the x-axis is an ordering over the features based on their mean absolute importance for the given method. For each slice at $X = i$, we see the mean and percentiles for the ith feature. For almost every feature across these methods, its empirical distribution of feature importance contains zero between its 2.5th and 97.5th percentiles. This means that even for features that are *on average* predictive of an ILC event occurring, i.e. the ‘useful’ features, these features are still often found to be predictive of no event. The way these methods explain these features is inconsistent; it may be locally informative but is not globally reliable.

Figure 2 shows a closer look at the distributions of individual features for the Gradient and LIME methods, using a few of the most important features from Table II. The LIME plots show an asymmetric bimodal distribution in three of the four examples, with the feature being labeled as either positive-predictive or negative-predictive, depending on the context. This includes “Demand release of persons or property (wk-2),” the overall most important feature for LIME. The Gradient plots show a nearly point-mass distribution; for “Threaten with administrative sanctions (wk-2),” over 90% of Gradient’s feature importance scores are between -1e-5 and 1e-5, and the remaining density is split 56% negative and 44% positive. While we omit LRP and $I \times G$ histograms, they are similar[1] to the Gradient plots.

These results indicate that for the local methods, whether and how we aggregate feature importance values can provide vastly different explanations. If a feature’s average importance is dominated by a few outliers and is otherwise symmetric and mean-zero, it may be difficult to interpret whether past explanations should influence our understanding of future model predictions. This may jeopardize the ability of these methods to instill trust in domain experts who do not understand the underlying black-box model.

In this context, a benefit of our proxy model approach is its *consistency*. In training our Proxy and L1 Proxy models, we assign a single importance score to each feature. This naturally aggregates the importance of each feature across the entire dataset. The addition of L1 regularization also greatly reduces the number of nonzero features, allowing the L1 proxy to summarize Crystal Cube’s global behavior as a collection of 40 features. Looking at features in Table II, we can see that the L1 Proxy accounts for the temporal correlation between Crystal Cube predictions by putting the most weight on the “previous prediction” feature. For a domain expert, this can be interpreted as saying: “Crystal Cube tends to reduce its prediction probability by 13% unless one of these other 39 features is present in this month’s data.” For the time series application we consider, this may be a helpful insight that might be otherwise unnoticed through the lens of local explanations.

VI. Discussion

We have introduced a proxy model approach for explaining a black-box time series RNN. Our experimental evaluation on a dataset of rare political events demonstrates that our method provides a holistic explanation of the model’s behavior across an entire dataset. In this setting, our method is several orders of magnitude faster than existing, local methods of explanation that use sampling or gradients. While local methods such as LIME provide inconsistent (local) explanations across instances, our approach provides a globally-consistent explanations: features are either explained as predictive of ILC events, predictive of no event, or unimportant. Our regularized proxy model highlights a small number of features in a transparent linear model while remaining faithful to the original model. This improves over local methods, for which a global explanation requires aggregating over a high-variance distribution of importance scores.

In an applied setting where a domain expert seeks to understand the predictions of a black-box model, our proxy model approach provides unique benefits. Its speed makes it easy to apply the analyses of Table I to a new black-box model and dataset, which clearly shows the proxy’s faithfulness to

[1] For “Threaten with administrative sanctions (wk-2),” 92% of LRP scores lie in (-1e-5, 1e-5) and the remainder are split 59% to 41%. For “Accede to demands for change in leadership (wk-2),” 96% $I \times G$ scores lie in (-1e-7, 1e-7) and the remainder are split 42% to 58%.

the trained model. In applications where model predictions are correlated across time, our proxy model explicitly accounts for this by putting a 'feature' weight on the previous prediction. If long-ago events have made an ILC more likely, it may be that there is a high 'baseline risk' rather than a specific event this month. Local methods do not handle this explicit structure. Once trained, the proxy model naturally aggregates the importance of features into a holistic and interpretable explanation. This is particularly true in our L_1-regularized case, where sparsity reduces the number of nonzero features. This provides a proxy that is simulatible and decomposible [16] – a domain expert can 'decompose' the coefficients of the linear proxy and understand (simulate) how they are individually combine to produce the prediction. For an expert who knows the features and their real-world relevance, such a holistic explanation allows them to easily compare their own judgment against the proxy model's explanation.

There are additional pros and cons to our approach. A benefit of our method is that it can be applied in settings without access to the black-box model itself, as long as its data inputs and predictions are available. While this was irrelevant to our Crystal Cube application, it may make our method particularly helpful in cases where the black-box model is proprietary or otherwise restricted. A drawback of our method is that it does not provide specific details about any individual data instance, which may make it less helpful for understanding edge cases of the black-box model's behavior. Unlike local methods, we require enough examples on which to train a proxy. Finally, there is no guarantee that a simple linear model can provide a faithful proxy for a given black-box model; however, we can easily evaluate faithfulness in a way LIME cannot [3]. Our results also suggest opportunities for future work in Explainable AI. A proxy model can provide fast, global explanations that could inform which examples require additional local explanation, by checking where the black-box model's predictions differ from those of the proxy. Proxy model importance scores could similarly inform LIME's sampling procedure by suggesting which features need to be perturbed to learn a reliable surrogate explanation. Future work should also explore human evaluations to understand domain expert's subjective opinions of proxy model explanations.

Acknowledgements

We acknowledge support provided by the Johns Hopkins Institute for Assured Autonomy. We thank Anna Buczak, Benjamin Baugher, and Adam Berlier for sharing the Crystal Cube model and data.

References

[1] M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K.-R. Müller, S. Dähne, and P.-J. Kindermans, "iNNvestigate neural networks!" *Journal of Machine Learning Research*, vol. 20, no. 93, pp. 1–8, 2019.

[2] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[3] P. Atanasova, J. G. Simonsen, C. Lioma, and I. Augenstein, "A diagnostic study of explainability techniques for text classification," in *EMNLP*, 2020, pp. 3256–3274.

[4] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, 2015.

[5] A. Beger, C. L. Dorff, and M. D. Ward, "Irregular leadership changes in 2014: Forecasts using ensemble, split-population duration models," *International Journal of Forecasting*, vol. 32, no. 1, pp. 98–111, 2016.

[6] A. L. Buczak, B. Baugher, C. Martin, M. Keiley-Listermann, J. Howard II, N. Parrish, A. Stalick, D. Berman, and M. Dredze, "Crystal cube: Forecasting disruptive events," *Under Review*, 2021.

[7] "World development indicators," DataBank. [Online]. Available: https://databank.worldbank.org/source/world-development-indicators

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

[9] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Communications of the ACM*, vol. 63, no. 1, pp. 68–77, 2019.

[10] S. Feng, E. Wallace, A. Grissom II, M. Iyyer, P. Rodriguez, and J. Boyd-Graber, "Pathologies of neural models make interpretations difficult," in *EMNLP*, 2018, pp. 3719–3728.

[11] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3681–3688.

[12] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] A. Jacovi and Y. Goldberg, "Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?" in *ACL*, 2020.

[15] K. Leetaru and P. A. Schrodt, "GDELT: Global data on events, location, and tone," in *ISA Annual Convention*, 2013.

[16] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[17] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: an overview," *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 193–209, 2019.

[18] N. H. Parrish, A. L. Buczak, J. T. Zook, J. P. Howard, B. J. Ellison, and B. D. Baugher, "Crystal cube: Multidisciplinary approach to disruptive events prediction," in *AHFE*. Springer, 2018, pp. 571–581.

[19] J. Pearl, *Causality*. Cambridge university press, 2009.

[20] N. Prentzas, A. Nicolaides, E. Kyriacou, A. Kakas, and C. Pattichis, "Integrating machine learning with symbolic reasoning to build an explainable ai model for stroke prediction," in *BIBE*. IEEE, 2019.

[21] C. Raleigh and C. Dowd, "Armed conflict location and event data project (acled)," Jun 2021. [Online]. Available: https://acleddata.com/

[22] M. T. Ribeiro, S. Singh, and C. Guestrin, "'why should I trust you?' explaining the predictions of any classifier," in *KDD*, 2016.

[23] J. Risch, R. Ruff, and R. Krestel, "Offensive language detection explained," in *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, 2020, pp. 137–143.

[24] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in bertology: What we know about how bert works," *TACL*, vol. 8, 2020.

[25] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[26] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," *arXiv:1605.01713*, 2016.

[27] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv:1312.6034*, 2013.

[28] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," in *AIES*, 2020, pp. 180–186.

[29] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: contextualizing explainable machine learning for clinical end use," in *Machine learning for healthcare conference*. PMLR, 2019.

[30] I. van der Linden, H. Haned, and E. Kanoulas, "Global aggregations of local explanations for black box models," in *FACTS-IR Workshop*, 2019.

[31] Z. Wood-Doughty, I. Cachola, and M. Dredze, "Faithful and plausible explanations of medical code predictions," *arXiv:2104.07894*, 2021.