

LIMREF: Local Interpretable Model Agnostic Rule-based Explanations for Forecasting, with an Application to Electricity Smart Meter Data

Dilini Rajapaksha¹, Christoph Bergmeir¹

¹Faculty of Information Technology
Monash University, Clayton Campus
Building 6, 29 Ancora Imparo Way
Monash University VIC 3800

dilini.rajapaksha@monash.edu, christoph.bergmeir@monash.edu

Abstract

Accurate electricity demand forecasts play a crucial role in sustainable power systems. To enable better decision-making especially for demand flexibility of the end user, it is necessary to provide not only accurate but also understandable and actionable forecasts. To provide accurate forecasts Global Forecasting Models (GFM) trained across time series have shown superior results in many demand forecasting competitions and real-world applications recently, compared with univariate forecasting approaches. We aim to fill the gap between the accuracy and the interpretability in global forecasting approaches. In order to explain the global model forecasts, we propose Local Interpretable Model-agnostic Rule-based Explanations for Forecasting (LIMREF), a local explainer framework that produces k-optimal impact rules for a particular forecast, considering the global forecasting model as a black-box model, in a model-agnostic way. It provides different types of rules that explain the forecast of the global model and the counterfactual rules, which provide actionable insights for potential changes to obtain different outputs for given instances. We conduct experiments using a large-scale electricity demand dataset with exogenous features such as temperature and calendar effects. Here, we evaluate the quality of the explanations produced by the LIMREF framework in terms of both qualitative and quantitative aspects such as accuracy, fidelity, and comprehensibility and benchmark those against other local explainers.

1 Introduction

Predicting electricity consumption and providing reasoning behind the predictions are important building blocks in the wider socio-economic challenge of transforming energy consumers to prosumers, needed to transform the energy sector to a carbon-emission-free industry. Forecasting the electricity demand on a smart meter level for particular households has been studied widely, and the effects of exogenous factors such as the temperature, daily, weekly, and yearly seasonalities, and other calendar effects, are well established. However, accurate energy forecasts are not enough to change the consumers' behaviour towards greater demand flexibility. We need to give understandable and actionable insights to the end user to enable behaviour change.

In the forecasting literature there are plenty of works on accurate electricity demand forecasting (Porteiro, Hernández-Callejo, and Nesmachnow 2020; Blum and Riedmiller 2013; Taylor 2003; Bedi and Toshniwal 2018; Al-Musaylh et al. 2018). The traditional energy demand forecasting methods are state-of-the-art univariate forecasting approaches such as ARMA, Trend Seasonal (TBATS) models and ARIMA models. The main limitation of these approaches is that they do not consider the information across multiple series. This cross-learning ability is particularly relevant when the time series are sharing key patterns and common structures among them. Global Forecasting Models (GFMs) (Januschowski et al. 2020) that are trained across all available time series have recently dominated traditional forecasting approaches in many cases including the M4 competition (Makridakis, Spiliotis, and Assimakopoulos 2018; Smyl 2018, 2020), the M5 competition (Makridakis, Spiliotis, and Assimakopoulos 2020), various Kaggle competitions (Bojer and Meldgaard 2020), and also real-world applications in companies like Amazon (Salinas et al. 2020), Walmart (Bandara et al. 2019), and Uber (Laptev et al. 2017).

However, there is very limited work yet to explain the forecasts of GFMs. More general, to explain model predictions in machine learning the concept of “local interpretability” has been proposed first in the work of Ribeiro, Singh, and Guestrin (2016) which produces feature importance based explanations for individual predictions rather than providing explanations globally across all instances. Subsequently, researchers proposed multiple approaches which provide explanations for the global model predictions based on feature-importance (Lundberg and Lee 2017), saliency maps (Shrikumar, Greenside, and Kundaje 2017; Baehrens et al. 2010), rule-based explanations (Ribeiro, Singh, and Guestrin 2018; Rajapaksha, Bergmeir, and Buntine 2020) and other methods (Samek et al. 2021). Out of these approaches only (Rajapaksha, Bergmeir, and Buntine 2020) produces actionable counterfactual explanations. In machine learning there are some other counterfactual explanations approaches, e.g., minimum distance counterfactuals (Wachter, Mittelstadt, and Russell 2017), feasible and least-cost counterfactuals (Ustun, Spangher, and Liu 2019), causally feasible counterfactuals (Mahajan, Tan, and Sharma 2019; Karimi et al. 2020) and counterfactual explanations

through data manifold closeness and sparsity (Verma, Dickerson, and Hines 2020; Poyiadzi et al. 2020). However, these approaches are not directly applicable to GFM and do typically not produce actionable explanations.

In this work, our aim is to propose a general methodology to explain the forecasts of a GFM. Then, we focus our method to the use case of electricity customers, to enable them to make better data-informed decisions using the generated explanations. In particular, we propose a framework called **LIMREF Local Interpretable Model-agnostic Rule-based Explanations for Forecasting** which produces different types of rules as guidance to support electricity consumption planning. Our proposed method has been developed for the FUZZ-IEEE Competition on Explainable Energy Prediction (Triguero et al. 2021). The problem posed in the competition is to provide accurate and self-explaining electricity demand predictions for a set of customers for each month over the coming year (January to December). Our solution to the competition used a GFM based on Expectile Regression (ER) (Sobotka and Kneib 2012) to generate accurate forecasts. LIMREF then provides rule-based local explanations, where, in line with the local interpretability literature, local means an explanation of the global model forecast for each series/customer separately. To the best of our knowledge, LIMREF is the first model-agnostic algorithm that provides local rule-based explanations in time series forecasting. Moreover, the provided rules explain the decision and also provide counterfactual rules which are often important as actionable insights to change the outcomes.

We evaluate our LIMREF framework and achieve competitive results on the competition dataset which contains half-hourly energy readings for 3248 smart meters, between January 2017 to December 2017, along with temperature data. Furthermore, we conduct both qualitative and quantitative experiments to assess the quality of the explanations and the global model, and are able to achieve competitive results. In summary, the contributions of this paper are as follows. (i) We propose a novel approach based on impact rules to generate local rule-based model-agnostic explanations for time series forecasting (ii) We further tailor the approach to an electricity demand prediction use case with global time series forecasting models. (iii) Our approach produces different types of rules, such as affirmative and counterfactual rules, which lead to actionable explanations.

2 Background and theoretical foundations

In this section, we discuss the relevant background from association rule mining research with respect to impact rules where the target of the dataset is a numeric variable (Webb 2001), as well as the background of local interpretability and how to translate it to forecasting.

Impact Rule Mining

Association rule mining is a rule-based machine learning approach that was first introduced by Agrawal, Imieliński, and Swami (1993) to detect interesting combinations of input and output variables in large databases. In this context, rules can be defined as follows.

Definition 1 Rule: A rule in our context follows the form $r_c = p \rightarrow q$. The left-hand side (LHS) of a rule, p , is also called the antecedent of the rule and consists of boolean conditions on feature values. The right-hand side (RHS) of a rule, q , is called the consequent of the rule and is the target value of the decision variable.

In association rules, the consequent is typically a class label. However, in our context of time series forecasting, the target outcome is a numeric value. A straightforward way to use class association rules for such a regression task, would be to use a discretized target. An advanced algorithm in this line is proposed in the work of Srikant and Agrawal (1996). However, Webb (2001) argues that the “provision of statistics relating to the frequency of one sub-range of values of a target variable does not directly address the issue of the impact on the general distribution of values for that variable.” Therefore, to mitigate the above drawbacks, Webb (2001) proposed impact rules where the segmentation of data is on the basis of a numeric value (e.g., mean, sum, median) rather than a distribution distorted by discretization. Thus, in impact rules, the impact of the target can be measured based on the following statistics.

- *Coverage*: Proportion of cases which satisfy the LHS.
- *Mean*: The mean value of the target for the instances covered by the LHS.
- *Sum*: The sum of the values of the target for the instances covered by the LHS.
- *Impact*: $\text{impact} = \text{sum} - \text{mean} * \text{absolute_coverage}$
Here, absolute coverage is the number of cases that fulfill the conditions of the LHS of the rule. This indicates the magnitude of the impact of the conditions in the LHS of the rule on the target beyond the sum expected if the LHS was independent of the target.

The main objective of discovering interesting rules is to identify the group that contributes most or least to a particular outcome. For example, considering the application of electricity demand forecasting, the conditions which explain to increase or decrease the monthly electricity demand are considered as the most interesting rules. Webb (2001) argues that identifying the group with a higher *mean* value of the target is not equal to identifying the group which contributes most or least to the total, as the group may be small. As a result, Webb (2001) concludes that rules generated based on aggregated measures such as *sum* and *impact* measures add greater interest generally. In this work, we consider the impact of the target based on the mean, which calculates the mean value of the target for the instances covered by the conditions of the LHS of the rule.

When it comes to mining the rules efficiently, in our work we use the OPUS_IR algorithm developed by Webb (2001). OPUS_IR is a statistically sound rule-mining algorithm which overcomes many of the shortcomings of other rule mining techniques by substantially pruning the search space in each pruning action (Webb 2001). OPUS search has been successfully used in both classification (Webb 1994) and association rule-mining (Webb 2000) tasks. Furthermore, it has effective control over the spurious correlations (Webb 2011). Most importantly, the algorithm achieves

a dramatic reduction of the computational time compared to other techniques while providing efficient and complete rules.

Local Interpretability in Forecasting

The concept of local interpretability in machine learning can be formally defined, and transferred to time series forecasting to explain GFM forecasts, as follows.

Let $X \in \mathbb{R}^{n \times m}$ be a training dataset with n instances and m features, and let z be an n dimensional vector of corresponding actual target values, and let g be a given global black-box background model that produces predictions \hat{z} from m dimensional input vectors x , that is, $g(x) = \hat{z}$, such that the error between z and \hat{z} is minimised. The goal of a local interpretability algorithm is to find an interpretable explainer model e , for example, a linear model, such that for a given instance x where the global model predictions \hat{z} need to be explained, $e(x) \approx \hat{z}$. To develop e , we first define a neighbourhood for the instance to be explained x following two steps. First, we select the most similar instances to x using a distance function. After, we create a set of newly generated instances that lie in the neighbourhood. We call this new set \tilde{X} . We then apply g to all these instances in \tilde{X} to obtain the global model predictions \tilde{z} . Here, we obtain \tilde{z} to identify the behaviour of the global model when the instances are similar to the instance to be explained. Thus, e is trained on \tilde{X} to resemble \tilde{z} as the target predictions \hat{z} . To generate \tilde{X} , usually in machine learning the interpretability algorithms first filter the most similar instances from the original training set X as instances that lie in the neighbourhood of x , into a set that we call X_{filt} . Then, X_{filt} is used to generate \tilde{X} by using sampling procedures and by adding noise to the features of X_{filt} .

When transferring the concept of local interpretability to global model forecasting, the global forecasting model f is considered as the background model that has been trained on a set of time series $Y = \{y_1, \dots, y_p\}$, where p is the total number of time series in the dataset. Let the global model forecasts of a particular time series y be \hat{y} , where $f(y) = \hat{y}$. The goal of this work is to develop a local explainer e which is an interpretable model, such that $e(y) \approx \hat{y}$. In our approach LIMREF to generate the neighbourhood \tilde{X} , we first filter the most similar (according to a distance measure to be chosen) time series from the original training set X as instances that lie in the neighbourhood of x . These series comprise X_{filt} . Then, X_{filt} is used to generate \tilde{X} by using bootstrapping procedures on X_{filt} . Finally, the local explainer e which is an interpretable surrogate model is fitted on \tilde{X} to provide local explanations for the forecasts \hat{y} of the time series to be explained.

3 Proposed LIMREF Approach

We first define a neighbourhood as time series that are similar to the series for which the forecasts are to be explained. Then, we use a bootstrapping technique to generate new time series from the ones in the neighbourhood. After, we generate global model forecasts for these series, namely for both

Algorithm 1: A Local Interpretable Model-agnostic Rule-based Explanations for Forecasting: LIMREF

Input : Y_t – training set without test data
 x – time series that need to be explained
 f – a global demand forecasting model
 $N_{synthetic}$ – # of new instances to be generated
 N_{filt} – # of nearest neighbours
Output: E_x – Rule-based explanations for the forecast of the instance to be explained

- 1: $X_{filt} \leftarrow \text{SelectNearestNeighbours}(Y_t, x, N_{filt})$
- 2: $X_{synthetic} \leftarrow \text{GenerateFromNeighbours}(X_{filt}, N_{synthetic})$
- 3: $\tilde{X} \leftarrow X_{filt} \cup X_{synthetic}$
- 4: $f(\tilde{X}) \leftarrow \text{GetPredictFromGlobalModel}(\tilde{X}, f)$
- 5: $\tilde{X}_{proxy} \leftarrow \text{GetSurrogateFeatures}(\tilde{X})$
- 6: $R_x \leftarrow \text{DiscoverImpactRules}(\tilde{X}_{proxy}, f(\tilde{X}))$
- 7: $E_x \leftarrow \text{GenerateGuidance}(R_x, x, f(\tilde{X}))$
- 8: **return** E_x

the bootstrapped series and the original series in the neighbourhood. Finally, we perform impact rule mining based on the approach proposed in the work of (Webb 2001) for the combined set of series and their global model forecasts to produce impact rules as the explanations. An overview diagram of the LIMREF architecture is shown in Figure 1. Our approach is furthermore summarized in Algorithm 1 and the most important steps are discussed in the following.

Select the time series most similar to the time series to be explained This step corresponds to the $\text{SelectNearestNeighbours}(Y_t, x, N_{filt})$ function in Algorithm 1. Several distance measures have been proposed in the literature for time series. The most popular one is arguably Dynamic Time Warping (DTW) (Keogh and Ratanamahatana 2005), which allows a non-linear mapping of two vectors by minimizing the distance between them. Following the work of Fu et al. (2008) we perform mean normalization of the time series before applying DTW to achieve better results. After, we select the top N_{filt} instances with the smallest DTW distance to the time series in consideration. We use the implementation `dtw.distance_matrix_fast` (Meert, Hendrickx, and Craenendonck 2020) in the `dtaidistance` package in Python.

Generate new series in the neighbourhood, and corresponding GFM predictions This corresponds to $\text{GenerateFromNeighbours}(X_{filt}, N_{synthetic})$ in Algorithm 1. Following the local interpretability paradigm, we generate new synthetic instances in the neighbourhood. As our instances are time series, we employ a time series bootstrapping procedure, namely the procedure proposed by Bergmeir, Hyndman, and Benítez (2016), Box-Cox and Loess-based decomposition bootstrap, to generate N_{filt} number of new series from each time series of the neighbourhood. We use the implementation in the `bld.mbb.bootstrap` function in the `forecast` package (Hyndman et al. 2018; Hyndman, Khandakar et al. 2007) in the R programming language. We then generate

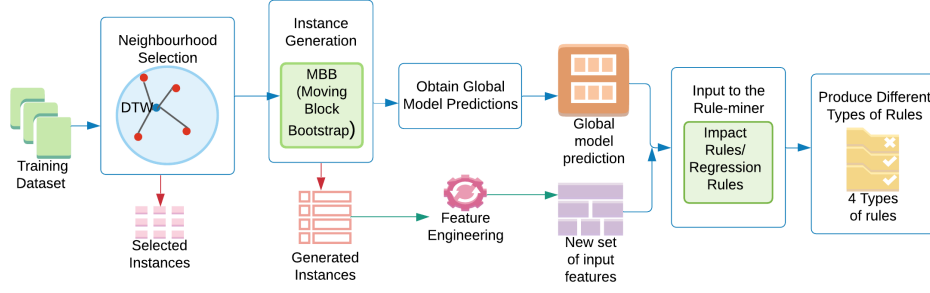


Figure 1: An overview diagram of our LIMREF architecture to generate different types of guidance in the form of rule-based explanations.

global model forecasts for all series in the combined dataset \tilde{X} , which consists of the original series in the neighbourhood, X_{filt} , and the bootstrapped series $X_{synthetic}$, as target values for the local explainer.

Generate surrogate feature set of inputs, and surrogate outputs (optional) A drawback of local interpretability algorithms such as LIME (Ribeiro, Singh, and Guestrin 2016) is that their explanations are feature importance values. Thus, if the original features are not interpretable, the explanations are not useful. In time series forecasting, we can be in a situation that the forecasts will be aggregated, e.g., the GFM provides daily forecasts that are subsequently aggregated to monthly forecasts. In this situation, explanations that use the daily input features and the daily forecasts are not useful. Thus, once all neighbourhood series are selected and generated, and forecasts are produced, to provide meaningful explanations in such a situation when the forecasts will be later aggregated, we can perform customised feature engineering tasks, which is corresponding to the $\text{GetSurrogateFeatures}(\tilde{X})$ function in Algorithm 1. This effectively generates a surrogate task with both aggregated inputs and outputs, to yield relevant explanations. As an example, in our electricity demand forecasting case study, we aggregate the input electricity consumption data and the exogenous variables such as temperature to generate a new set of features such as mean electricity demand, minimum electricity demand, maximum electricity demand, mean temperature etc., also see Figure 2.

Generate impact rule-based explanations This step corresponds to the $\text{DiscoverImpactRules}(\tilde{X}_{proxy}, f(\tilde{X}))$ function in Algorithm 1. We generate statistically sound impact rules using OPUS_IR (Webb 2001). As discussed before, by using OPUS_IR, LIMREF is able to provide explanations that contribute large amounts to the total by using *sum* and *impact* measures as the filtering option. The filtering option based on the *sum* will be useful if the target directly measures the end objective, such as the true profit from the electricity consumption. However, in the electricity demand study we aim to explain the monthly electricity consumption rate. Thus, we filter our rules based on the *impact* since our target is an intermediate variable (electricity consumption).

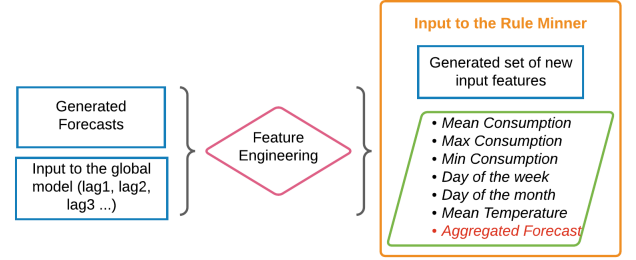


Figure 2: Feature engineering step for aggregated demand forecasting explanations.

Generate different types of guidance When explaining global classification models locally using association rules, Rajapaksha, Bergmeir, and Buntine (2020); Rajapaksha et al. (2021) propose to use four types of explanations, defined by a contingency table of the LHS and RHS of the rule. Similarly, as the final step of our LIMREF framework, we categorize our rules into four types with respect to a contingency table of the LHS and RHS of the impact rules, which will finally produce different types of guidance. This step corresponds to the function $\text{GenerateGuidance}(R_x, x, f(\tilde{X}))$ in Algorithm 1. Then, we identify the rule with the highest impact for each type of guidance as the output of LIMREF.

The LHS of a rule is considered as *true* if the antecedent of the rule applies to the surrogate feature values of the time series to be explained. In contrast to classification, in forecasting, the RHS of the rule and the global model forecast are not in categorical form. Consequently, we cannot verify directly if the consequent of the rule agrees with the forecast provided by the global model of the time series to be explained. Therefore, in our approach, the RHS of the rule is considered as *true* if the global model forecast falls within one standard deviation of the neighbours whose feature values agree with the LHS of the rule.

Let us again consider an example from our monthly electricity demand forecasting scenario, where for the meter id x the global model predicts the January electricity consumption would be (p). Assume from the neighbours of the time

series to be explained, the mean of the global model forecasts whose features agree with the LHS of the rule is \tilde{x} , and δ is a constant value. To explain this forecast LIMREF produces four types of rules which produces finally six types of guidance as follows.

Current Supporting rules ($\mathcal{R}_{c\alpha}^+$): The rules that support the forecast of the global model. *Definition:* IF LHS is true, THEN RHS is true ($p - \delta \leq \tilde{x} \leq p + \delta$). *Guidance type G1:* Explains the current practises that result in obtaining the predicted consumption p .

Current Contradicting Rules ($\mathcal{R}_{c\beta}^-$): The rules that contradict the prediction of the global model.

Type 1 Definition: IF LHS is true, THEN RHS is false, ($\tilde{x} > p + \delta$). *Guidance type G2:* Explains the current practises that indicate a result in increasing the predicted consumption p . *Benefits:* Observing the Current Contradicting Rules-Type 1, customers will be acknowledged about their ongoing risky practises which may result in an increase of the given prediction. Therefore, they should try to avoid those practises.

Type 2 Definition: IF LHS is true, THEN RHS is false, ($\tilde{x} < p - \delta$). *Guidance type G3:* Explains the ongoing practises that result in decreasing the predicted consumption p . *Benefits:* Observing the Current Contradicting Rules-Type 2, the customers are informed about which factors they should not change.

Hypothetically Supporting Rules ($\mathcal{R}_{h\alpha}^+$): The rules would increase the probability of the prediction of the global model. *Definition:* IF LHS is false, THEN RHS is true ($p - \delta \leq \tilde{x} \leq p + \delta$). *Guidance type G4:* Explains the potential practices to follow to maintain the predicted consumption p . Determine the conditions that are currently not satisfied by x (they are hypothetical) that would further result in obtaining the predicted consumption p . *Benefits:* Observing the Hypothetically Supporting Rules, if the customers are happy about the predicted consumption p , following these conditions will further maintain p .

Hypothetically Contradicting Rules (Counterfactual Rules, $\mathcal{R}_{h\beta}^-$): The rules that may invert the prediction of the global model.

Type 1 Definition: IF LHS is false, THEN RHS is false ($\tilde{x} > p + \delta$). *Guidance type G5:* Explains the potential practises that result in increasing the predicted consumption p . Determine the conditions that are currently not satisfied by x (they are hypothetical) that would further increase the predicted consumption p . *Benefits:* Observing the Hypothetically Contradicting Rules-Type 1, the customers should try to avoid these risky practises which may increase the given prediction.

Type 2 Definition: IF LHS is false, THEN RHS is false ($\tilde{x} < p - \delta$). *Guidance type G6:* Explains the potential practises that result in decreasing the predicted consumption p . Determine the conditions that are currently not satisfied by x (they are hypothetical) that would decrease the predicted consumption p . *Benefits:* Observing the Hypothetically Contradicting Rules-Type 2, the customers should try to follow these non-risky practises which may decrease the given prediction.

4 Application in Electricity Demand Forecasting

In this section we describe the dataset, decision making approach and the evaluation criteria of the proposed LIMREF framework used for our experiments.

Dataset

We use the dataset provided in the FUZZ-IEEE competition on Explainable Energy Prediction 2021 (Triguero et al. 2021) for our experiments. This dataset consists of historical half-hourly energy readings for 3248 smart meters for a one year period. The time series have different lengths, as customers may have joined at different times during the previous year. Moreover, together with the electricity consumption dataset, the competition provided access to the full-year temperature for each meter_id or customer at a daily resolution, including the average temperature, minimum and maximum of each day.

Electricity Demand Forecasting Model

Our electricity demand forecasting and decision making approach consists of two major phases: (1) developing an accurate electricity demand forecasting model, and (2) generating local explanations using the LIMREF framework to explain the forecasts of the global electricity demand forecasting model.

Though our paper presents an algorithm for local interpretability and accurate forecasting models are not the main interest for us, a forecasting model is needed to underpin our work. Our forecasting approach is based on the approach used by the team that won 4th place in the previous competition IEEE-CIS technical challenge on Energy Prediction from Smart Meter Data on the same dataset where the goal was to generate the most accurate forecasts, without regarding explainability. The following sections briefly describe the main components of the approach.

Data pre-processing *Data Aggregation:* First we aggregate the half-hourly data into daily data to reduce the impact of the large proportion of missing values in the dataset. Therefore, our forecasting model will generate a daily forecast for 365 days ahead as the intended forecast horizon of this competition is 12 months. *Missing values imputation:* To do the further imputation of missing values, a seasonal imputation approach is performed by considering the median energy consumption of each day of the week of each household. *Temperature estimation:* The competition dataset has provided temperatures only for the year 2017. A preliminary experiment which estimated temperature values with a Moving Block Bootstrapping procedure did not increase the accuracy of the global model significantly. Therefore, to estimate the temperature values for the forecasting period, we simply considered the same 2017 temperature values as 2018 values. *Normalizing data:* Since the electricity consumption varies across the different households as per the recommendations of Hewamalage, Bergmeir, and Bandara (2021) we performed a mean scaling normalization strategy.

Global forecasting model From preliminary experiments on the dataset, we found indications of systematic overestimation or underestimation of the final forecasts for yearly and monthly predictions, respectively. To correct for this, we used Expectile Regression (ER) (Sobotka and Kneib 2012) as the GFM. Based on empirical findings we set a quantile of 0.57 for longer series and 0.39 for shorter series. To implement the ER global model we used the `expectreg` (Otto-Sobotka et al. 2019) package in the R programming language. We apply the moving window forecasting strategy with a window size of 20, and we iteratively generate daily forecasts for the whole next year, by feeding back the forecasts as inputs for the subsequent time steps. We finally aggregate into monthly and yearly forecasts.

Experiments

In this section, we evaluate the LIMREF framework with two other interpretable models in both qualitative and quantitative aspects.

Select interpretable models for baseline comparisons

As LIMREF, to the best of our knowledge, is the first local rule-based model agnostic interpretability algorithm in forecasting, there are no baselines readily available. Therefore, as the baseline comparisons, we propose to obtain explanations and the forecasts by fitting existing interpretable algorithms such as linear regression (LR) and decision tree (DT) models. The LR model produces coefficients that explain the most important features as the explanations for the global model forecasts whereas the DT produces a decision tree that we consider interpretable. The LR approach is comparable to LIME (Ribeiro, Singh, and Guestrin 2016), adapted to our time series application.

We fit all three local explainers (LIMREF, LR, and DT) on the same neighbourhoods. Therewith, as LIMREF, LR, and DT are deterministic algorithms, the results are directly comparable as the neighbourhood generation is then the only computation step that adds a certain degree of randomness to the explanations.

For the LR and DT models we use a cross-validation approach to select the best models, and we use the implementations from the `glmnet` package (Simon et al. 2011) and the `rpart` package (Therneau and Atkinson 2019) in R, respectively.

Parameter setup of the local explainers Following computational feasibility considerations, we set N_{filt} to 50 which is the number of neighbours from the training set and we set $N_{synthetic}$ to 100, which is the number of synthetic instances generated by the bootstrapping procedure from each selected neighbour of the time-series to be explained. Therewith, for 3428 households the size of the bootstrapped dataset is $3428 \times 50 \times 100$, so that we consider over 17 million time series in total in the experiments.

Evaluating local explainer models

Following the literature on local explainability (Yang and Kim 2019; Carvalho, Pereira, and Cardoso 2019), we evaluate our local explainers based on the comprehensibility, fidelity, and the accuracy. Based on a qualitative analysis we

evaluate the comprehensibility of the explanations. Then, based on a quantitative analysis the fidelity and the accuracy of the local explainers are examined, which assesses how well the explainer approximates the forecasts of the GFM and the real values, respectively.

Qualitative analysis of the local explainers In this section, we illustrate with an example the comprehensibility of LIMREF, LR, and DT local explainers. We consider a household where the electricity demand forecast of the global model (ER) for the month of February is 129.27kWh. The feature values of the time series that needs to be explained are as follows.

Table 1: Original Values of the Instance to be Explained

Mean_cons	Max_cons	Min_cons	Temp	Month	ER Forecast
16.22kWh	30.75kWh	0kWh	6.01C	Feb	568.93kWh

In the following, we discuss the possibilities of changes to the electricity consumption of the given instance for the month of February using the local explainer models.

What are current practises that result in obtaining the predicted consumption 568.93kWh? This question can be answered using the current supporting rules \mathcal{R}_{ca}^+ of the LIMREF algorithm. $\mathcal{R}_{ca_LIMREF}^+$: *Your predicted consumption is 568.94kWh. Because you have mean consumption > 14.74kWh.*

Advantage: This type of rule explains the current practices that result in obtaining the predicted consumption. Hence, having mean electricity consumption greater than 14.74kWh in the past 20 days is why the predicted consumption is 568.94kWh.

Table 2: Coefficients of local explainer LR model

Mean_cons	Max_cons	Min_cons	Temp	Month
28.4	1.26	-6.86	24.14	-0.52

To answer the same question, the LR explainer provides the feature importance in the form of the coefficients of the local linear model. Accordingly, Table 2 shows that for the particular household the mean consumption and the temperature factors positively support the predicted consumption of 568.94kWh.

The DT explainer provides a decision tree as the explanation. According to Figure 3, the DT explainer suggests that having mean consumption between 13kWh and 19kWh supports to have the predicted consumption of 568.94kW in February.

What are the current conditions which result in increasing or decreasing the predicted consumption 568.94kWh? To answer this question LIMREF is providing (\mathcal{R}_{cb}^-), as follows: *The conditions that currently exist that indicate an opportunity of a decreased consumption by 170.45kWh for the particular month are $5.95 < \text{average temperature} \leq 6.01$.*

Advantage: Observing the Current Contradicting Rules-Type 2, the customers are informed about which factors they

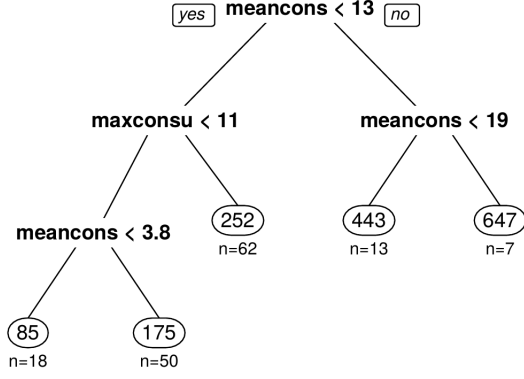


Figure 3: Local interpretable explanation of DT model

should not change. In particular, the customer understands that having an average temperature between 5.95C and 6.01C, there can be a decreased consumption of 170.45kWh. In this case, the rules are not actionable since the outside temperature is not manageable.

To answer the same question with LR, Table 2 shows that the minimum consumption and the month features are negatively supported to the predicted mean consumption. In other words, if we increase the minimum energy consumption it will negatively affect the predicted consumption. However, the outcome of the LR explainer is arguably less informative as the LIMREF output. The DT explainer cannot answer this question.

What are the hypothetical conditions that would result in obtaining the predicted consumption 568.93kWh? Only LIMREF is capable of answering this question. For that, it uses the Hypothetically supporting rules $\mathcal{R}_{h\alpha}^+$ as follows: *The conditions that need to be satisfied to maintain the monthly predicted consumption would be mean consumption > 14.74 & min consumption > 2.88.*

Advantage: By observing these rules, the customers can understand what they need to follow to maintain the current energy consumption if they are happy with their electricity bill. These rules are actionable.

What are the hypothetical conditions that would result in increasing or decreasing the predicted consumption 568.93kWh? (Counterfactual explanations) Both LIMREF and DT explainers are able to answer the above question. The LIMREF approach uses the hypothetically contradicting rules ($\mathcal{R}_{h\beta}^-$) to provide counterfactual explanations: *If you have $6.01 < \text{average temperature} \leq 6.31$ & mean consumption > 14.74 & max consumption > 24.02 it will increase your consumption by 110.64kWh. To reduce your consumption by 216.6kWh the conditions would be $6.01 < \text{average temperature} \leq 6.31$ & $15.94 < \text{max consumption} \leq 24.02$ & $0.00 \leq \text{min consumption} \leq 2.88$.*

Advantage: The customers should try to avoid these practices which may increase the given prediction. In particular, the customer should not exceed the mean consumption greater than 14.74kWh and maximum consumption greater than 24.02kWh when the average temperature is within the

given range.

For the DT, according to Figure 3, if the mean consumption is > 19, there will be an increment of the electricity consumption of the considered household. In contrast, if the mean consumption < 13, the electricity consumption of the selected household will be decreased.

Quantitative analysis of the local explainers In this section, we analyse fidelity and accuracy of the local explainers. For this analysis, we consider three error measures that are widely used in forecasting and regression. Since in the competition, the evaluation process was based on Relative Absolute Error (RAE), in our experiments we consider RAE as an error metric. Further, we consider Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) error to increase the reliability of our results.

Table 3 shows the fidelity and the accuracy of the local explainer models by calculating the error between the explainer and the global model (fidelity) and the real values (accuracy), respectively. Here, low error rates denote high fidelity and high accuracy. We see that the LR model achieves the highest fidelity by obtaining the least error across all three error metrics. LIMREF is second, and DT third in this comparison. With regards to the accuracy aspect, our approach outperforms both LR and DT explainers.

Table 3: Fidelity and the accuracy measures of the local explainers

Performance	Fidelity			Accuracy		
Explainer	RAE	RMSE	MAE	RAE	RMSE	MAE
LR	0.348	73.528	45.454	1.285	269.277	183.263
DT	0.425	94.506	55.491	1.283	268.233	182.921
LIMREF	0.390	79.522	50.945	1.259	257.605	179.552

In addition to the fidelity and the accuracy aspects, it is interesting to examine the feature importance of the surrogate features across all the instances in the dataset. Therefore, for LR and DT explainers we calculate the feature importance for each instance and average across the instances, see Table 4.

Table 4: Average feature importance of the explainers (LR, DT), and average feature usage frequency for each rule type of LIMREF.

Feature	LR	DT	$\mathcal{R}_{c\alpha}^+$	$\mathcal{R}_{c\beta}^-$	$\mathcal{R}_{h\alpha}^+$	$\mathcal{R}_{h\beta}^-$
Max_cons	1.565	2.034	0.395	0.587	0.399	0.610
Mean_cons	0.785	2.299	0.357	0.526	0.368	0.611
Min_cons	7.721	0.387	0.505	0.499	0.610	0.610
Month	2.140	0.079	0.516	0.495	0.556	0.688
Temp	2.314	0.643	0.372	0.624	0.606	0.658

According to Table 4, in the LR explainer, knowing the minimum electricity consumption is considered as the most important feature while the maximum electricity consumption is considered as the least important feature when providing explanations. However, when providing explanations based on the DT explainer knowing the mean consumption

is considered as the most important factor whereas the forecasted month has the least importance.

Table 4 furthermore illustrates how often the LIMREF explainer uses the surrogate features when providing explanations for each rule type. It shows that for $\mathcal{R}_{c\alpha}^+$, and $\mathcal{R}_{h\alpha}^+$ rules, the minimum consumption, month, and the temperature features are used most frequently while rule types $\mathcal{R}_{c\beta}^-$ and $\mathcal{R}_{h\beta}^-$ use the features mostly equally.

Model-Agnostic Global Explanations

We also perform an analysis to provide a global explanation of the forecasting model, to assess whether local explanations are needed in the first place, or a global explanation may be sufficient. Therefore, first, we generate the surrogate feature set from the whole actual dataset using the procedure for surrogate feature generation outlined earlier. After obtaining global explanations we fit LR, DT, and LIMREF explainers on the generated dataset of surrogate features and the global model forecasts. We call these models LR_g , DT_g and $LIMREF_g$ in the following.

Table 5: Fidelity and accuracy of the global explainers

Performance	Fidelity			Accuracy		
Explainer	RAE	RMSE	MAE	RAE	RMSE	MAE
LR_g	0.355	74.157	46.388	1.294	274.493	184.468
DT_g	0.367	76.094	47.897	1.310	274.033	186.807
$LIMREF_g$	0.522	95.211	68.238	1.233	257.949	175.862

According to Table 5, LR_g approximates the global model forecasts best by acquiring the highest fidelity while $LIMREF_g$ achieves highest accuracy. Comparing Tables 3 and 5, we see that the local explainer models perform better than the global explainers in terms of both fidelity and accuracy.

5 Limitations

Although the proposed LIMREF framework produces model-agnostic rule-based local explanations for the forecasts produced by the global forecasting models, it certainly has limitations. The main limitation of the current work is that the chosen application is limited by the competition dataset that we considered in our experiments. As such, we could only use the available features, and, more importantly, the forecasting horizon was very long (a full year). As such, the surrogate model’s features are limited through the dataset and problem statement and can only consider past consumption and average temperatures on a relatively high aggregation level. Also, the four types of explanations are a direct mapping from the theory, of what the algorithm generates. As they are quite general, they may not be specific enough for a particular application.

6 Conclusions and Future Work

Electricity demand forecasting is important for many applications, and global forecasting models that train across series have been proposed to perform this task accurately. However, accurate forecasts are oftentimes not enough to

make informed and good decisions. For example, customers do not know what they should and should not do to decrease their electricity consumption. In this paper, we propose the LIMREF framework, which is a method developed for the FUZZ-IEEE competition on Explainable Energy Prediction and also, to the best of our knowledge, the first model-agnostic rule-based algorithm which explains global forecasting models locally. LIMREF generates six types of guidances in the form of rule-based explanations for global forecasting models. These rule-based guidances could be communicated to energy customers, e.g., in their invoices, to help them understand their bills and what they can potentially do to lower their energy costs. We have compared LIMREF with two other approaches regarding fidelity, accuracy, and comprehensibility, showing its unique strengths and competitiveness.

In the future, we hope to validate the robustness of the LIMREF framework by applying it to different forecasting applications outside of the particular energy forecasting application case. Moreover, as with other ML models, this model is susceptible to biases in the data. However, it is commonly hypothesized that post-hoc explanations can assist in detecting model biases. Therefore, our explainer model can be able to help uncover such biases. This has not been addressed in the current paper but would be an excellent research question for future work. Furthermore, the “right” level of locality for explanations is largely unexplored in the time series forecasting domain, which can be explored more in the future.

7 Acknowledgements

We sincerely thank Hansika Hewamalage for supporting us in the global modelling part by replicating the global forecasting model developed for the previous competition, IEEE-CIS technical challenge on Energy Prediction from Smart Meter Data on the same dataset. We would also like to thank Isaac Triguero and his team for organising the FUZZ-IEEE competition on Explainable Energy Prediction. This research was supported by the Australian Research Council under grant DE190100045.

References

- Agrawal, R.; Imieliński, T.; and Swami, A. 1993. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, 207–216.
- Al-Musaylh, M. S.; Deo, R. C.; Adamowski, J. F.; and Li, Y. 2018. Short-term electricity demand forecasting with MARS, SVR and ARIMA models using aggregated demand data in Queensland, Australia. *Advanced Engineering Informatics*, 35: 1–16.
- Baehrens, D.; Schroeter, T.; Harmeling, S.; Kawanabe, M.; Hansen, K.; and Müller, K.-R. 2010. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11: 1803–1831.
- Bandara, K.; Shi, P.; Bergmeir, C.; Hewamalage, H.; Tran, Q.; and Seaman, B. 2019. Sales demand forecast in e-commerce using a long short-term memory neural network methodology. In *International Conference on Neural Information Processing*, 462–474. Springer.
- Bedi, J.; and Toshniwal, D. 2018. Empirical mode decomposition based deep learning for electricity demand forecasting. *IEEE access*, 6: 49144–49156.
- Bergmeir, C.; Hyndman, R. J.; and Benítez, J. M. 2016. Bagging exponential smoothing methods using STL decomposition and Box–Cox transformation. *International journal of forecasting*, 32(2): 303–312.
- Blum, M.; and Riedmiller, M. 2013. Electricity demand forecasting using gaussian processes. In *Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence*.
- Bojer, C. S.; and Meldgaard, J. P. 2020. Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*.
- Carvalho, D. V.; Pereira, E. M.; and Cardoso, J. S. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8): 832.
- Fu, A. W.-C.; Keogh, E.; Lau, L. Y. H.; Ratanamahatana, C. A.; and Wong, R. C.-W. 2008. Scaling and time warping in time series querying. *The VLDB Journal*, 17(4): 899–921.
- Hewamalage, H.; Bergmeir, C.; and Bandara, K. 2021. Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, 37(1): 388–427.
- Hyndman, R.; Athanasopoulos, G.; Bergmeir, C.; Caceres, G.; Chhay, L.; O’Hara-Wild, M.; Petropoulos, F.; Razbash, S.; Wang, E.; and Yasmeeen, F. 2018. forecast: Forecasting functions for time series and linear models. R package version 8.4. URL: <https://CRAN.R-project.org/package=forecast>.
- Hyndman, R. J.; Khandakar, Y.; et al. 2007. *Automatic time series for forecasting: the forecast package for R*. 6/07. Monash University, Department of Econometrics and Business Statistics.
- Januschowski, T.; Gasthaus, J.; Wang, Y.; Salinas, D.; Flunkert, V.; Bohlke-Schneider, M.; and Callot, L. 2020. Criteria for classifying forecasting methods. *International Journal of Forecasting*, 36(1): 167–177.
- Karimi, A.-H.; Barthe, G.; Balle, B.; and Valera, I. 2020. Model-agnostic counterfactual explanations for consequential decisions. In *ICAIS*, 895–905.
- Keogh, E.; and Ratanamahatana, C. A. 2005. Exact indexing of dynamic time warping. *Knowledge and information systems*, 7(3): 358–386.
- Laptev, N.; Yosinski, J.; Li, L. E.; and Smyl, S. 2017. Time-series extreme event forecasting with neural networks at uber. In *International conference on machine learning*, volume 34, 1–5.
- Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, 4768–4777.
- Mahajan, D.; Tan, C.; and Sharma, A. 2019. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*.
- Makridakis, S.; Spiliotis, E.; and Assimakopoulos, V. 2018. Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PloS one*, 13(3): e0194889.
- Makridakis, S.; Spiliotis, E.; and Assimakopoulos, V. 2020. The M5 accuracy competition: Results, findings and conclusions. *Int J Forecast*.
- Meert, W.; Hendrickx, K.; and Craenendonck, T. 2020. wannesm/dtaidistance v2. 0.0.
- Otto-Sobotka, F.; Spiegel, E.; Schnabel, S.; Schulze Waltrup, L.; Eilers, P.; Kauermann, G.; and Kneib, T. 2019. *expectreg: Expectile and Quantile Regression*. R package version 0.50.
- Porteiro, R.; Hernández-Callejo, L.; and Nesmachnow, S. 2020. Electricity demand forecasting in industrial and residential facilities using ensemble machine learning. *Revista Facultad de Ingeniería Universidad de Antioquia*.
- Poyiadzi, R.; Sokol, K.; Santos-Rodriguez, R.; De Bie, T.; and Flach, P. 2020. FACE: Feasible and actionable counterfactual explanations. In *AAAI/ACM*, 344–350.
- Rajapaksha, D.; Bergmeir, C.; and Buntine, W. 2020. LoRMiK: Local rule-based model interpretability with K-optimal associations. *Information Sciences*, 540: 221–241.
- Rajapaksha, D.; Tantithamthavorn, C.; Bergmeir, C.; Buntine, W.; Jiarapakdee, J.; and Grundy, J. 2021. SQAPlaner: Generating data-informed software quality improvement plans. *IEEE Transactions on Software Engineering*.
- Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. ACM.
- Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.
- Salinas, D.; Flunkert, V.; Gasthaus, J.; and Januschowski, T. 2020. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191.

- Samek, W.; Montavon, G.; Lapuschkin, S.; Anders, C. J.; and Müller, K.-R. 2021. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3): 247–278.
- Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, 3145–3153. PMLR.
- Simon, N.; Friedman, J.; Hastie, T.; and Tibshirani, R. 2011. Regularization Paths for Cox’s Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, 39(5): 1–13.
- Smyl, S. 2018. *M4-methods*. Uber Technologies.
- Smyl, S. 2020. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, 36(1): 75–85.
- Sobotka, F.; and Kneib, T. 2012. Geoadditive expectile regression. *Computational Statistics & Data Analysis*, 56(4): 755–767.
- Srikant, R.; and Agrawal, R. 1996. Mining quantitative association rules in large relational tables. In *Proceedings of the 1996 ACM SIGMOD international conference on Management of data*, 1–12.
- Taylor, J. W. 2003. Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of the Operational Research Society*, 54(8): 799–805.
- Therneau, T.; and Atkinson, B. 2019. *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-15.
- Triguero, I.; Alonso, J. M.; Magdalena, L.; Wagner, C.; and Bernab  -Moreno, J. 2021. FUZZ-IEEE Competition on Explainable Energy Prediction. <https://ieee-dataport.org/competitions/fuzz-ieee-competition-explainable-energy-prediction>.
- Ustun, B.; Spangher, A.; and Liu, Y. 2019. Actionable recourse in linear classification. In *Proc. Conf. on Fairness, Acc. and Trans.*, 10–19.
- Verma, S.; Dickerson, J.; and Hines, K. 2020. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*.
- Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31: 841.
- Webb, G. I. 1994. Recent progress in learning decision lists by prepending inferred rules. In *SPICIS’94: Proceedings of the Second Singapore International Conference on Intelligent Systems*, B280–B285.
- Webb, G. I. 2000. Efficient search for association rules. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 99–107.
- Webb, G. I. 2001. Discovering associations with numeric variables. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 383–388.
- Webb, G. I. 2011. Filtered-top- k association discovery. *WIREs Data Mining Knowl Discov*, 1(3): 183–192.
- Yang, M.; and Kim, B. 2019. BIM: Towards quantitative evaluation of interpretability methods with ground truth. *arXiv preprint arXiv:1907.09701*.