The Importance of Incorporating Data Context When Developing XAI Frameworks

Saif Anwar
saif.anwar@warwick.ac.uk

Nathan Griffiths
nathan.griffiths@warwick.ac.uk

Abhir Bhalerao
abhir.bhalerao@warwick.ac.uk

Thomas Popham
thomas.popham@warwick.ac.uk

Mark Bell
mbell@trl.co.uk

Abstract

There is increasing attention and importance being to Machine Learning (ML) techniques, due to their ability to make accurate decisions in many applications. Despite this, the trustworthiness of these techniques is difficult to assess, which is important in high-risk or ethically sensitive scenarios. Many techniques using complex ML models, such as artificial neural networks, are treated as a 'black-box' where the reasoning or criteria for the final decision is opaque to the user. Explainable AI (XAI) research aims to 'unbox' complex models and provide reliable explanations to support adoption in high-risk fields such as healthcare, finance and criminal justice. Some existing XAI approaches approximate model behaviour using perturbed data. However, such methods have been criticised for ignoring feature dependencies, with explanations being generated using potentially unrealistic data. This paper highlights the importance of incorporating data context into XAI, by showing how its absence produces inaccurate explanations. We propose CHILLI, a novel approach for incorporating data context into XAI by generating contextually appropriate data perturbations, which maintain faithfulness to the existing data used to train the black-box model being explained. This is shown to improve both the soundness and accuracy of the explanations, increasing trust in both the explanation and the AI model.

1. Introduction

Machine Learning (ML) and Artificial Intelligence (AI) are increasingly being used to tackle problems in a variety of domains because of their exceptional performance for automated decision-making. Some of these domains have high associated risks, such as financial systems [1, 4], healthcare [14, 17] and criminal justice [31]. Incorrect decisions in these fields can have significant repercussions, such as intentional or inadvertent biases leading to discrimination and other social consequences [27, 38]. Therefore, it is essential that decisions made by an ML model can be trusted before being acted upon. The foundation of such trust is dependent on both developers and end users understanding the reasoning for model decisions.

Due to the complexity of many ML techniques, it is often difficult to understand why a model has made a particular decision. They are treated as a 'black-box' which maps input features to an output. Understanding the behaviour of an AI system allows users to detect biases in data, assess the vulnerabilities of a model, and ensure a model meets regulatory standards [10] and societal expectations. Explainable AI (XAI) methods aim to increase confidence in AI systems, supporting their adoption. Properties of XAI, such as explainability and interpretability, do not have fixed definitions in existing literature. However, the National Institute of Standards

and Technology (NIST) proposes principles that encompass the core concepts and goals of XAI [26]. These principles allow us to define key terms that are used throughout this paper.

- **Explainability:** An AI system must provide evidence or reasoning for all outputs.
- **Interpretability:** The explanations must be understandable to users.
- **Faithfulness:** The explanations should accurately reflect the behaviour of the system.

There exist a range of XAI methods, which attempt to address the various concerns associated with AI systems, and ML techniques in particular. Some ML methods are inherently interpretable [37], such as decision trees, where a model is designed with XAI goals in mind and can implicitly explain model behaviour [3]. Post-hoc XAI methods attempt to form explanations after a prediction has been made. These are often developed with the aim of being general purpose and applicable to a wide range of ML methods and contexts [9, 19, 28], with some explanation methods being model-agnostic. However, evaluations of such approaches have concluded that, just as ML systems are tailored to specific contexts, XAI systems should also be tailored to the appropriate deployment domain [44, 36]. Numerical data typically represents quantitative information, such as quantities, measurements, and statistics, which makes it challenging to convey any meaningful content or context in its raw values alone. To extract insights and meaning from numerical data, it often requires further interpretation that can provide context. For example, a numerical value of 0.5 may represent a probability of 50% or a value halfway between 0 and 1. This is a common problem in XAI, where the raw values of features are used to explain model predictions without contextual knowledge regarding the data. This is a problem that has been highlighted in the literature [44, 36]. Images and text data, on the other hand, may contain salient objects or word embeddings which provide context to facilitate explanations. Earlier works [13, 34] discuss the importance of context sensitivity for computer systems. This is crucial for XAI frameworks, since we rely on them for a trustworthy understanding of complex methods in high-risk scenarios. An XAI framework explaining a model trained on tabular data requires underlying domain knowledge to incorporate the appropriate semantics into its explanation. In this paper we explore the effects on faithfulness and interpretability, of incorporating contextual domain knowledge into XAI frameworks. The aim of this paper is to highlight the importance of data context when explaining black-box model predictions by improving the faithfulness of an explanation towards the model being explained. We aim to demonstrate this in both an intuitive and quantitative manner, to increase understanding and trust in explanations of complex model behaviour. The main contributions of this paper are as follows.

- We present an analysis of an existing XAI framework (LIME [28]) to explore the impact of disregarding data context on explanations in terms of their interpretability and faithfulness.
- We propose a method for incorporating data context into a post-hoc XAI framework when using a proxy model.
- Finally, an algorithm is presented for generating local contextually appropriate data samples for use when fitting a proxy model.

2. Background

2.1. Existing XAI Methods

Explanations formed using XAI methods can be classified through their scale [18]. Global explanations attempt to explain the behaviour of an overall model using techniques such as model visualisation [40] or decision rules [11]. On the other hand, local explanations are limited to a certain part of the model or around particular data instances using local approximations [28, 29] or saliency maps [43, 2]. Illustrations of a global and local scale are shown in Figure 1, which shows a binary classifier with the red and green background representing the decision boundary.

2.1.1. Inherently Interpretable Models

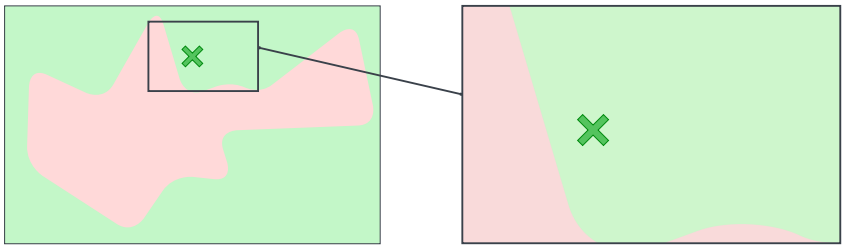
Certain types of models can be classed as inherently interpretable, i.e., their structure is interpretable without additional effort. Examples of these are simpler model types such as linear regression, where feature coefficients can be observed [23], and decision trees, where the decision path can be directly traced. They are often referred to as 'glass-box' models given their inherent transparency, due to the relatively small number of features and parameters involved. We define transparency, which is often used interchangeably with explainability, to be a property related to how visible and open the inner mechanisms of a ML system are. In these cases, the explanation is the model itself and is therefore completely faithful to the behaviour of the model. As a result, such models are often favoured in high-risk scenarios [32].

2.1.2. Post-hoc Approaches

It may not always be appropriate to use an inherently interpretable model, for example if the aim is to explain an existing model which has been tuned over a long time. Moreover, for some applications the best performing models are highly complex and contain a large number of parameters [35] and are often not inherently interpretable [41].

Although there is increasing research in high performing inherently interpretable models [37], post-hoc XAI methods have been developed to explain existing models. These methods are generally model-agnostic and exploit feature correlations using only input and output data to understand model behaviour. Some methods use counterfactual examples involving modifying an input instance to assess the consequences on predictions [39]. Other XAI methods form explanations by presenting feature contributions, using techniques such as Shapley values [16] or proxy models.

Proxy models attempt to approximate the behaviour of a more complex model in a simplified form, such as a decision tree [33] or linear regression model [28]. This is done by fitting a simpler model on the existing training data used to fit the original complex model. The proxy model is used to understand behaviour in an interpretable form, without sacrificing the performance of the complex model. A global scale proxy model needs to encapsulate the complete behaviour of the complex model. However, achieving high fidelity to a complex model also increases the complexity of the proxy model, thus reducing its interpretability. It is generally accepted that there is a trade-off between faithfulness and interpretability [7]. Therefore, some XAI approaches use a local proxy model to reduce the coverage of the approximation, that is fit on data in a smaller region of the model space. This allows for a more faithful explanation in the covered area, whilst maintaining a lower complexity [42].

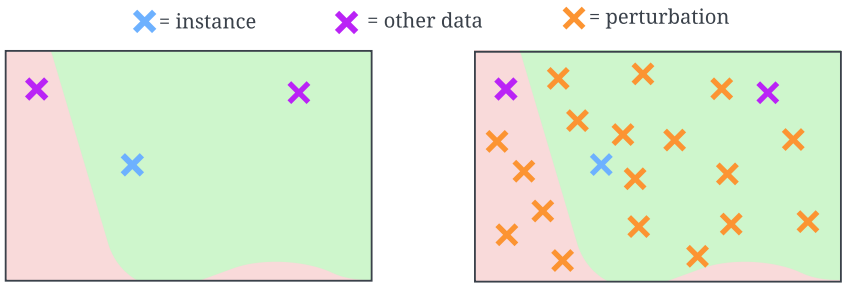


(a) (b)

Figure 1: (a) Depiction of the overall decision space, with the red and green backgrounds representing the decision boundary between the two classes and (b) a depiction of the local area around a single instance in the decision space.

2.1.3. Perturbation Based Methods

In some cases, there may not be enough training data to fit a proxy model. In these cases, perturbation based methods can be used to generate data samples which are similar to the original instance. Local Interpretable Model-Agnostic Explanations (LIME) [28] is a method which approximates the behaviour of a complex model around a particular instance by fitting a proxy model in the locality of the instance. Algorithms like LIME involve fitting a local model on a set of perturbations generated around the instance whose prediction is being explained, as illustrated in Figure 2. To ensure that the explanation is local to the instance, each perturbation is assigned a weighting based on some distance metric between the instance and the perturbation. This weighting affects the contribution of each perturbation towards the fitting of the proxy model.



(a) (b)

Figure 2: (a) A selected instance (denoted by a blue cross) in a local area of the decision space shown in Figure 1, along with other datapoints (purple crosses) which fall in this locality. (b) A set of perturbations (orange crosses) is generated which will be used to fit a local proxy model. Perturbations falling in the red region will be classified as such by the base model and vice versa for perturbations in the green region. A proxy model will be fit in the locality using the model predictions as target values.

In a review of model-agnostic XAI approaches, Molnar et al. [21] highlight that perturbation-based methods, which require extrapolation of data, tend to ignore feature dependencies. This means that perturbation-based

methods may extrapolate in areas which are outside the original data distribution, and are therefore unknown to the original model [22]. While generalising to unknown environments is a desirable property of an ML model, an explanation based on perturbations that cannot be associated to existing domain knowledge is undesirable, and is potentially dangerous in high-risk scenarios [12].

Molnar et al. also suggest that ignoring feature dependencies when extrapolating may lead to unrealistic data, since contextual constraints may not have been considered. For example, if creating perturbations of data with a feature representing a person's age, the perturbation method must consider that the feature value cannot be negative or unreasonably large. A proxy model which is fit to unrealistic data will not be trustworthy or reliable since it is not representative of the behaviour of the model being explained. It is suggested that additional information regarding the dependency between features should be included to avoid such potential pitfalls [22].

Many of the problems faced by model-agnostic approaches arise in situations where a method designed for a specific purpose is applied in an unsuitable context [21], particularly when feature dependence is not taken into account. Feature dependence may lead to inconsistent explanations for locally similar instances when using approaches such as LIME, which makes for an unstable and unreliable XAI framework [30, 15, 36, 44]. In this paper, we address the importance of feature dependence by proposing a framework for incorporating dependency information using prior domain knowledge, and evaluating its effect on the performance of explanations.

2.2. Evaluting XAI Methods

The desiderata of XAI involve subjective properties relating to trust, ethics and understanding. Given the variation in expertise of users, it is not possible to ensure interpretability for all users [18]. We argue that the terms used to present an explanation should be predetermined for a particular level of expertise. For example, a medical practitioner may be more familiar with the terminology used in the medical domain, while a layperson may not. An explanation intended for the layperson should not use complex medical terminology.

A satisfactory explanation provides transparency, allowing users to understand decisions, data usage, privacy features and so on. Quantifying properties such as transparency and interpretability is difficult, and therefore many XAI evaluations use human participants to measure performance. Doshi-Velez and Kim [6] outline a method testing human participants' understanding of a model through a range of tasks, such as asking users to predict a model's output given an input and explanation.

When evaluating trust using human participants, it has been shown that users tend to favour explanations that are persuasive in nature and simpler to understand rather than more complex, yet more accurate, representations of model behaviour [8]. However, it is inappropriate to favour a simplified explanation to promote acceptance when the limitations of the simplified explanation are not understood by users.

Given that the notion of interpretability is subjective, it becomes costly and time-consuming to evaluate XAI systems using human participants. In this paper, we will instead focus on assessing the faithfulness of XAI approaches by measuring how accurately an explanation represents the behaviour of the model being explained. In the case of proxy models, this can be numerically determined. The faithfulness of a proxy model, g, which is an explanation for a base model, f, can be characterised using an error metric such as Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE), by comparing the predictions made by f and g on some set of input datapoints, $\mathcal{X}$. Calculating the error over $|\mathcal{X}|$ instances, as shown in Equations 1 & 2, allows us to

evaluate the faithfulness of the explanation. For a global explanation, the inputs used for evaluation may be from the dataset used to train the base model, f [42].

$$\text{MAE}(\mathcal{X}) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} |f(x) - g(x)| \tag{1}$$

$$\text{RMSE}(\mathcal{X}) = \sqrt{\frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} (f(x) - g(x))^2} \tag{2}$$

When creating a local explanation around an instance, x, there may not be enough data to evaluate the faithfulness effectively. Therefore, methods such as LIME [28] form a set of perturbations, $\mathcal{Z}$, around an instance when fitting a proxy explanation model in this locality. The performance of the explanation is evaluated by comparing predictions from f and g on the perturbations, as shown by Equation 3 if RMSE was used as an error metric.

$$\text{RMSE}(\mathcal{Z}) = \sqrt{\frac{1}{|\mathcal{Z}|} \sum_{z \in \mathcal{Z}} (f(z) - g(z))^2} \tag{3}$$

Using this metric, an explanation, g, which is more faithful to the behaviour of f will have a lower error. From a set of possible proxy models, G, the proxy model with the lowest error is selected as the explanation.

2.3. Sythentic Minority Over-Sampling Technique (SMOTE)

SMOTE is a well-regarded sampling technique used to generate synthetic data when there is an imbalance between classes in a dataset [5]. For the minority class, a random point, x, is selected and its k-nearest-neighbours are located, for a predetermined integer, k. A random instance, $\hat{x}$, is selected uniformly from the nearest neighbours. To create a synthetic instance, each feature of x is interpolated in the direction of $\hat{x}$. The amount of interpolation is defined by a random value between 0 and 1, also selected uniformly at random and denoted as I. The difference between the feature vectors is calculated and multiplied by I to determine the value of each feature of the new synthetic datapoint, s. This is carried out for a number of different instances, x, from the minority class until a specified amount of over-sampling has been achieved. The process for generating a single synthetic datapoint, s, around a single minority class instance, x, is outlined in Algorithm 1.

Algorithm 1 SMOTE

Input: Minority class samples T; Selected Sample x; Number of nearest neighbours k

Output: Synthetic instance s

N_x = set of k nearest neighbours around x in T

Uniformly at random select a neighbour $\hat{x} \in N_x$

Uniformly at random select some value between 0 and 1 $\rightarrow I$

F = Features of x

for each f in F **do**

    $f_s = f_x + I(f_{\hat{x}} - f_x)$

end for each

3. Contextual Feature Dependency for Perturbation-Based XAI

3.1. Contextually Appropriate Proximity Measures

To explore the behaviour of features with contextually dependent characteristics in post-hoc XAI approaches, we base our approach on LIME [28], because of its wide acceptance within the research community. As mentioned in Section 2.1.3, a set of perturbations, $\mathcal{Z}$, is generated in the locality of an input instance, x, whose output prediction from some complex model, f, is being explained. A local proxy model, g, is then fit to these perturbations by minimising some loss function, where the loss contribution of each perturbation, z, is weighted by a proximity measure, $\pi_x(z)$, according to some distance function $\mathcal{D}(x, z)$. LIME calculates the proximity of two points based on the distance between their respective feature vectors using some distance function, $\mathcal{D}$, which by default is Euclidean distance [28]. The distance between points $\mathbf{p}$ and $\mathbf{q}$ is calculated using Equation 4, where d is the number of features. It is assumed that all data has been normalised, so that the contribution of each feature to the overall proximity is equal. The proximity between the two points, $\pi_p(q)$, is then calculated as shown in Equation 5, where σ is a hyperparameter which defines the locality within which an explanation is defined.

$$\mathcal{D}(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i \in d} (p_i - q_i)^2} \tag{4}$$

$$\pi_p(q) = exp\left(\frac{-\mathcal{D}^2(\mathbf{p}, \mathbf{q})}{\sigma^2}\right) \tag{5}$$

In this formulation, all proximity calculations are based on Euclidean distance, irrespective of the feature type. In contexts where for some features the absolute difference between two values does not reflect their distance, this proximity measure is not appropriate. For example, this is the case if the units of a value are not equidistant, or a feature is not measured linearly, such as magnitude which is often recorded on a logarithmic scale. Such distance measures are also unsuitable for cyclic temporal features e.g., time of day where the raw values for 23:00 and 00:00 appear to have a distance of 23, but our domain knowledge informs us that they are consecutive hours.

We propose that for XAI the context of features should be considered independently when calculating the distance between two values of a feature, by considering the scale and bounds of the feature to ensure the calculated distance is truly representative. Consider the points $\mathbf{p}$ and $\mathbf{q}$, represented by feature vectors of d dimensions. The distance between $\mathbf{p}$ and $\mathbf{q}$ can be calculated individually across each dimension, using a custom distance function for each feature, which is then averaged, as shown in Equation 6, with each feature having equal contribution to the overall distance due to normalisation.

$$\mathcal{D}(\mathbf{p}, \mathbf{q}) = \frac{1}{d} \sum_{i \in d} \mathcal{D}_i(p_i, q_i) \tag{6}$$

Using Equation 5, the distance is converted into its respective proximity measure as follows:

$$\pi_p(q) = exp\left(\frac{-(\frac{1}{d}\sum_{i \in d} \mathcal{D}_i(p_i, q_i))^2}{\sigma^2}\right) \tag{7}$$

Since the proximity measure is used by LIME when quantifying the performance of each explanation, g, from the set of possible explanations, G, an accurate proximity measure is essential to ensure the most faithful explanation is selected.

The value of the locality hyperparameter, σ, may be adjusted to vary the locality of an explanation in the model space. As σ increases, the proximity tends to 1, as shown by Figure 3 for a range of distance values. For a high enough value of σ, all perturbations, regardless of distance, will be assigned a proximity of 1 and considered equally when selecting the proxy model that is the best fit to the perturbations. Conversely, a smaller value of σ will result in greater variation between proximities for perturbations of differing distance to the instance being explained. This leads to the selection of a proxy model that performs better on perturbations of closer proximity, thus reducing the locality of the explanation.

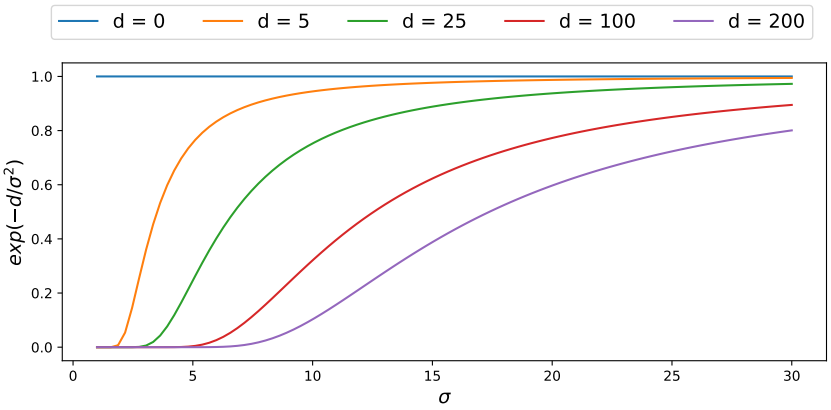


Figure 3: A comparison of the effect on proximity between two points of fixed distance, d, as the locality parameter, σ, is varied.

3.2. Domain Representative Perturbation Generation

Existing perturbation-based XAI methods, such as LIME, do not consider any prior contextual knowledge when generating the perturbations which are the foundation for creating an explanation [28, 30, 44, 36]. When generating perturbations of an instance, x, categorical features are perturbed by uniformly sampling the distribution of the feature across the entire dataset. This ignores the notion of locality since perturbations may contain feature values that are not local to x. If perturbations contain values across the range of possible values, rather than those local to x, a proxy model being fit to the perturbations, and target values obtained from the base model, will focus on all possible data relationships rather than those specifically relevant to the area around x. Therefore perturbations that are not local to x will give an explanation which is generalised to the overall behaviour of the model. Presenting this as a local explanation for a more complex model provides false insight into the local behaviour of the model and is therefore an untrustworthy explanation.

Non-categorical features are sampled in LIME from a Normal distribution with mean and standard deviation taken from the training data. Samples are not taken around the instance of interest, but from the training data's center. These are then scaled around the instance according to the covariance of the feature against the target variable which can be problematic since the samples do not represent the immediate locality around

the instance being explained [20]. Some features may have real-world bounds, such as 'Time' or 'Age' which can only take positive values. Without considering these bounds, perturbations of non-categorical features may contain unrealistic values. Moreover, given that each feature is perturbed independently, perturbations may be formed which ignore feature dependencies. For example, it can be assumed that as the population of a city grows, traffic congestion will increase. Since these two features are correlated, an XAI method that ignores this correlation may create a perturbation which combines high congestion with a low population. The omittance of such dependencies may result in an unrealistic set of perturbations.

When creating an explanation, it is important that the model is fit on data that conforms to any contextual bounds and is representative of the training data for the model being explained. An explanation which is fit to unrealistic data would not describe behaviour of the underlying model, resulting in an explanation which lacks reliability and trust [12]. An XAI framework should generate perturbations that are representative of real-world data, such as the training data, and are local to the instance being explained.

In this paper, we present a novel framework for generating local and contextually conforming perturbations which can be used to construct a local explanation of a complex model. Our method takes inspiration from the SMOTE algorithm [5], which is used to interpolate between datapoints of a minority class.

To ensure that perturbations fall within realistic bounds, perturbations in all feature dimensions are produced by interpolating between an instance being explained, x, and some other instance, x', in the training data. Each feature value in x is interpolated independently in the direction of the corresponding feature of x'. This ensures that the perturbed value lies within the range of appropriate values. For categorical data, the space in which we interpolate is segmented into possible feature values.

To maintain locality of the perturbations, the selection of x' is based on the proximity of each datapoint to x. Using Equation 7, the proximity of x to every other datapoint in $\mathbf{X}$ is calculated and normalised such that $P(x' = x_i) = \pi_x(x_i)$ and $\sum_{x_i \in \mathbf{X}} P(x' = x_i) = 1$.

This means that perturbations are more likely to contain values in closer proximity to x. The process for generating a set of N perturbations is outlined in Algorithm 2.

3.3. Contextually Enhanced Interpretable Local Explainable AI

We propose Contextually Enhanced Intepretable Local Explainable AI (**CHILLI**); a combination of the methods outlined in Sections 3.1 & 3.2. The goal of CHILLI is to satisfy potential contextual constraints and consider limitations of numerical data when fitting proxy models that are being used as explanations. Through this approach, explanations may be deemed more trustworthy since the underlying intuition is sound, and an explanation may be fit to perturbation data that is representative of the data used to train the complex base model. The method for generating contextually appropriate perturbations proposed in Section 3.2 is dependent on contextually appropriate proximity measures, such as those discussed in Section 3.1. These proximity measures are also used when quantifying the performance of an explanation to select the best proxy model, g, from the set of possible proxy models, G.

4. Evaluation Methodology

We compare the functionality and performance of our proposed novel method, CHILLI, with that of LIME [28] to explore the effect of incorporating context into a perturbation-based XAI framework used to explain

Algorithm 2 Contextual Perturbation Generation

Input: Number of perturbations to generate N; Data instance to perturb x; Training dataset $\mathbf{X}$

Output: Set of perturbations $\mathcal{Z}$

F = Features of x
Initialise empty set of perturbations, $\mathcal{Z} = [\ ]$
Calculate $\pi_x(x^i)$ for each $x^i \in \mathbf{X}$
Assign a probability to each x^i where $P(x' = x^i) = \frac{\pi_x(x^i)}{max(\pi_x(x^i) \forall x^i \in \mathbf{X})}$
while $|\mathcal{Z}| \leq N$ **do**
    Uniformly select some value between 0 and 1 $\rightarrow I$
    Select some $x^i \in \mathbf{X}$ based on probability for each $x^i \rightarrow x'$
    for each f in F **do**
        f^z = $f^x + I(x'_f - x_f)$
    end for each
end while

a black-box model making predictions on numerical data. We develop a non-inherently interpretable base predictive model to perform a regression task and make predictions on a test set of a larger dataset. From these predictions, we randomly select a set of instances for which we form explanations to understand the underlying behaviour of the model. We do this using both CHILLI and LIME, which produce explanations in the same format. For each instance being explained, an explanation is a set of linear coefficients for a linear regression proxy model which is fit over the perturbations of the instance. A sample explanation for a model which has been fit to data with 5 features is shown in Figure 4. The proxy model in this case has 5 feature coefficients, each corresponding to the respective contribution of the feature towards the prediction. Coefficients of larger magnitude signify a greater contribution. A positive contribution indicates a positive correlation between the feature and the target variable while a negative contribution indicates the opposite.

A user may consider the explanation and their prior knowledge to make a decision on whether the model is behaving as expected. For example, in a medical diagnosis setting a doctor may use the explanation to decide whether the correct factors are being considered when making a diagnosis. If the correct factors are being considered according to their prior knowledge, they may come to a reasonable conclusion that the model is behaving as expected. On the other hand, if the explanation shows that irrelevant features, such as the patient's name or ID, are being prioritised by the model in its decision-making, the doctor can infer that the model is not to be trusted.

This paper is not focused on the underlying behaviour described by an explanation, but rather on the reliability and accuracy of the explanation itself, since if an explanation is not in itself trustworthy, then it cannot be used to evaluate the trustworthiness of the model it is explaining.

We evaluate the performance of explanations through empirical measures, alongside manually observing the operation of the explanation generation framework, to monitor whether the behaviour aligns with our domain knowledge. As discussed in Section 2.2, the accuracy of an explanation can be measured through its faithfulness

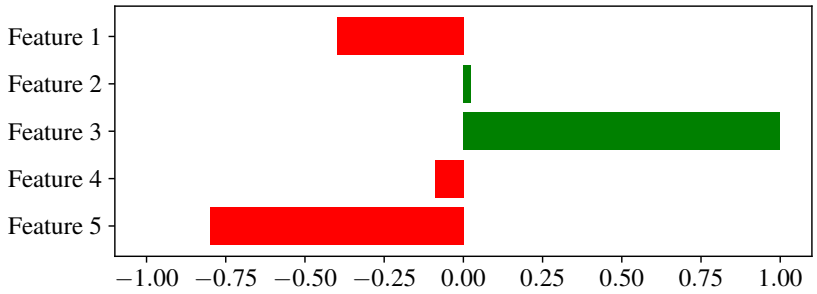


Figure 4: An example explanation depicting feature contributions obtained from the coefficients of a linear regression proxy model. Features with a red bar indicate a negative correlation with the target variable while a green bar indicates a positive correlation. The length of the bar indicates the strength of the correlation. A stronger correlation indicates a more prominent contribution towards a base model prediction.

to the base model, which may be quantified using an error metric. A smaller error indicates a more faithful explanation as discussed in Section 2.2.

Alongside this, we examine the perturbations produced by both LIME and CHILLI, of data instances being explained. Ultimately, these perturbations are the core component in the formation of the explanation since the proxy model is fit to them. Therefore, ensuring that the perturbations are representative of the data the base model was trained on is crucial.

4.1. Datasets

WebTRIS and MIDAS are both datasets containing quantitative numerical data for which we develop a black-box regression model to make a prediction on a target variable. WebTRIS [24] is a data series recorded by Highways England which collects traffic data at many motorway sites around England. We restrict the dataset to data collected at two sites (M60/9094A & M6/7570A) and between 01/01/2016 and 01/01/2017. Each site is modelled individually to monitor performance consistency across the dataset. The data for each site includes several features, a subset of which are used to predict the volume of traffic flow in a given 15 minute interval. The selected features, along with their descriptions, are listed in Table I. Features 1-6 will be input variables to the developed model, where features 1-5 are taken directly from the dataset and feature 6 is inferred from the 'Report Date' feature in the dataset. Feature 7 is the target variable. also taken directly from the dataset. The dataset for each site is split into a training and test dataset containing 80% and 20% of the data respectively.

MIDAS [25] is a dataset published by the UK Meteorological Office recording hourly weather observations at multiple land surface stations across the UK. We use data from a station located at Keswick and 3 neighbouring stations (St. Bees Head, Shap and Warcop Firing Range) to predict 'Air Temperature' at Keswick at a given time. It is expected that observations from surrounding areas will be related to the upcoming weather at Keswick, and therefore the data used from neighbouring stations is offset by 1 hour e.g., the Shap wind speed at 9:00 is recorded alongside the 10:00 Keswick air temperature. A map of the locations of these stations is shown in Figure 5 and the features used for training are listed in Table II.

	Feature	Description
1	Time Interval	The time interval as an integer where 00:00-00:15 is 0 and 23:45-00:00 is 95
2	0 - 520 cm	The number of vehicles of length between 0 and 520cm which passed this site in the time interval
4	521 - 660 cm	The number of vehicles of length between 521 and 660cm which passed this site in the time interval
3	661 - 1160 cm	The number of vehicles of length between 661 and 1160cm which passed this site in the time interval
5	Avg mph	Average speed (mph) of vehicles passing this site in the time interval
6	Day	The integer representation of the day of the week where Monday is 0 and Sunday is 6
7	Total Volume	The total number of vehicles which passed this site in the time interval

TABLE I: Features used from the WebTRIS dataset. Features 1-5 are raw values taken from the dataset, feature 6 is inferred from the 'Report Date' feature in the WebTRIS data, and feature 7 is the target variable.

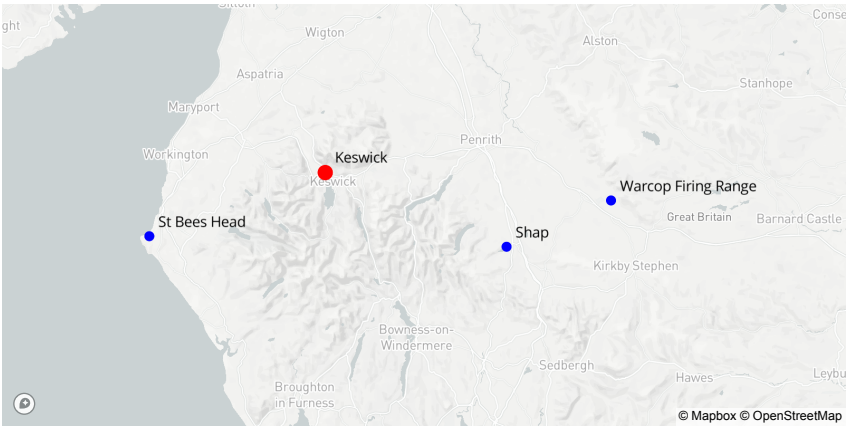


Figure 5: A terrain map of the Cumbria (UK) area showing the geographical locations of Keswick, Shap, Warcop Firing Range and St Bees Head.

4.2. Predictive Models

For the WebTRIS data, a model in the form of a Support Vector Regressor (SVR) is fit to the training dataset to predict the 'Total Volume'. The raw values from the test data for the M6/7570A/9094A site are shown in blue in Figure 6. The distribution of each feature against the target variable is shown in an individual subplot. The trained SVR is used to make predictions on the test data, with the predicted values shown in pink, and its performance being quantified using Root Mean Squared Error (RMSE). Although our focus is not on the predictive performance of the base model, the distribution of the data and predictions may provide some preliminary insight into what should be expected from the explanations. It can be observed that some features are more linearly correlated with the target variable with the target variable than others across the entire distribution of the feature, such as '0-520cm'. We expect these features to have a strong linear coefficient in all localities. Other features, such as '521-660cm', may exhibit a strong linear relationship in some areas but

	Feature	Description
1	Keswick Wind Speed	Average estimated wind speed at Keswick measured in knots
2	Keswick Wind Direction	Average wind direction at Keswick measured in true degrees
3	Keswick Total Cloud Cover	Observed cloud cover at Keswick measured in Okta
4	Keswick Cloud Base Height	Average cloud base height at Keswick measured in decametres
5	Keswick Visibility	Average visibility at Keswick measured in decametres
6	Keswick MSL Pressure	Average mean sea level pressure at Keswick measured in hectaPascals
7	Keswick Dewpoint	Dewpoint temperature at Keswick measured in °C
8	Keswick Relative Humidity	Calculated relative humidity at Keswick as a percentage
9	Keswick Air Temperature	Air temperature at Keswick measured in °C
10	St Bees Head Wind Speed	Average estimated wind speed at St Bees Head measured in knots
11	St Bees Head Wind Direction	Average wind direction at St Bees Head measured in true degrees
12	St Bees Head MSL Pressure	Average mean sea level pressure at St Bees Head measured in hectaPascals
13	St Bees Head Dewpoint	Dewpoint temperature at St Bees Head measured in °C
14	St Bees Head Relative Humidity	Calculated relative humidity at St Bees Head as a percentage
15	Warcop Range Wind Speed	Average estimated wind speed at Warcop Range measured in knots
16	Warcop Range Wind Direction	Average wind direction at Warcop Range measured in true degrees
17	Warcop Range MSL Pressure	Average mean sea level pressure at Warcop Range measured in hectaPascals
18	Warcop Range Dewpoint	Dewpoint temperature at Warcop Range measured in °C
19	Warcop Range Relative Humidity	Calculated relative humidity at Warcop Range as a percentage
20	Shap Wind Speed	Average estimated wind speed at Shap measured in knots
21	Shap Wind Direction	Average wind direction at Shap measured in true degrees
22	Shap MSL Pressure	Average mean sea level pressure at Shap measured in hectaPascals
23	Shap Dewpoint	Dewpoint temperature at Shap measured in °C
24	Shap Relative Humidity	Calculated relative humidity at Shap as a percentage
25	Date	Date and time of recorded observations measured to the hour

TABLE II: Features used from the MIDAS dataset, where feature 9 is the target variable.

not in others depending on the locality of the instance.

A Recurrent Neural Network (RNN) is trained using a subset of the MIDAS data to predict the 'Keswick Air Temperature'. Figure 7 shows the distribution of the test dataset against the target variable, along with the predictions from the RNN for each data instance. The performance in this case is quantified using MAE since the error in individual predictions will generally be smaller than 1. We can hypothesise about expected feature importance due to the linearity of the feature relationships against the target variable. Since LIME scales perturbations according to the covariance for each feature against the target variable, we explore the effect of removing generally linear features (features 6,7,8,13,14,18,19,23 & 24 in Table II) on the local explanations.

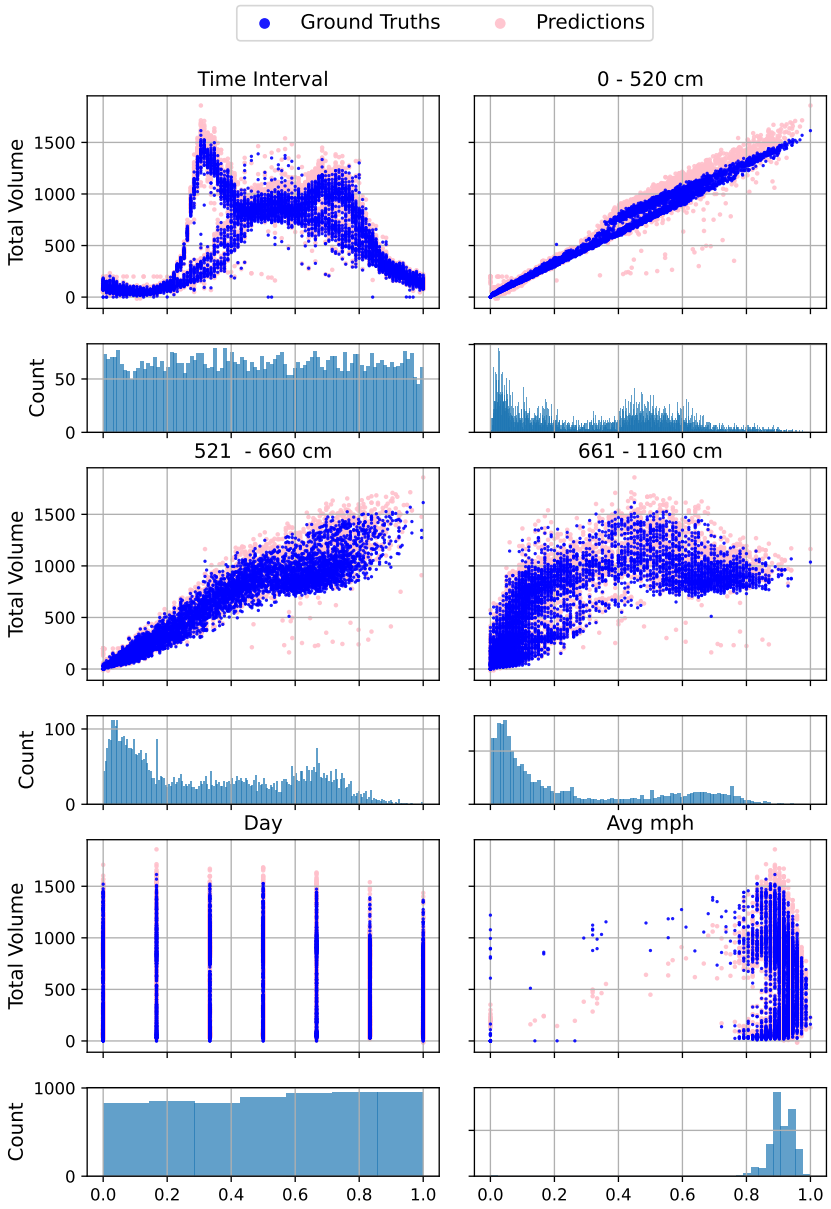


Figure 6: The predictions made by a SVR on the test WebTRIS data from site M6/7570A shown in individual feature dimensions. Each instance in the test dataset contains a value for each feature which is shown in the individual subplots. The value of the feature is shown against the value of the target variable, 'Total Volume', for that instance. The histogram of values for each feature is also shown.

Using prior knowledge regarding weather patterns, it is expected that there will be some geospatial relationship between the features from neighbouring stations, such as wind direction, and the 'Keswick Air Temperature'. An appropriate model should consider this, with its corresponding explanation highlighting the use of these features by a model when making a prediction.

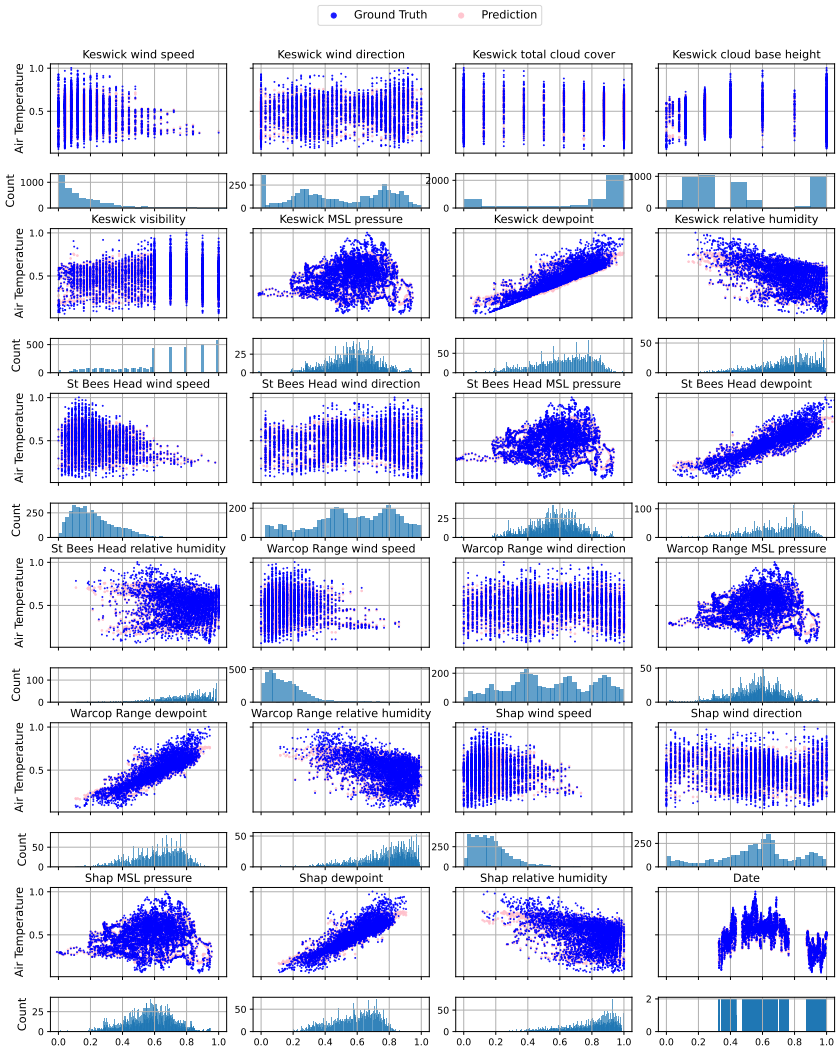


Figure 7: The predictions made by a RNN on the test MIDAS data shown in individual feature dimensions. Each instance in the test dataset contains a value for each feature which is shown in individual subplots. The value of the feature is shown against the value of the target variable, 'Air Temperature', for that instance. The histogram of values for each feature is also shown.

4.3. Forming Explanations

Random instances from the test dataset are selected, for which the prediction from a trained complex black-box model are explained using CHILLI and LIME, with the performance of both techniques then compared. Given that both CHILLI & LIME generate local explanations, it is important to vary the instance being explained so as to compare performance across the model space and in different local regions. Therefore, explanations are generated for a number of instances, selected uniformly at random, to ensure fluctuations in performance are sufficiently evaluated. Explanations generated by each method for the same instances are then compared empirically by inspecting their error. It is also important to confirm that the intuition used to generate an explanation is sound. Therefore, we observe the perturbations generated by each framework and assess whether they are suitable for fitting a trustworthy explanation.

As well as varying the regions of the model space considered, we also explore the effect of varying the locality parameter, σ, on the explanation performance for each method. As discussed in Section 3.1, this parameter affects the weighting assigned to a data point based on its distance to the instance being explained, particularly when fitting an explanation to the perturbations. It is expected that varying σ will lead to a variation in the generality of the explanation produced with respect to the model space.

5. Results

Here we present the results of using LIME & CHILLI to fit a local proxy model which is used to explain a prediction produced by a base black-box model for a given instance. The performance of each explanation is quantified by its error, which represents the faithfulness of an explanation towards the underlying behaviour of the base model being explained.

5.1. WebTRIS

Figure 8a shows an explanation produced by LIME for a prediction made by the SVR base model, f, for a randomly selected instance, x, from the WebTRIS M6/7570A test dataset, $\mathcal{X}$. The instance is shown as the red point in each of its feature dimensions against the target 'Total Volume' variable in the subplots of Figure 8b. The perturbations of the instance generated by LIME are shown as green points in the same subplots. The 'Total Volume' of each perturbation, $f(z)$, was predicted by the base model and is represented on the vertical axis. Notice that the feature values of some perturbations do not fall within the appropriate normalised ranges of 0 - 1, with many perturbations resulting in the base model predicting a negative value of 'Total Volume'. The opacity of each perturbation signifies its calculated proximity weighting, $\pi_x(z)$, to the instance being explained.

The linear proxy model which is best fit to the set of perturbations, $\mathcal{Z}$, and base model predictions, $f(\mathcal{Z})$, was selected as the explanation. The set of perturbations with predicted 'Total Volume' by the explanation model, $g(\mathcal{Z})$, are shown as orange points. Again, the opacity of the points indicate the proximity, $\pi_x(z)$, to the instance. The RMSE of the explanation predictions, $g(\mathcal{Z})$, against the base models predictions, $f(\mathcal{Z})$, was 36.699. The explanation was used to make a prediction for the randomly selected instance, x, where a completely faithful explainer will always predict the same value as the base model i.e., $g(x) = f(x)$. For the randomly selected instance shown in Figure 8b, the predicted value for x from the base model and explanation were 71 and 92 respectively, with the true value being 69. Similar plots, created using CHILLI instead of LIME to generate an explanation for the same instance, are shown in Figure 9. The prediction made by the explanation produced by CHILLI on the same randomly selected instance, x, in Figure 8, was 83 with the RMSE of the explanation across all perturbations being 13.21. The RMSE of the explanation produced by CHILLI was significantly lower than that of LIME, with the prediction for x made by the explanation being closer to that of the base model.

Figure 10 shows a comparison of the empirical performance of explanations produced by LIME and CHILLI over 25 instances, selected uniformly at random, in both the M6/7570A and M60/9094A test datasets. Figure 10a compares the RMSE achieved by the explanation produced by both methods for each evaluated instance where Figure 10b shows the average RMSE achieved over the 25 instances at each site. From Figure 10a it can be seen that, for both sites, CHILLI outperforms LIME by achieving lower RMSE in every instance. It can also be seen that there is greater consistency in the RMSE achieved by CHILLI across the instances. Explanations generated by LIME exhibited an RMSE variance of 41.6 and 73.2 for the two sites, while explanations generated

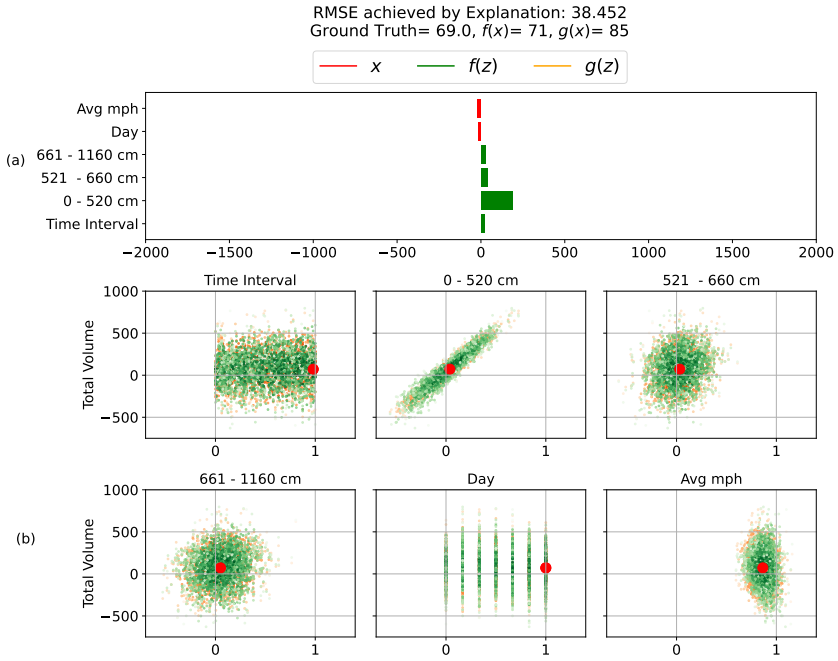


Figure 8: (a) Explanation and (b) perturbations produced by LIME when explaining the prediction from the SVR on the WebTRIS M6/7570A instance shown in red. Each subplot represents a single feature dimension of the perturbations. Predictions of the target variable from both the base model and proxy model used for the explanation are shown in green and orange respectively.

by CHILLI had a variance of 18.4 and 8.9. Looking at the average RMSE achieved by both methods in Figure 10b, CHILLI leads to a reduction in RMSE of 69% and 75% on the M60/9094A and M6/7570A datasets respectively.

5.2. MIDAS

Figure 11 shows the MIDAS data with a uniformly randomly selected instance being explained shown as the red point. Figures 12 and 13 show the perturbations of the instance produced by LIME and CHILLI respectively, in a similar format to Figures 8b & 9b. Again, each subplot represents a single feature dimension against the target variable, in this case 'Air Temperature'. Feature values of some perturbations generated by LIME fall outside the normalised ranges whereas those produced by CHILLI do not. The explanations produced by CHILLI and LIME are shown in Figures 15 and 14 respectively. The explanation produced by CHILLI exhibits a lower MAE than LIME, with the explanations from the respective methods achieving an MAE of 0.154 and 0.557 on their respective generated perturbations.

To test the fitting of explanations to general data trends rather than local trends, we consider the variation in contribution of all features across 25 explanations in areas of the data space selected uniformly at random, since this is expected to utilise different regions of the model space, thus enacting various model behaviours. The random instances are shown as navy points in Figure 11. Figure 16 shows the range of values, along with the median and quartile ranges, attributed to each feature across the explanations for the selected instances. It is clear that there is little variation in the explanations produced by LIME compared to those produced by CHILLI.

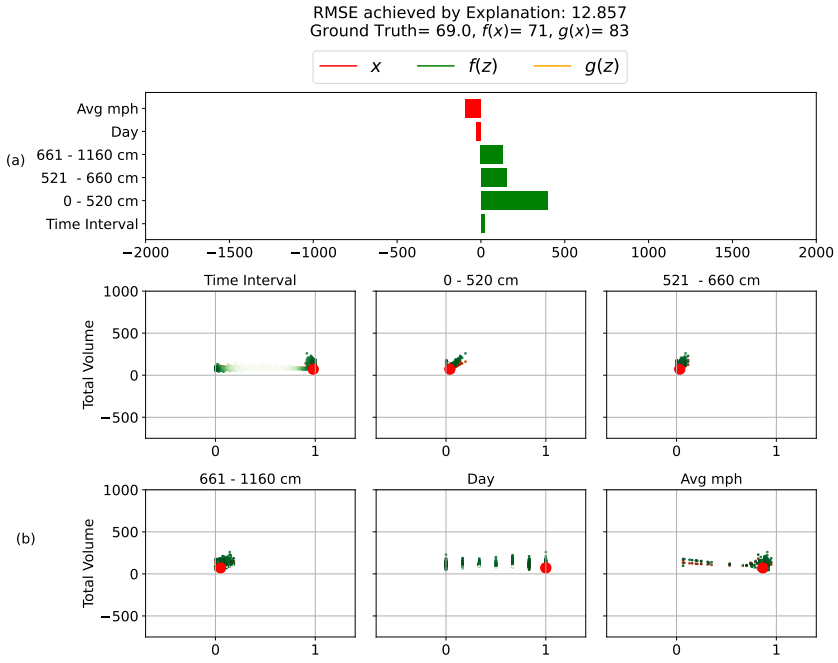


Figure 9: (a) Explanation and (b) perturbations produced by CHILLI when explaining the prediction from the SVR on the WebTRIS M6/7570A instance shown in red. Each subplot represents a single feature dimension of the perturbations. Predictions of the target variable from both the base model and proxy model used for the explanation are shown in green and orange respectively.

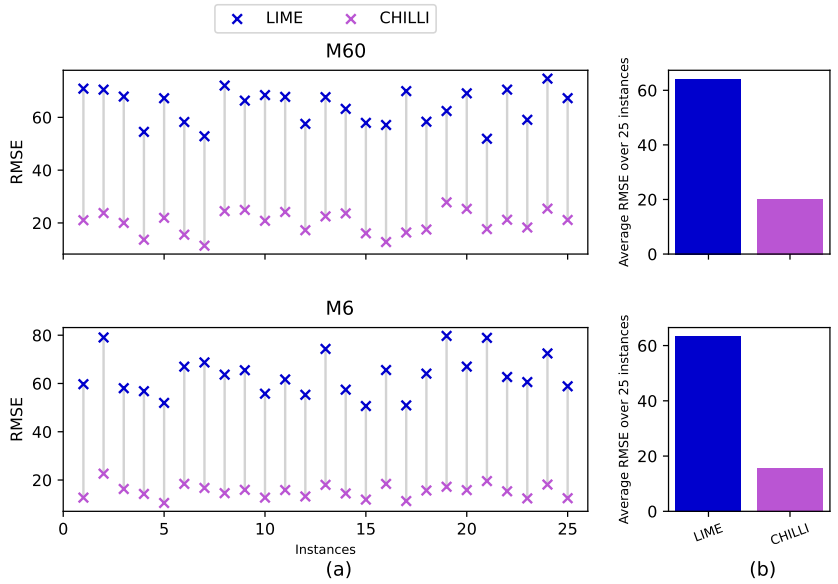


Figure 10: (a) RMSE achieved by explanations for 25 randomly selected instances for both the original LIME framework and CHILLI at both WebTRIS data sites M60/9094A and M6/7570A (b) Average RMSE achieved by the original LIME framework and CHILLI over the 25 instances shown in (a).

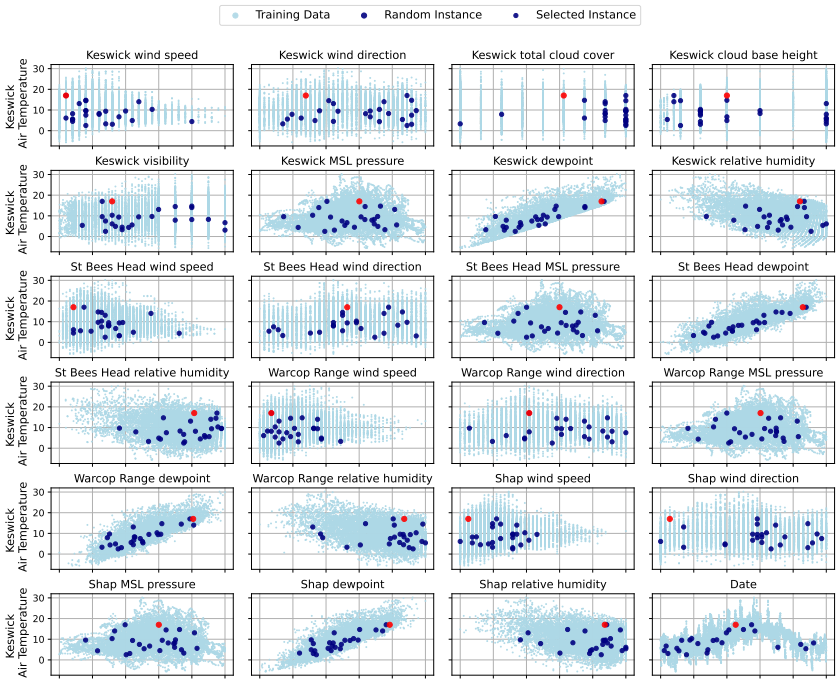


Figure 11: Distribution of the MIDAS training data shown in individual feature dimensions against the target variable 'Air Temperature'. Random instances for which explanations will be generated are shown in navy. A single instance whose explanation will be evaluated in further detail is shown in red.

Figure 17a shows a comparison of the MAE achieved by the explanations produced by LIME and CHILLI for each of the 25 instances shown in Figure 11. It can be seen that the MAE achieved by the explanations produced by CHILLI was consistently lower than those produced by LIME. There was a 58% average reduction in MAE when using CHILLI over LIME.

After removing the generally linear features in the MIDAS dataset, explanations were produced for the same 25 uniformly randomly selected instances shown in Figure 11. The range of values attributed to each feature in the 25 explanations is shown in Figure 18. It can be noticed that there is a greater variation in the explanations produced by CHILLI compared to those produced by LIME. It can also be noticed that there is greater variation in the contributions of some features after removing the generally linear features. Figure 19a shows a comparison of the MAE achieved by the explanations produced by LIME and CHILLI for each of the 25 instances shown in Figure 11 where Figure 19b shows the average MAE achieved across the 25 instances. It can be seen that explanations prouced by CHILLI consistently achieved a lower MAE than those produced by LIME with an 87% average reduction in MAE when using CHILLI over LIME.

As mentioned in Section 3.1, the value of the locality parameter, σ, is used to scale the proximity of points in the data space. The importance of locality on the performance of the explanation can be understood by observing the effect of varying σ on MAE. Figure 20 shows a comparison of the MAE achieved by both LIME and CHILLI when explaining the selected instance in Figure 11 with different values of σ.

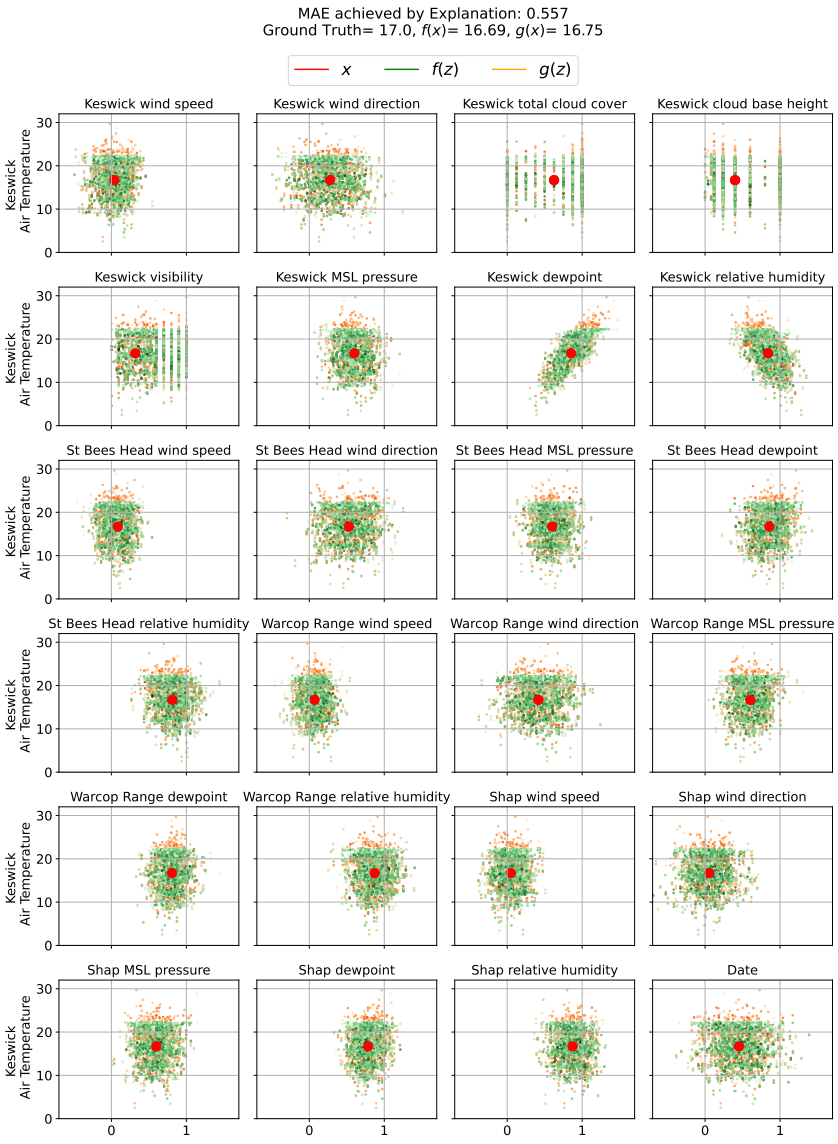


Figure 12: Perturbations produced by LIME when explaining the prediction from the RNN on the MIDAS instance shown in red. Each subplot represents a single feature dimension of the perturbations. Predictions of the target variable from both the base model and proxy model used for the explanation are shown in green and orange respectively.

6. Discussion

6.1. Perturbation Generation

From a visual inspection of Figures 8, 9, 12 & 13, it is clear that the two methods produce quite contrasting perturbations of the instances being explained. It can be seen that the perturbations of all features generated by LIME, do not follow the data distribution shown in Figures 6 & 11. Moreover, since each feature is perturbed independently, feature values in a single perturbation do not take into consideration existing dependencies, as discussed in Section 3.2, which leads to unrealistic perturbations being generated. For example in the WebTRIS data, a perturbation may exist where the 'Time Interval' is 03:00, however the value of '0-520cm'

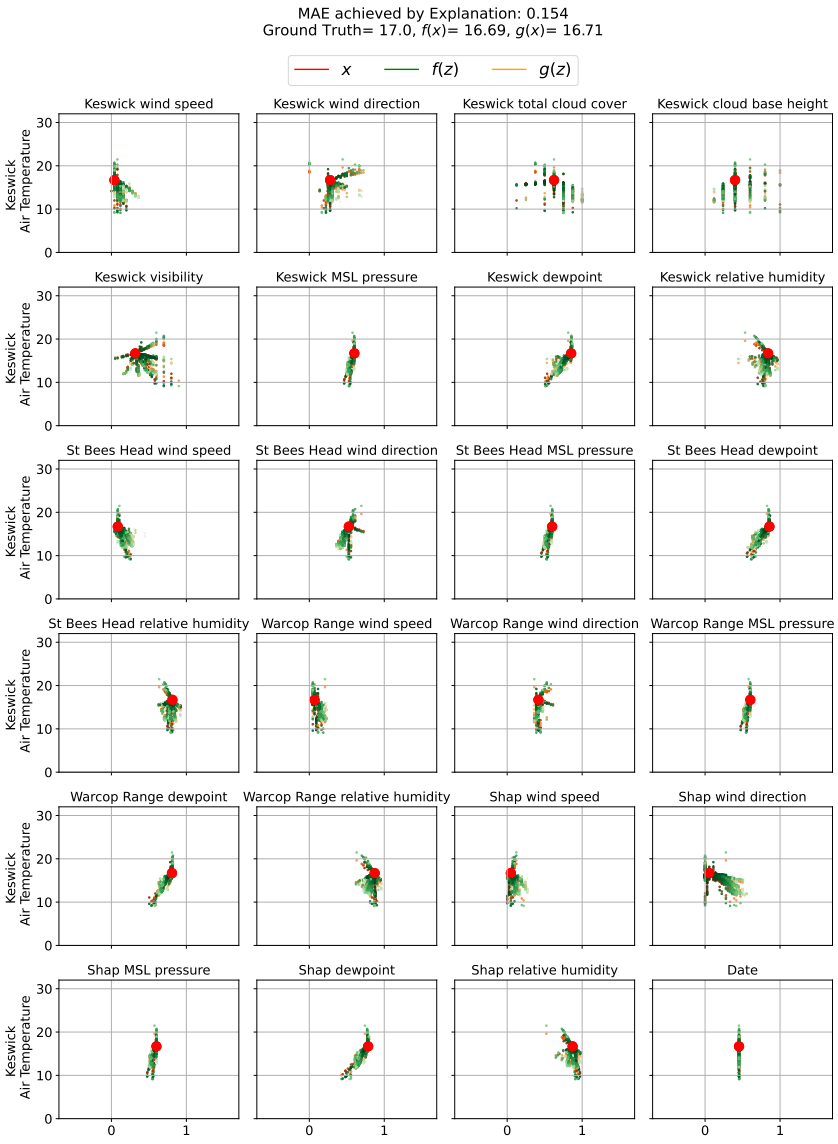


Figure 13: Perturbations produced by CHILLI when explaining the prediction from the RNN on the MIDAS instance shown in red. Each subplot represents a single feature dimension of the perturbations. Predictions of the target variable from both the base model and proxy model used for the explanation are shown in green and orange respectively.

may correspond to a number of vehicles that would be observed at rush hour.

In addition to the lack of consideration of feature dependency, the possible bounds of features have not been considered either. By simply sampling from a Normal distribution around the mean for non-categorical features, there is no prior contextual knowledge regarding the possible range of values a feature can take. It can be seen from Figures 8 & 12 that all non-categorical features exhibit perturbed values which fall outside the normalised range of 0-1. In the WebTRIS instance perturbations, negative values of '0-520cm', '521-660cm' & '661-1160cm' imply a negative number of vehicles of the respective sizes passing in the corresponding time interval, which of course is not possible. This is also present in the perturbations of the MIDAS data.

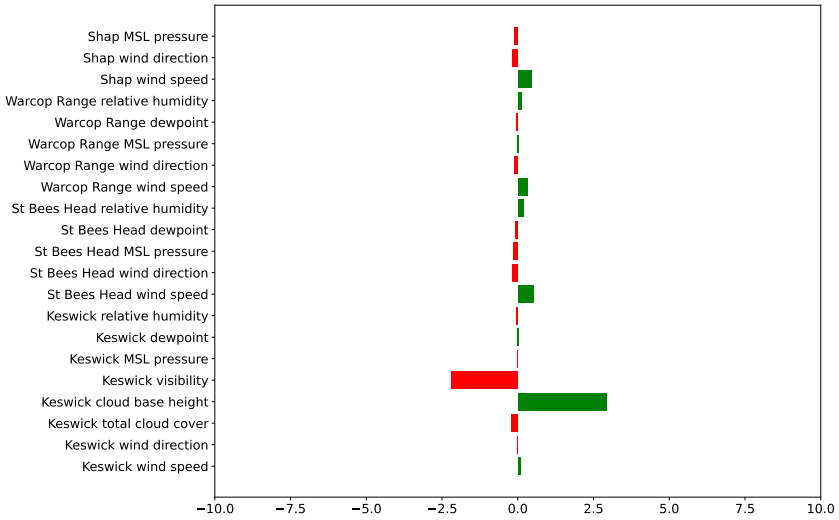


Figure 14: Explanation produced by LIME when explaining the prediction from the RNN on the MIDAS instance shown in red in Figure 11.

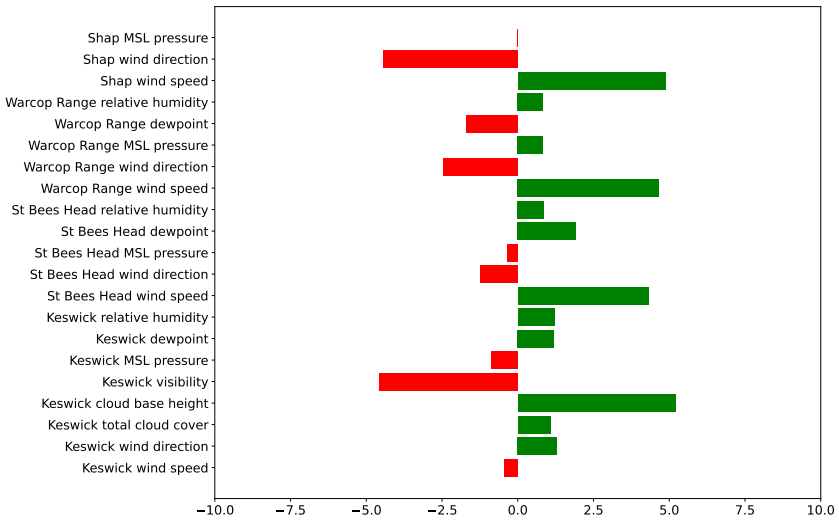


Figure 15: Explanation produced by CHILLI when explaining the prediction from the RNN on the MIDAS instance shown in red in Figure 11.

Relative humidity can take a maximum value of 1, indicating the air is fully saturated with moisture, however perturbations are generated that exceed this value. This leads to a set of perturbations that do not represent real world data, and therefore do not represent the training data. The negative impact of such inappropriate perturbations can be observed from the predicted values from the base model, which often predicts a negative volume of traffic flow which is not possible. This further reinforces the lack of conformity of the perturbations to the data used to train the base model, since all training data have positive 'Total Volume'. An explanation that is then fit on these inappropriate perturbations will not correctly represent the true behaviour of the complex base model. Therefore, in this case, explanations formed using LIME will not be trustworthy.

Upon further visual inspection of the opacities of the perturbations, it can be noticed that perturbations which

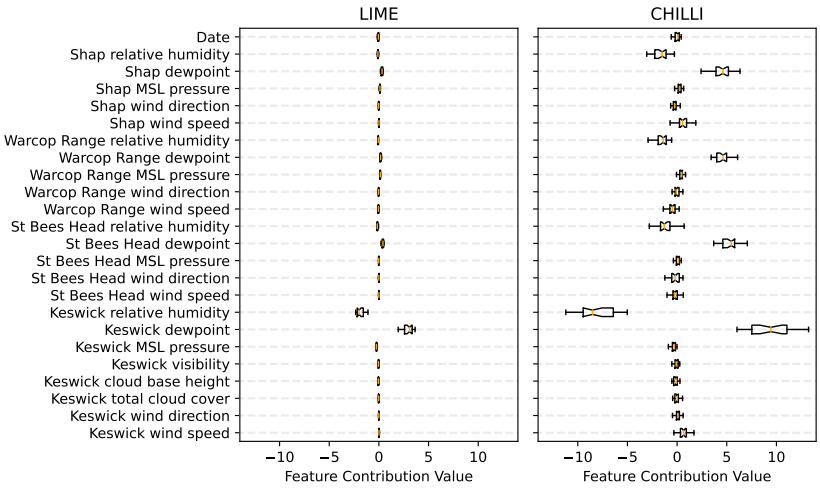


Figure 16: Variation in feature contributions presented in explanations produced by LIME and CHILLI across the 25 instances shown in Figure 11. The median value and quartile ranges from the set of contribution values are shown for each feature.

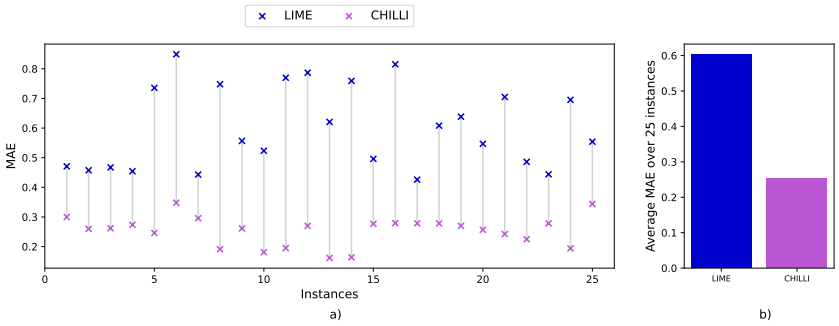


Figure 17: (a) MAE achieved by explanations for the 25 random instances shown in Figure 11, for both the original LIME framework and CHILLI (b) Average MAE achieved by the original LIME framework and CHILLI over the 25 instances shown in Figure 11.

are further from the instance are sometimes assigned a higher proximity weighting than perturbations which are closer to the instance. Given that this indicates the contribution of each perturbation to the selection of the best fit LR model, the produced explanation will not be focused on the immediate locality around the instance, and is instead a generalised explanation across all the perturbations.

On the other hand, if we consider the distribution of perturbations produced by CHILLI, shown in Figure 9, it can be observed that they much more closely follow the distribution of the real WebTRIS data shown in Figure 6, particularly in the locality of the instance being explained. The CHILLI algorithm generates perturbations using a sampling technique that considers the data the base model was trained on and any existing dependencies. Therefore, the produced perturbations not only conform to the distribution of the training data, but they are also realistic combinations of feature values that fall within the appropriate feature bounds. CHILLI also generates perturbations with greater density around the instance being explained, which can be seen from the increase

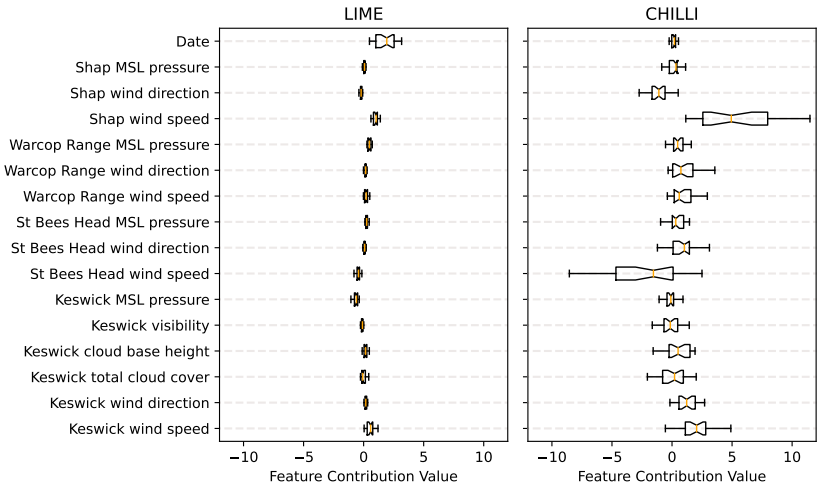


Figure 18: Explanation produced by LIME when explaining the prediction from the RNN on the MIDAS instance shown in red in Figure 11 when features 6,7,8,13,14,18,19,23 & 24 from Table II are excluded.

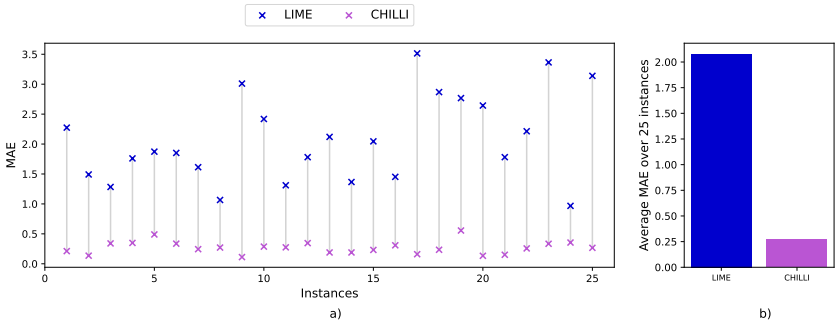


Figure 19: (a) MAE achieved by explanations for the 25 random instances shown in Figure 11, for both the original LIME framework and CHILLI (b) Average MAE achieved by the original LIME framework and novel method over the 25 instances shown in Figure 11 when features 6,7,8,13,14,18,19,23 & 24 from Table II are excluded.

in concentration of green points around the instance, which decreases as distance from the instance increases. The proximity of perturbations, indicated by the opacity, is clearly more suitable since there is a concentration of darker points around the instance being explained. This is also the case for features of a cyclic nature, such as 'Time Interval' in WebTRIS, where 0 and 1 are adjacent values. This leads to an explanation that is fit with greater emphasis on perturbations of closer locality.

6.2. Explanation Performance

Shifting focus from the perturbations to the explanations shown in Figures 8a, 9a, 14 & 15, it can be observed that CHILLI produces an explanation with greater disparity between feature coefficients. A coefficient of large magnitude indicates a strong linear relationship in the perturbed values of that feature. It is expected for explanations produced by CHILLI to have larger feature coefficients than LIME since the perturbations generated by LIME are based on a Normal Distribution which naturally does not exhibit any linear correlation.

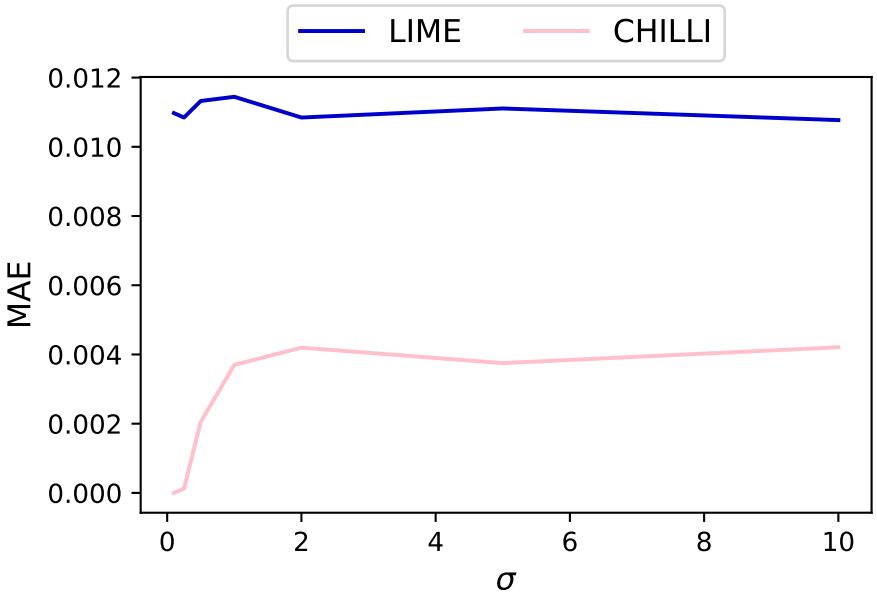


Figure 20: A comparison of the MAE achieved by explanations produced using LIME and CHILLI for a single instance of the MIDAS data, whilst varying the locality parameter, σ.

Due to the covariance scaling of the LIME perturbations towards the training data, the distribution of the perturbations is transformed to the aggregated linear correlation of the feature values in the training data. In some cases, this is effective, such as '0-520cm' which exhibits a strong and consistent linear correlation with 'Total Volume' across all values of the feature in the WebTRIS data, as can be seen in Figure 6. LIME recognises this and identifies it as the most significant feature contribution, as shown in Figure 8. Similarly, features in the MIDAS data containing 'dewpoint' and 'relative humidity' have a linear correlation with 'Keswick Air Temperature' across their range of values, as can be seen from 7. As a result, the covariance scaling applied by LIME transforms these feature perturbations in a way which conveys some linear relationship between the perturbed feature values and the target variable. For other features, where there is not such a prominent general linear correlation, LIME fits much smaller coefficients. This is clearly unsuitable since general feature trends are not relevant in an explanation framework focused on locality. As discussed earlier, explanations produced by CHILLI are fit to perturbations that are concentrated on the data distribution within close proximity to the instance being explained. Therefore, feature contributions are more representative of the base models behaviour in the instance locality rather than being more general.

CHILLI also emphasises the contribution of generally linear features in its explanation, which is understandable since it is expected that the most significant contributing features will consistently be those with a strong linear relationship across all instances. Therefore, the base model is expected to use these features consistently across the entire data space irrespective of the locality of the instance being explained. To some extent, this makes it appear that LIME manages to correctly encapsulate the dominating feature contributions in the locality, while in actuality it is fitting an explanation to the full data space.

As noted in Section 2.2, a lower error indicates that an explanation is more faithful to the base model

behaviour. The consistent reduction in error of the explanations produced by CHILLI, compared to LIME, supports the fact that CHILLI produces a more faithful explanation and is more representative of the base model's true behaviour. The lower variance in error for explanations produced by CHILLI indicates that it focuses on tuning each explanation to the locality of the instance being explained, rather than fitting a general explanation across the entire data space.

Observing the feature contribution variance in 16, only overall linear features noticeably deviate from a value of 0 in explanations produced by LIME. This implies that LIME only considers these features in its explanations and can be attributed to their general linear trend. Comparing this to CHILLI, there is much more variation in the contribution values which indicates a variation in explanation based on the locality of the instance being explained. This is expected since the shape of the perturbations each explanation is fit differs based on the locality of the instance being explained. However, it can still be observed that most feature contributions are close to 0, except those that are overall linear since these relationships outweigh any other less correlated features.

6.3. Removing Linear Features

To further explore the impact of features which exhibit generally linear behaviour i.e., those describing dew-point and relative humidity, we remove them from the dataset and form new explanations. Again there is much greater variation in the explanations produced by CHILLI, as can be noticed in Figure 18, which reinforces the focus CHILLI has on instance locality. Where LIME generates explanations based on general linear trends in the overall data, removing features with overall linear correlation leads to LIME not being able to establish any significant trends, which can also be seen in Figure 18 where all contributions are close to 0.

Figure 19 shows a comparison of the MAE obtained by explanations generated using LIME and CHILLI when overall linear features are omitted. Whilst focusing only on local trends in the data, CHILLI substantially outperforms LIME in every instance. Since LIME cannot detect local model behaviour due to its insufficient perturbation generation method, it performs poorly when locality is of importance. CHILLI, on the other hand, is able to detect local trends in the data and therefore achieves a significantly lower average MAE. This is expected since CHILLI is able to generate perturbations that are more representative of the local data environment.

6.4. Locality Hyperparameter Exploration

When varying the locality hyperparameter, σ, the MAE achieved by LIME is generally consistent across all values of σ, since LIME has been shown to form general explanations that do not consider the local data environment of the instance being explained. CHILLI, on the other hand, achieves lower MAE for lower values of σ before it stabilises at higher values of σ. As the defined locality increases, perturbations begin to be considered that are further in proximity from the instance being explained. Therefore, features which do not exhibit linear relationships on a broader scale are difficult to describe using a linear proxy model. This leads to an explanation that does not perform as well since it attempts to generalise behaviour, such as is the case for LIME, rather than explaining local trends. Even as MAE stabilises when using CHILLI, it still outperforms LIME since the underlying intuition regarding perturbation generation is sound, while, as discussed previously, LIME does not base explanations on realistic perturbation data.

7. Conclusion and Future Work

In this paper, we explored the effect of incorporating prior contextual knowledge of the data domain into an existing model-agnostic local perturbation-based XAI approach, namely LIME, when explaining black-box models used to make predictions on tabular data. We proposed a novel method for implementing contextually appropriate proximity measures, which consider the characteristics of individual features, to ensure locality is accurately defined and constrained. We also proposed a novel method for generating perturbations that consider the contextual limitations of data features, as well as ensuring that the perturbations used to fit an explanation are representative of the data used to train the base model which is being explained. Combining these methods, we propose a new framework, CHILLI, for generating local explanations for black-box ML models which appropriately considers contextual limitations of the features and data used. We compared the functionality and performance of CHILLI with LIME, an XAI technique held in high-regard due to its model-agnostic approach. We tested both approaches on the WebTRIS and MIDAS datasets and explored the difference in performance when features with overall linear correlation are removed.

It was found that LIME does not appropriately measure proximity between data points, due to its lack of contextual awareness regarding features being used. This resulted in poor definition of locality around a data instance being explained. Due to this, proxy models generated by LIME were found to identify only general trends in the data, thus producing a generalised explanation. For features without overall linear trends, this resulted in insignificant feature contribution on a general scale whilst ignoring local impact. It was also found that LIME does not generate perturbations in a way which is representative of the training data used. Not only were the distributions not representative, but the perturbations used to fit the explainer contained unrealistic values that would never occur in reality.

When using CHILLI to generate explanations, it was found that the perturbations used to fit the proxy models were representative of the data used to train the black-box model and were of local proximity to the instance being explained. Therefore, CHILLI's explanations had relatively larger feature contributions as opposed to those produced by LIME.

We also compared the empirical performance of an explanation by calculating an error metric when comparing a proxy model's predictions with that of the black-box model. This error is representative of the faithfulness of an explanation towards the model it is explaining. CHILLI consistently outperformed LIME, achieving a lower error across all explained instances. Explanations produced by CHILLI, when compared with those from LIME, reduced RMSE by an average of 72% across 25 instances selected uniformly at random from the WebTRIS data. Similarly, 25 randomly selected instances being explained in the MIDAS data exhibited an average reduction in MAE of 49%. When features with overall linear trends were removed, the difference between the two methods was more significant, with CHILLI resulting in an average 96% reduction in MAE when explaining 25 random instances from the MIDAS data.

Through both empirical and intuitive evaluation of LIME and the proposed novel method, CHILLI, we can conclude that incorporating contextual domain knowledge regarding data features used for generating explanations improves faithfulness of the explanation towards the model it is explaining, ultimately increasing trust in both the explanation and explanation framework. This is achieved by ensuring that locality is appropriately

defined and constrained, and that perturbations used to fit the proxy model are representative of the data used to train the black-box model.

We would like to further this work by exploring how improving the performance of local explanations affects overall trust in a model. We would also like to explore the effect of our novel method when proxy models of a different form are used, such as decision trees or small-order polynomial regressors.

Acknowledgments

How to structure acknowledgements?

References

[1] Maher Ala'raj and Maysam F. Abbod. "Classifiers consensus system approach for credit scoring". en. In: *Knowledge-Based Systems* 104 (July 2016), pp. 89–105. ISSN: 09507051. DOI: 10.1016/j.knosys.2016.04.013. URL: https://linkinghub.elsevier.com/retrieve/pii/S0950705116300569 (visited on 09/06/2022).

[2] David Baehrens et al. "How to Explain Individual Classification Decisions". In: arXiv:0912.1128 (Dec. 2009). arXiv:0912.1128 [cs, stat]. URL: http://arxiv.org/abs/0912.1128.

[3] Leo Breiman. *Classification and Regression Trees*. New York: Routledge, Oct. 2017. ISBN: 978-1-315-13947-0. DOI: 10.1201/9781315139470.

[4] Ajay Byanjankar, Markku Heikkila, and Jozsef Mezei. "Predicting Credit Risk in Peer-to-Peer Lending: A Neural Network Approach". en. In: *2015 IEEE Symposium Series on Computational Intelligence*. Cape Town: IEEE, Dec. 2015, pp. 719–725. ISBN: 978-1-4799-7560-0. DOI: 10.1109/SSCI.2015.109. URL: http://ieeexplore.ieee.org/document/7376683/ (visited on 09/06/2022).

[5] N. V. Chawla et al. "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16 (June 2002). arXiv:1106.1813 [cs], pp. 321–357. ISSN: 1076-9757. DOI: 10.1613/jair.953. URL: http://arxiv.org/abs/1106.1813 (visited on 10/03/2022).

[6] Finale Doshi-Velez and Been Kim. "Towards A Rigorous Science of Interpretable Machine Learning". In: arXiv:1702.08608 (Mar. 2017). arXiv:1702.08608 [cs, stat]. URL: http://arxiv.org/abs/1702.08608.

[7] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. "Explainable artificial intelligence: A survey". In: *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. May 2018, pp. 0210–0215. DOI: 10.23919/MIPRO.2018.8400040.

[8] Leilani H. Gilpin et al. *Explaining Explanations: An Overview of Interpretability of Machine Learning*. Tech. rep. arXiv:1806.00069. arXiv:1806.00069 [cs, stat] type: article. arXiv, Feb. 2019. URL: http://arxiv.org/abs/1806.00069 (visited on 09/29/2022).

[9] Alex Goldstein et al. *Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation*. Tech. rep. arXiv:1309.6392. arXiv:1309.6392 [stat] type: article. arXiv, Mar. 2014. URL: http://arxiv.org/abs/1309.6392 (visited on 09/28/2022).

[10] Bryce Goodman and Seth Flaxman. "European Union regulations on algorithmic decision-making and a "right to explanation"". In: *AI Magazine* 38.3 (Oct. 2017). arXiv:1606.08813 [cs, stat], pp. 50–57. ISSN: 2371-9621, 0738-4602. DOI: 10.1609/aimag.v38i3.2741. URL: http://arxiv.org/abs/1606.08813 (visited on 08/03/2022).

[11] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. "Interpretable Decision Sets: A Joint Framework for Description and Prediction". en. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco California USA: ACM, Aug. 2016, pp. 1675–1684. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939874. URL: https://dl.acm.org/doi/10.1145/2939672.2939874.

[12] Thibault Laugel et al. "The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations". en. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. Macao, China: International Joint Conferences on Artificial Intelligence Organization, Aug. 2019, pp. 2801–2807.

ISBN: 978-0-9992411-4-1. DOI: 10.24963/ijcai.2019/388. URL: https://www.ijcai.org/proceedings/2019/388 (visited on 06/30/2022).

[13] H. Lieberman and T. Selker. "Out of context: Computer systems that adapt to, and learn from, context". In: *IBM Systems Journal* 39.3.4 (2000). Conference Name: IBM Systems Journal, pp. 617–632. ISSN: 0018-8670. DOI: 10.1147/sj.393.0617.

[14] Muhammad K Lodhi et al. "Predicting Hospital Re-admissions from Nursing Care Data of Hospitalized Patients". In: *Advances in data mining. Industrial Conference on Data Mining* 2017 (2017), pp. 181–193. DOI: 10.1007/978-3-319-62701-4_14. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5665368/ (visited on 09/06/2022).

[15] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. *Consistent Individualized Feature Attribution for Tree Ensembles*. arXiv:1802.03888 [cs, stat]. Mar. 2019. URL: http://arxiv.org/abs/1802.03888 (visited on 10/04/2022).

[16] Andreas Messalas, Yiannis Kanellopoulos, and Christos Makris. "Model-Agnostic Interpretability with Shapley Values". In: *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*. July 2019, pp. 1–7. DOI: 10.1109/IISA.2019.8900669.

[17] Karl Øyvind Mikalsen et al. *An Unsupervised Multivariate Time Series Kernel Approach for Identifying Patients with Surgical Site Infection from Blood Samples*. Tech. rep. arXiv:1803.07879. arXiv:1803.07879 [cs, stat] type: article. arXiv, Mar. 2018. URL: http://arxiv.org/abs/1803.07879 (visited on 09/06/2022).

[18] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems". en. In: *ACM Transactions on Interactive Intelligent Systems* 11.3–4 (Dec. 2021), pp. 1–45. ISSN: 2160-6455, 2160-6463. DOI: 10.1145/3387166.

[19] Christoph Molnar. *Interpretable Machine Learning*. en. Google-Books-ID: jBm3DwAAQBAJ. Lulu.com, 2020. ISBN: 978-0-244-76852-2.

[20] Christoph Molnar. "Interpretable Machine Learning". en. In: ().

[21] Christoph Molnar et al. *General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models*. Tech. rep. arXiv:2007.04131. arXiv:2007.04131 [cs, stat] type: article. arXiv, Aug. 2021. URL: http://arxiv.org/abs/2007.04131 (visited on 09/29/2022).

[22] Christoph Molnar et al. *Model-agnostic Feature Importance and Effects with Dependent Features – A Conditional Subgroup Approach*. Tech. rep. arXiv:2006.04628. arXiv:2006.04628 [cs, stat] type: article. arXiv, June 2021. URL: http://arxiv.org/abs/2006.04628 (visited on 09/29/2022).

[23] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL: probml.ai.

[24] National Highways. *WebTRIS*. http://webtris.highwaysengland.co.uk/. Jan. 2017.

[25] Met Office. *MIDAS Open: UK daily weather observation data, v202207*. en. 2022. DOI: 10.5285/4B44CEC2F9A846F39D5007983B7EAAAB. URL: https://catalogue.ceda.ac.uk/uuid/4b44cec2f9a846f39d5007983b7eaaab (visited on 01/03/2023).

[26] P. Jonathon Phillips et al. *Four Principles of Explainable Artificial Intelligence*. en. preprint. Aug. 2020. DOI: 10.6028/NIST.IR.8312-draft. URL: https://nvlpubs.nist.gov/nistpubs/ir/2020/NIST.IR.8312-draft.pdf (visited on 09/28/2022).

[27] Reuters. *Amazon ditched AI recruiting tool that favored men for technical jobs*. en-GB. Oct. 2018. URL: https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine (visited on 03/20/2023).

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. en. arXiv:1602.04938 [cs, stat]. Aug. 2016. URL: http://arxiv.org/abs/1602.04938 (visited on 07/05/2022).

[29] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Anchors: High-Precision Model-Agnostic Explanations". en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (Apr. 2018). ISSN: 2374-3468, 2159-5399. DOI: 10.1609/aaai.v32i1.11491. URL: https://ojs.aaai.org/index.php/AAAI/article/view/11491.

[30] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. *Model-Agnostic Interpretability of Machine Learning*. arXiv:1606.05386 [cs, stat]. June 2016. URL: http://arxiv.org/abs/1606.05386 (visited on 10/04/2022).

[31] Cristopher Rigano. "Using Artificial Intelligence to Address Criminal Justice Needs". en. In: *US NIJ Journal* 280 (Jan. 2019). www.nij.gov/journals/280/Pages/using-artificial-intelligence-to-address-criminal-justice-needs.aspx.

[32] Cynthia Rudin. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead". en. In: arXiv:1811.10154 (Sept. 2019). arXiv:1811.10154 [cs, stat]. URL: http://arxiv.org/abs/1811.10154.

[33] G.P.J. Schmitz, C. Aldrich, and F.S. Gouws. "ANN-DT: an algorithm for extraction of decision trees from artificial neural networks". In: *IEEE Transactions on Neural Networks* 10.6 (Nov. 1999). Conference Name: IEEE Transactions on Neural Networks, pp. 1392–1401. ISSN: 1941-0093. DOI: 10.1109/72.809084.

[34] T. Selker and W. Burleson. "Context-aware design and interaction in computer systems". In: *IBM Systems Journal* 39.3.4 (2000). Conference Name: IBM Systems Journal, pp. 880–891. ISSN: 0018-8670. DOI: 10.1147/sj.393.0880.

[35] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. en. arXiv:1409.1556 [cs]. Apr. 2015. URL: http://arxiv.org/abs/1409.1556 (visited on 07/05/2022).

[36] Kacper Sokol et al. *bLIMEy: Surrogate Prediction Explanations Beyond LIME*. en. arXiv:1910.13016 [cs, stat]. Oct. 2019. URL: http://arxiv.org/abs/1910.13016 (visited on 07/05/2022).

[37] Agus Sudjianto and Aijun Zhang. *Designing Inherently Interpretable Machine Learning Models*. Tech. rep. arXiv:2111.01743. arXiv:2111.01743 [cs, stat] type: article. arXiv, Nov. 2021. URL: http://arxiv.org/abs/2111.01743 (visited on 09/29/2022).

[38] Latanya Sweeney. "Discrimination in Online Ad Delivery". In: arXiv:1301.6822 (Jan. 2013). arXiv:1301.6822 [cs]. URL: http://arxiv.org/abs/1301.6822.

[39] Sahil Verma, John Dickerson, and Keegan Hines. *Counterfactual Explanations for Machine Learning: A Review*. Tech. rep. arXiv:2010.10596. arXiv:2010.10596 [cs, stat] type: article. arXiv, Oct. 2020. URL: http://arxiv.org/abs/2010.10596 (visited on 09/29/2022).

[40] Xiting Wang et al. “TopicPanorama: A Full Picture of Relevant Topics”. en. In: (), p. 14.

[41] Sandareka Wickramanayake, Wynne Hsu, and Mong Li Lee. *Towards Fully Interpretable Deep Neural Networks: Are We There Yet?* Tech. rep. arXiv:2106.13164. arXiv:2106.13164 [cs] type: article. arXiv, June 2021. URL: http://arxiv.org/abs/2106.13164 (visited on 09/29/2022).

[42] Zach Wood-Doughty, Isabel Cachola, and Mark Dredze. “Proxy Model Explanations for Time Series RNNs”. en. In: *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. Pasadena, CA, USA: IEEE, Dec. 2021, pp. 698–703. ISBN: 978-1-66544-337-1. DOI: 10.1109/ICMLA52953.2021.00117. URL: https://ieeexplore.ieee.org/document/9680082/.

[43] Matthew D. Zeiler and Rob Fergus. “Visualizing and Understanding Convolutional Networks”. In: arXiv:1311.2901 (Nov. 2013). arXiv:1311.2901 [cs]. URL: http://arxiv.org/abs/1311.2901.

[44] Yujia Zhang et al. *”Why Should You Trust My Explanation?” Understanding Uncertainty in LIME Explanations*. en. arXiv:1904.12991 [cs, stat]. June 2019. URL: http://arxiv.org/abs/1904.12991 (visited on 07/05/2022).