Machine Learning (2022) 111:487–517
https://doi.org/10.1007/s10994-021-06051-1

Non-technical losses detection in energy consumption focusing on energy recovery and explainability

Bernat Coma-Puig[1] · **Josep Carmona**[1]

Received: 30 September 2020 / Revised: 6 August 2021 / Accepted: 27 August 2021 /
Published online: 29 September 2021

Abstract
Non-technical losses (NTL) is a problem that many utility companies try to solve, often using black-box supervised classification algorithms. In general, this approach achieves good results. However, in practice, NTL detection faces technical, economic, and transparency challenges that cannot be easily solved and which compromise the quality and fairness of the predictions. In this work, we contextualise these problems in an NTL detection system built for an international utility company. We explain how we have mitigated them by moving from classification into a regression system and introducing explanatory techniques to improve its accuracy and understanding. As we show in this work, the regression approach can be a good option to mitigate these technical problems, and can be adjusted in order to capture the most striking NTL cases. Moreover, explainable AI (through Shapley Values) allows us to both validate the correctness of the regression approach in this context beyond benchmarking, and improve the transparency of our system drastically.

Keywords Non-technical losses · Explainability · Regression · Classification · Robustness

1 Introduction

The services provided by energy companies are essential to societies, but are rather expensive: the necessary infrastructure to provide them includes power plants, kilometres of pipes and lines, and millions of meters, whose economic cost is covered by the bills paid by the companies' customers and, in many cases, also taxes.

Another less visible cost that these companies face are the energy losses, i.e., the gap between the energy provided and the energy billed to the customers. The energy losses caused by the physical properties of the power system components are referred to as Technical Losses

Editors: João Gama, Alípio Jorge, Salvador García.

✉ Bernat Coma-Puig
bcoma@cs.upc.edu

Josep Carmona
jcarmona@cs.upc.edu

[1] Universitat Politècnica de Catalunya, Barcelona, Spain

and cannot be easily avoided. In contrast, the losses caused by meter malfunctions and fraudulent customer behaviours, known as non-technical losses, correspond to losses that the company aims to eradicate.

Companies usually perform a pre-selection of suspicious cases of NTL to be visited by a technician (an activity known as *campaign*) to check if the installation is correct. In the past, the customers' pre-selection was based on simple rules indicating an abnormal consumption behaviour according to the stakeholder's knowledge (e.g., an abrupt decrease of consumption). This approach usually had a low success rate because these behaviours can often be explained by other reasons besides fraud, for instance, a long convalescence in a hospital. Nowadays, in the era of big data and machine learning, utility companies exploit the data available in their information systems and combine them with other contextual information to design more accurate campaigns, including statistical and machine learning-based techniques.

One of these systems is the NTL-Detection classifier system we have implemented for an international utility company from Spain (Coma-Puig et al. 2016; Coma-Puig and Carmona 2019). This consists of a supervised classification approach in which the system learns from historical NTL cases (and non-NTL cases) a model to predict how suspicious a customer is at present. As we explain in Sect. 2, this approach is very common in the literature, as it allows automating the generation of campaigns.

After several years of working in our system, we detected that this approach faces several challenges that cannot be easily solved, and which compromise the quality and fairness of the predictions. From a technical point of view, our system lacked robustness due to data-related problems, a common problem in the existing NTL literature (Glauner et al. 2017). Remarkably, the trade-off between the energy recovered (i.e. the energy that should be charged for the NTL cases detected) and the campaign cost (sending technicians to check selected meters) was often unsatisfactory. Finally, the use of black-box algorithms compromised the transparency of our system.

This work proposes a novel approach to detect NTL cases: a predictive regression system, where the prediction target is the amount of energy recovered for each NTL case. In theory, the change from classification to regression means establishing a priority among our NTL cases, making the supervised algorithm focus on the variables related to customer consumption. According to the experiments described in Sect. 4, the results confirm that the regression approach is a valid alternative to classification when there exist problems in terms of energy recovered and system robustness. Classification is the most common approach in the literature.

Moreover, our analysis beyond benchmarking confirms the correctness of the regression model in terms of explainability: we report that regression learns better and more reliable patterns than our previous classification system. To this end, we analyse both models using the SHAP library (Lundberg and Lee 2017) to determine the contribution of each feature value in each prediction through Shapley Values (Shapley 1953). This work is the first one to show the use of Shapley Values in analysing the correctness of an NTL detection model.

Finally, Sect. 6 concludes the paper and analyses the benefits of using regression and an explainability algorithm in detecting NTL. It also provides research lines for the future.

2 Related work

2.1 Related work in NTL

Our approach of using a black-box classification algorithm to detect NTL cases is very common in the literature.

From the approaches that use Ensemble Tree Models, we would like to highlight (Buzau et al. 2018), a similar approach to ours (it uses Gradient Boosting models and is also implemented in Spain). Another option used to detect NTL is the Support Vector Machine (SVM). In general, SVM-based solutions use as kernel the radial basis function (e.g. (Nagi et al. 2009) and its update (Nagi et al. 2011), the latter including Fuzzy Rules to improve the detection), and the sigmoid kernel (e.g. Depuru et al. 2013). In Costa et al. (2013), Pereira et al. (2013) and Ford et al. (2014) three examples of using neural networks to detect NTL are described. Artificial Neural Networks (ANN) are very popular in Machine Learning, and this is apparent in the NTL detection literature, where several examples of systems that use ANN can be found. The classical approach in the literature is the ANN with several layers and back-propagation, but there are also examples of extreme learning machines (i.e. feedforward neural networks with nodes that are not tuned) (Nizar et al. 2008). Finally, there exist in the literature several examples of using Optimal-Path Forest Classifier (Papa et al. 2009) to detect NTL. The works in Ramos et al. (2011b, 2018) are examples of this rather new non-parametric technique that is grounded on partitioning a graph into optimum-path trees and shares similarities with the 1-Nearest Neighbour Algorithm (Souza et al. 2014). Other algorithms used to build supervised models to detect NTL are the k-nearest neighbour (a technique used in general as a baseline model to compare the proposed approach, as can be seen in Ramos et al. 2011a), or Rule Induction (e.g. León et al. 2011)

In contrast to the aforementioned supervised techniques, there are also other different approaches to detect NTL cases; in Badrinath Krishna et al. (2015) and Angelos et al. (2011) there are two examples of using clustering; in Cabral et al. (2008) there is an example of using unsupervised neural networks (Self-Organizing Maps). In Spirić et al. (2015) and Liu and Hu (2015) there are two examples of using unsupervised methods that focus on statistical control to detect NTL cases, and in Monedero et al. (2012) a Bayesian Network is implemented, an approach that guarantees the interpretability of the directed acyclic graph.

In addition to the previous data-oriented solutions, the existence of sensors and smart grids allow other non-data solutions. For instance, Kadurek et al. (2010) presents an approach for analyzing the load flow; and in Xiao et al. (2013) a group structure is proposed with a head smart grid (referred to as *inspector*) that controls the sub-meters (i.e. the customers' meters), an approach that facilitates the detection of NTL in highly populated cities.

In Messinis and Hatziargyriou (2018) there is a survey that summarises the approaches seen in the literature, including data-oriented solutions (i.e. supervised and non-supervised approaches), network-oriented and hybrid.

In conclusion, several complementary techniques are available in the literature. We believe some of them can be combined (e.g. outlier detection as a preprocessing step) to improve NTL detection's overall performance.

2.2 Related work in robustness and explainability

The main challenge that a predictive model faces is the quality of data. If the data does not properly represent reality, then it is challenging to guarantee reliability, accuracy or fairness in the predictive model (Saria and Subbaswamy 2019; Yapo and Weiss 2018). In some cases, the problem is bias-related, and if there is a feedback loop (i.e. the model learns based on its previous predictions), the bias is aggravated in each new prediction made (Mehrabi et al. 2019; Mansoury et al. 2020). In other cases, the problem is related to the fact that the dataset evolves over time. Therefore, the labelled instances from the past could not represent the actual customers at present (i.e. Concept Drift Tsymbal 2004). Moreover, there are also model-related problems that could hinder the robustness of a predictive model. The main reason for these problems is that the algorithm does not learn causal patterns but correlations (Pearl and Mackenzie 2018). These correlations might not be robust patterns. All these problems cannot be easily controlled and mitigated if the predictive model is a black-box algorithm (e.g., Deep Learning or the Ensemble Tree Models).

Over the last few years we have seen an effort in the machine learning community to build methods and algorithms to explain through human-understandable information (i.e. textual, numerical or visual explanation) how the black-box algorithms learn. The process of explaining a prediction can be summarised as follows: being M the supervised model trained with labelled data $\{(x_1, y_1), ..., (x_n, y_n)\}$, where $x_i = \{v_{i1}, v_{i2}, ..., v_{im}\}$ is the feature vector and y_i the label to predict, the explanatory model E aims to provide an explanation of how each v_i influenced the prediction, i.e., if the value of the feature was relevant to the prediction made. This generic description fits differently, as explained in Arrieta et al. (2020), depending on the algorithm used, the task to be automatised and the method used to explain the model. For this work we focus our explanation on the methods tested for our system, i.e. Feature Importance, LIME and SHAP, post-hoc solutions for an Ensemble Tree Model. Hereunder there is a brief description of each method:

Feature Importance In Tree Models (e.g. a boosting of trees), the Feature Importance method provides a generic approach to how each feature influenced the training process. This naive definition includes the method implementation from *Scikit-learn*Pedregosa et al. (2011) (that evaluates the Gini impurity of the samples of the nodes decrease after a split using that feature), or the *LossFunctionChange* and *PredictionValueschange* from *Catboost*, two methods that evaluate how the loss function or the prediction change with or without the inclusion of the feature. Other approaches to measuring the importance of a feature consist of counting the split occurrences, i.e. how many times the feature has been used in the splitting process. All these approaches can only provide modular explanations.

LIME A Local Surrogate model is a simple interpretable model L that replicates the prediction made by a black-box algorithm M for one specific instance x (i.e. it provides local explanations). Once achieved that $L(x) \simeq M(x)$, then $L(x)$ can be used to explain the prediction from M, keeping the complexity of L as low as possible, for example using as few features as possible to provide a simple and interpretable explanation. LIME is a model-agnostic state-of-the-art implementation on this explanatory approach and has different implementations to explain tabular data, text and images. In Coma-Puig and Carmona (2018), we analysed the use of LIME as a rule-based double-checking method to discard high-scored customers with unreliable explanations from LIME.

SHAP Shapley Values (Shapley 1953) is a method to analyse the importance of each player in a cooperative game to reasonably determine the importance of each player for

the payoff. SHAP (Lundberg and Lee 2017) adapts this idea to determine how much the value of each feature of x influences the prediction $M(x)$. From a Base Value that corresponds to the mean of the labelled instances in the training set, SHAP analyses how each feature in each instance increases or decreases this Base Value to achieve the final prediction from M finally.

The Shapley Values of a feature value in instance x is usually defined as:

$$\psi_i = \sum_{S \subseteq \{x_1,\ldots,x_m\} \setminus \{x_i\}} \frac{|S|!(p-|S|-1)!}{p!} \Big(val\big(S \cup \{x_i\}\big) - val(S) \Big)$$

where p corresponds to the number of features, S a subset of the features from the instance and *val* corresponds to the function that indicates the payout for these features. In the equation, the difference between the *val* corresponds to the marginal value of adding the feature in the prediction for a particular subset of features S. The summand denotes all the possible subsets S that can be done without including the feature from which the Shapley Values is calculated, i.e., v_j. Finally, $\frac{|S|!(p-|S|-1)!}{p!}$ corresponds to the permutations that can be done with subset size $|S|$, to properly distribute the marginal values between all the features of the instance. All possible subsets of features are considered, and the effect in the prediction of including the feature to each subset is observed.

SHAP is model agnostic and provides different methods to compute the Shapley Values, depending on the predictive algorithm used. In our system, we use the Tree Explainer, the specific method to extract the Shapley Values from Tree Models (Lundberg et al. 2018). Some examples of the use Shapley Values are; Lundberg et al. (2018), an example of using the Shapley Values to prevent hypoxaemia during surgery; Galanti et al. (2020), an example of using Shapley Values to explain LSTM models in predictive process monitoring (business process management); or Posada-Quintero et al. (2020), a social science work in which the Shapley Values are used to understand the risk factors associated with teacher burnout.

From a more theoretical point of view on how to guarantee reliable models and explanations, we highlight (Rudin 2019), representing a rigorous analysis of the current explainability approaches in the literature, and their lacks that alleges for the use of interpretable algorithms when possible. Doshi-Velez and Kim (2017) provides a vision of the role of the explainable methods to obtain fairness and robustness in our predictive models. Finally, Molnar (2019) analyses the pros and cons of most of the interpretable models and the state-of-the-art of explanatory model agnostic algorithms.

3 Our NTL detection system

3.1 The system

Over the last few years, the Universitat Politecnica de Catalunya has developed an NTL detection system for the international utility company Naturgy. The system built can be summarised as follows (see Fig. 1):

1. *Campaign configuration* The stakeholder delimits the scope of the campaign (e.g. region and tariff), and the system extracts the required data from the data sources (i.e. com-

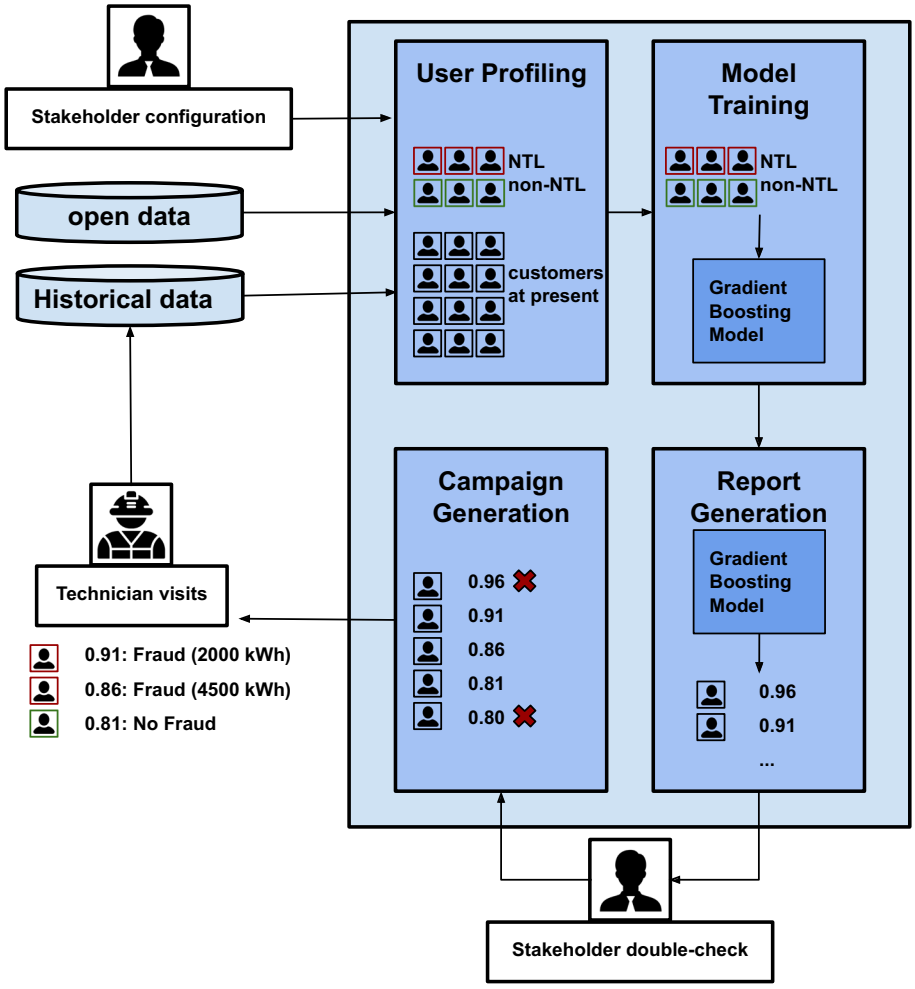


Fig. 1 The NTL detection Framework: after the stakeholder configures the campaign to be carried out, the system loads the data, trains the supervised model and predicts the binary scores, the stakeholder builds the campaign based on these scores (discarding those that, according to their knowledge, should not be included in the campaign), and the technician visits the meter installation. The campaign results (i.e. if there exist NTL and an estimation of the energy to be recovered) are updated in the data sources

pany's databases and pre-processed open data). The information from the company's databases used for this work is updated once a month.

2. *User profiling and Model Training* The system profiles both the customers in the past (when they were visited) to train a Catboost (Prokhorenkova et al. 2017) supervised classification model, and scores the profiles of the customers at present, assigning a probability score of committing fraud or having an energy loss.

3. *Report generation and campaign generation* Once the scores are assigned, the stakeholders analyse the high-scored customers. If the stakeholders validate the scores assigned (i.e. no biases or undesired characteristic are detected, like for instance the scores being

biased towards a particular region), the company builds a campaign based on these scores. Customers who have been recently visited or controlled by other means (e.g. the recidivist customers are controlled in specific campaigns) are dropped from the final list.

4. *Feedback* The result of the inspection (i.e. if the customer committed fraud or had an NTL case, or if the installation was correct, or the impossibility of checking the meter installation), as well as the estimation of the amount of energy loss that should be charged in a back-payment, is included in the system.

Each customer is profiled with around 150 features, which can briefly be explained as follows:

Consumption-Related Features The consumption-related features are the most important information in the profile since they should reflect abnormal consumption behaviour. The consumption features included in the profile can be divided into several groups:

- Raw Information: Consumption of the customers in kWh in a period of time. We include long-term features (e.g. the consumption of the customer during the last 12 months or the previous 12 months) to provide information of the customer's consumption at present, i.e. the consumption of the customer during the last three months.
- Processed Information: These features aim to represent changes in the consumption behaviour that could indicate suspicious behaviour. We include features that compare the consumption of the customer at present in comparison to itself in the past (e.g. to detect an abrupt decrease of consumption), and also features that compare the consumption of a customer in a period of time in comparison to the expected consumption (i.e. the consumption of similar customers in terms of Tariff and Region); this allows us to detect both periods of low consumption, but also abnormal consumption curves.

To build these features, we use the customer's meter readings, the billing information, and some processed information from the company.

Visit-Related Features Another important group of features are the visit features that indicate visit-related information of the customer:

- NTL cases: Information related to the NTL cases of the customer, including how many times the customer has committed fraud (or had a meter malfunction) or the last time the customer committed fraud.
- Non-NTL cases: Similarly to the NTL features explained above, we also record and represent with features how many times the customer has been visited with no NTL case, and also the last time there was this type of visit.
- Impossible visits: When a visit could not be carried out, the result of the visit is neither an NTL case nor a non-NTL case. However, we include this information in different features because it can be representative of abnormal behaviour: the customer would not facilitate the meter reading to continue committing fraud.
- Threats: Finally, we also include features about how many times a customer threatened a technician during the check and the last time of a threat. These features are clearly related to suspicious behaviours.

Static Features Less important features are the static information (i.e. contractual information that does not usually change over time). These features include the customer's tariff, the meter location, or the property of the meter. We do not consider these features key information in terms of NTL patterns, but they are included to contextualise the consumption-related and visit-related features. For instance, in case of an abrupt decrease in consumption, a customer with the meter inside the house should be more suspicious than a customer that has the meter accessible.

Sociological Features We included information related to each town's inhabitants' average income, the unemployment in that town or the proportion of inhabitants that lived in conflicting neighbourhoods. This information helped us to determine economically depressed areas in Spain. The sociological data have a similar role to the Static features, i.e. to contextualise the consumption and visit-related features. For instance, given two similarly suspicious customers, the customer that lives in a poorer region with higher unemployment may be considered more suspicious of committing fraud.

3.2 System goals and challenges

The system explained in Sect. 3 has been successful as an NTL detection system. Nevertheless, several problems were detected. These problems are explained below.

3.2.1 Technical challenges

In general, our system has achieved good results, especially considering that it is implemented in a European region with a very low ratio of NTL cases. However, the robustness of our system campaigns varied depending on the type of campaign. For instance, our system is accurate in certain types of campaigns where the type of customer was predefined (e.g. customers with no current contract,[1] or customers with long periods of no consumption).[2] However, in more generic campaigns (i.e. campaigns that included hundreds of thousands of customers) the system underperforms in robustness, i.e. the system cannot consistently provide good results.

According to our experience and knowledge, two fronts explain these problems: the existing biases in the labelled instances available from the company and the difficulty of properly benchmarking a model using a validation dataset.

Regarding the data-related problems, we have already explained in Coma-Puig and Carmona (2019) how we detected different types of biases and other data-related problems in our data. These problems are a direct consequence of using observational data produced for other purposes. Therefore, the available information does not reliably represent reality, and it is a challenge to ensure generalisability since the assumption that the labelled and the unseen instances are i.i.d, i.e. independent and identically distributed, is not met. For instance, the fact that the company visits more customers suspected of NTL leads to an over-representation of these customers, meaning that *average* customers with a normal consumption are grossly under-represented in the system. A similar problem is that the

[1] Customers with no contract refers to the customers that had a contract in the past, but the contract is currently cancelled. In many cases, these customers maintain the wire and meter installation and, therefore, can commit fraud. Our system has achieved many campaigns of around 50% of precision.

[2] As people in Spain move to cities, many villages become empty. This is a problem for the company as they do not know how to differentiate a house without consumption because it is a second home with punctual consumption or a fraudulent client. Our system was able to detect NTL cases for these types of customers with a precision of up to 36%.

company generates more campaigns in those regions where it has historically achieved better results, making the quality of the labelled information in under-visited regions very low. Therefore, it is a challenge to continually build robust models when the labelled dataset does not correctly represent reality.

Our first efforts consisted of implementing classical machine learning techniques, e.g. to modify the model's regularisation and tuning, but no improvement in the campaigns was observed. Similarly, we attempted to improve the labelled information used to train the model, e.g. by weighting the customers according to their representativeness, balancing the class imbalance typical of fraud detection problems, or implementing a cost-sensitive solution. However, after applying these solutions, the results were inconclusive: some of the experiments validated in our labelled information had initially unsuccessful results in real campaigns. Moreover, the company's demand for having short-term results made us rule out the generation of exploratory campaigns with these techniques that could offer us a long-term improvement of the system. All of this evidenced the difficulty of benchmarking our NTL system on validation datasets and a scalar metric (Drummond and Japkowicz 2010).

At this point, we discarded the most complex methods and introduced some simple solutions that could be easily validated. For example, in Coma-Puig and Carmona (2019) we explained how we segmented the customers to build more targeted campaigns to mitigate imbalance-related problems. For benchmarking, we used the Average Precision Score,[3] which provides a good generic vision of how well a model ranks, without setting a threshold when the data is highly imbalanced (Davis and Goadrich 2006). These solutions improved our system. Nevertheless, the system was still not sufficiently reliable for its industrial adoption.

3.2.2 Economic efficiency

The use of machine learning solutions to generate campaigns is justified if it provides a better solution than a random selection of customers or a baseline non-smart method (e.g. a basic rule system consisting of visiting those customers that have had an abrupt decrease of consumption). The term *better solution* includes different aspects from the company's point of view but can be summarised in the following two dimensions:

- The machine learning solution is more precise than other solutions, i.e. the proportion of True Positives is higher than the random selection or the rule-based approaches.
- The machine learning solution recovers more energy than other solutions, i.e. the energy estimated that the NTL cases have not paid (and should be charged in the near future to those customers) is higher than the energy recovered from random selection or rule-based campaigns.

Therefore, a campaign with a low precision but a large amount of energy recovered would be considered a successful campaign. Similarly, a campaign with fairly low energy recovered would also be considered a good campaign if many NTL cases are discovered, as it

[3] The Average Precision Score is the scalar value that results from summarising a precision-recall curve as the weighted mean of precisions at each threshold, using as weight the increase in recall from the previous threshold.

would prevent energy loss in the future. Understandably, an excellent campaign would be able to combine both good precision and a high amount of energy recovered.

To better understand what would be considered a good campaign in terms of energy recovered, it is necessary to note that the average annual electricity consumption per household in Spain is about 3500 kWh. In addition, the distribution company can legally invoice the NTL for one year: "*... the distribution company will invoice an amount corresponding to the product of the contracted power, or that should have been contracted, for six hours of daily use during one year,...*".[4] Under these circumstances, the following classification has been considered for the purpose of analysing the NTL cases detected according to the energy recovered:

- $>$ 3500 kWh recovered: The detection of these customers is a priority due to the amount of energy lost.
- Between 3500 kWh and 2000 kWh recovered: These NTL cases are also important. As in the previous example, the consumption curve should reflect an abnormal behaviour that the predictive system should be able to detect, e.g. a long period of low consumption.
- Between 2000 and 500 kWh recovered: These NTL cases should have some abnormal consumption behaviour (e.g. a recent abrupt decrease of consumption). However, their detection should not be prioritised over the customers with an NTL case estimated to recover energy $>$ 2000 kWh.
- 500 kWh or less: Although these are NTL cases, their consumption behaviour might not properly represent the NTL behaviour (e.g. an abnormal consumption curve or an abrupt decrease of consumption). Therefore, these NTL cases might not be prioritised over the previous NTL cases, as they might include in some cases noise or biases in the system.

From the company's point of view, our system tended to detect NTL cases with low energy to recover. For this reason, some machine learning techniques were implemented to increase the amount of energy to recover (e.g. weighting the customers according to the energy recovered). However, the results obtained after applying these solutions were inconclusive and, in many cases, seemed to aggravate some of the existing data biases (e.g. by oversampling the customers from specific regions).

3.2.3 System transparency

Although it is generally accepted in the literature that the black-box algorithms are more accurate than other more interpretable approaches, their use poses a clear problem in terms of transparency, which greatly hampered the development of our system. The problems explained above and the lack of conclusive results in our tests were a direct consequence of the impossibility of understanding how the methods implemented impacted our system.

This lack of transparency affected the company's stakeholders in different ways. On the one hand, the stakeholders historically in charge of generating the NTL campaigns could not validate the patterns learned by the model. As widely analysed in the literature (Pearl and Mackenzie 2018; Pearl 2009; Arrieta et al. 2020), the supervised methods only detect

[4] Real Decreto 1955/2000, de 1 de Diciembre, art. 87.

correlations, and therefore human supervision is necessary to validate them as reliable causal patterns (or, at least, reliable correlations in the company's context). The use of a black-box algorithm made this task challenging, so they could neither easily detect undesired patterns nor suggest system improvements. On the other hand, managers in charge of setting company guidelines had to make decisions regarding the use of the system (i.e. whether to have confidence in the system and use it to generate campaigns) in a blind manner, based solely on their results.

As explained in more detail in Sect. 5, our first approaches (i.e. to use Feature Importance and LIME) to provide explainability to our system (and therefore to make our system more transparent for the stakeholders) were insufficient.

3.3 Regression and explainability to improve our system

In this work we propose a novel approach to detecting NTL: to use as a label the amount of energy recovered in an NTL (considering that a non-NTL has a 0 label). The benefits of using the regression approach in our context (i.e. in an NTL system with biased information due to observational data) are discussed in the following sections, where this small change in our NTL system allowed us to achieve better campaigns in terms of energy recovered. Moreover, as we explain in Sect. 5, we introduce Shapley Values to obtain robust and reliable explanations from our system. As explained in this work, the purpose of using Shapley Values is to compare each approach beyond benchmarking (an approach that, as explained earlier, has given many inconclusive results) and improve the transparency of our system.

4 The regression approach for NTL detection

4.1 From classification to regression in NTL detection

The classification and regression models are two supervised methods that can be defined as follows: being X the labelled instances $\{(x_1, y_1), ..., (x_n, y_n)\}$, where x_i is the feature vector that represents an instance and y_i the value to be predicted, the supervised model aims to learn the function $f, Y = f(X)$, wherein a classification model Y is either 0 or 1 (or $0 \leq Y \leq 1$ if the model provides probabilities), and in a regression model the value to predict is continuous (i.e., $Y \in \mathbb{R}$).

The classification approach to detect NTL is widely seen in the literature (see, for instance, the examples from Related Work, Sect. 2 or our work explained in Coma-Puig et al. 2016). This approach, despite the good results that it can achieve (in Coma-Puig and Carmona 2019 we explain how we have achieved campaigns with an accuracy higher than 50%), oversimplifies the representation of the reality in our NTL detection system since it equalises the importance of each NTL case: both the customer that has been committing NTL for one year and has stolen 3000 kWh and the customer that had a meter problem for a few weeks (and therefore the energy loss is low), have the same label, even though the former case is much more important for training a supervised model for NTL detection. The higher the energy recovered, the better, as already introduced in Sect. 3.2.2, is true for several reasons.

– On equal terms, it is preferable to recover more energy at once in each visit from an economic point of view.

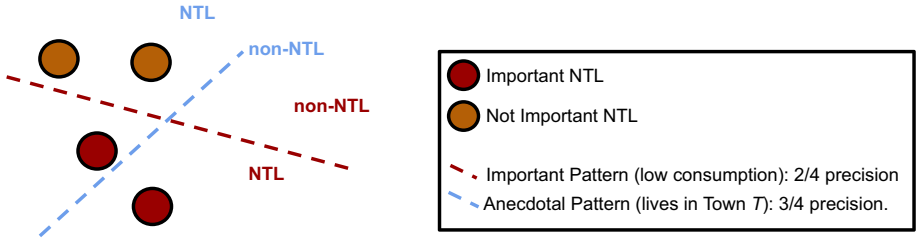


Fig. 2 With the binary classification we are equating the importance of each NTL, learning undesired patterns: if we do not prioritise the darker red instances (NTL cases with a large amount of energy recovered and, therefore, better representatives of the behaviour of an NTL case), we might prioritise undesired patterns like the one represented in a blue pattern. The result is a biased model that cannot robustly detect NTL cases

– The company usually detects short-term NTL cases through smart meter sensors. That is, if the smart meter detects a manipulation, it sends a signal to the company to warn about that manipulation, taking some days (or weeks) to include that customer in a campaign. Focusing on detecting these cases through data analysis may overlap the sensor NTL detection method. However, the long-term NTL cases are NTL cases that remain undetected.

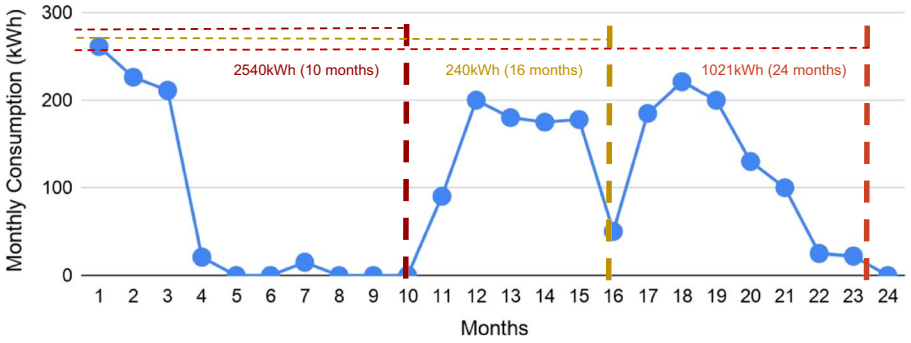
– The company might have problems recovering all the NTL from long-term fraudulent customers due to legal reasons. For this reason, companies focus their efforts on detecting these long-term fraudulent customers to reduce the difference between the energy loss and the energy they will be able to bill.[5]

Moreover, as we explain in Sect. 3.2.1, we work with observational data, i.e. data produced for other purposes that has not been prepared nor randomly sampled to properly represent the actual customers. The fact that the labelled information available corresponds to customers visited to control abnormal behaviour (or correct a meter problem), altogether with other company-related decisions that aim to maximise the campaign results (e.g. the companies usually over-control the customers that constantly commit fraud), lead the training dataset available to train the model to not represent the reality of the company's customers properly, disserving the machine learning process. Consequently, we are dealing with the existence of dataset-shift, i.e. the joint distribution of inputs and outputs differs between the training and test datasets: if $P_{population}(x)$ and $P_{labeled}(x)$ denote the real population and labelled (train) fraud distributions, it often happens that $P_{labeled}(x) \neq P_{population}(x)$, since $P_{labeled} = P_{population}(x|s = 1)$, where s is the binary condition that indicates if the customer is included in the training dataset, in our case if the customer was visited. All these problems cause the robustness degradation of our classification approach, explained in Sect. 1 and visually represented in Fig. 2.

In this work, we propose to use the energy to recover as the value to be predicted by the model, i.e., to convert our classification approach with a LogLoss function model, i.e.

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n}[y_i \cdot log_e(\hat{y}_i) + (1 - y_i) \cdot log_e(1 - \hat{y}_i)]$$

[5] For instance, not detecting a long-term NTL customer (e.g. 20 months of energy loss) will increase the energy stolen by the customer. A customer that has been committing NTL for three months will also steal energy, but the company will still be able to bill all the stolen energy if it is detected during the next nine months.



Profile analysis

2540 kWh: Abrupt decrease of consumption, several months with no consumption. Detected because of its consumption profile.
240 kWh: No significant abrupt decrease of consumption, average consumption during the last months relatively high. Probably detected due to the previous NTL detection. Non-relevant consumption patterns.
1021 kWh: Gradual decrease of consumption, average consumption during the last months low but not close to 0. More relevant NTL case than the 240 kWh instance

Fig. 3 Twenty-four months consumption curve from a recidivist customer that has committed fraud three times (each vertical line corresponds to the moment the company detected that the customer was committing fraud, with the amount of energy recovered). A binary approach would label each case equally (i.e. as a positive instance), overlooking the fact that each NTL detection is different, and needs to be contextualised. The RMSE regression approach would set the desired priority

into a regression problem, where the value to predict is the amount of energy recovered in the NTL case. With this fundamental change, we aim at improving our system by focusing on learning better patterns that generalise better on unseen data, as we explain below:

– By breaking the NTL/non-NTL binary representation of the NTL case, we implicitly indicate to the system that it should focus on learning patterns from high NTL cases whose profile should have clearer abnormal consumption feature values (e.g. low consumption during the last year).

– Moreover, we avoid learning patterns from over-represented customers in the observational data due to business-related decisions (e.g. the recidivist customers) if it does not entail greater energy recovery.

If we look again at the example in Fig. 2, using the energy to recover as the target variable means that the system is going to learn the important pattern first rather than the other.

The two most typical regression Loss Functions are the Root Mean Square Error (RMSE), i.e.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

and the Mean Absolute Error (MAE), defined as

$$MAE = \frac{1}{n}\sum_{i=1}|y_i - \hat{y}_i|$$

The difference between the RMSE and the MAE loss function is the square of the errors, i.e. the higher errors have more weight in the RMSE (as exemplified in Fig. 3). Therefore, the RMSE fits better in ranking problems, in recommender systems, or in our purpose of learning patterns from the higher NTL instances from our training dataset.

4.2 Experiments: classification versus regression benchmarking in real data

In this subsection we compare both the classification and the regression model for NTL detection and confirm the expected benefit of using regression when the organisation's aim is to recover energy without visiting too many customers.

4.2.1 Preliminaries

Data For the experiments, we will use four different datasets from two regions (*A* and *B*), with two different tariffs (1, the most common tariff for houses and apartments in Spain, and tariff 2, an equivalent tariff to 1 but with hour price discrimination. The regions are anonymous to protect the privacy of the data.).[6] The customers must have less than 10kwh of Contracted Power to be on these tariffs. The domain D_{A1} (i.e. the customers from region *A* and tariff 1) has more than 1,000,000 customers, and domain D_{B2} has less than 50,000 customers. The other two datasets fall between these two datasets in terms of population. The proportion of the NTL cases in each domain is lower than 5%. We have around 300,000 labelled instances for the D_{A1} domain, several thousand cases for D_{A2} and D_{B1}, and several hundred cases for D_{B2}.

Model For the classification and the regression predictions, we have trained two different CatBoost models. Each model is trained using the same 80% of the positive instances and 80% of the negative instances. We split in half 20% of instances left, keeping the positive/negative ratio, to build the validation dataset (i.e. the data used to tune the model), and the test dataset (i.e., the training, the validation and the test dataset are stratified). The random partition is chosen over considering the timestamp (e.g. the last 10% of NTL cases as the test dataset) to guarantee diversity and reduce the differences between the datasets due to company decisions. To avoid overfitting, the metric used for early stopping to establish the optimal number of trees is the Average Precision Score for the classification model and the RMSE for the regression model. Both models use the same customer profile, with the only difference that for the classification approach we use a binary target (NTL/non-NTL), while in the regression approach we use the amount of energy to recover (information that, as we explain in Sect. 3, is provided by the technician when an NTL is detected).

Benchmarking A good benchmarking metric to use if we aim at recovering more energy in our campaigns is the Normalized Discounted Cumulative Gain ($NCDG_n$) (Järvelin and Kekäläinen 2002). It is a measure of ranking quality that evaluates our output's correctness with a value between 0 and 1 (1 being the perfect order of the

[6] A tariff with price discrimination involves charging a different price for the electricity depending on when the electricity is consumed. More specifically, electricity would be cheaper at night but more expensive during the day. The potential customer of this tariff is the customer that has an electric car and charges it at night.

NTL cases, and 0 otherwise). This metric allows us a global vision of the correctness of the predictions made, without considering one specific threshold (i.e. the top 100 customers): in many cases, the number of customers to be included in a campaign is unknown when the campaign is being built.

The $NCDG_n$ is defined as

$$NDCG_n = \frac{DCG_n}{IDCG_n}$$

where DCG_n is defined as

$$DCG_n = \frac{\sum_{i=1}^{n} Rel_i - 1}{log_2(i+1)}$$

Rel_i being the relevance (i.e. the score in the ranking, in our case the amount of energy recovered), and $IDCG_n$, i.e. the *ideal DCG*, corresponds to a perfect ordered DCG for the top n elements of the list.

In addition to the $NDCG_n$ metric, we use the amount of energy recovered from the top n scored customers to compare approaches. In both cases we provide four different results (i.e. four different n threshold values): n = (NTL cases in test)/2, n = (NTL cases in test)/5, n = (NTL cases in test)/10 and n = (NTL cases in test)/25; each threshold aims to represent different types of campaigns: from very small campaigns where just a few customers are visited to big campaigns where hundreds of customers are included in the campaign.

4.2.2 Benchmarking results

In Table 1 we report the comparison, in terms of energy recovered and *NDCG* metrics, for the regression and classification approach in the four datasets (for each n threshold).

In terms of *NDCG*, the regression models always score better than the classification models, meaning that the regression approach is able to order better the test customers according to its consumption. Therefore, we recover more energy at the very top of the list, confirming in terms of benchmarking its superiority over the classification approach. This superiority is especially true for small campaigns, where the *NDCG* value for the classification approach is extremely low.

In terms of energy recovered, the regression approach is superior to the classification approach; the amount of energy recovered in our results is usually higher than the energy recovered with the classification models, especially for small-sized campaigns. Recovering more energy is the desired outcome: accumulating very high NTL cases at the very top of the list would allow the company to generate more fruitful campaigns.

With large or medium-sized campaigns, the benefits in terms of *NDCG* and energy recovered of the regression approach is not as clear as in small-sized campaigns, as we can see in Fig. 4: the regression model ranks higher the high-NTL cases (i.e. the NTL cases in which more energy can be recovered, in purple and in red) in comparison to the classification model, but then this advantage fades slightly, and the energy recovered by both approaches becomes more similar.

Table 1 The table at the top compares classification and regression in terms of energy recovered (i.e. the kWh recovered in each threshold n)

Energy recovered from an n-sized campaign (kWh)				
Domain D_{AN}	**n = 528**	**n = 211**	**n = 106**	**n = 42**
Reference	1112625	798198.3	582480.8	366088.1
Classification	434531	196407	97659	37838
Regression	468496 (+7%)	267121 (+36%)	164814 (+69%)	73092 (+93%)
Domain D_{AD}	**n = 186**	**n = 74**	**n = 37**	**n = 15**
Reference	362877.4	273622.2	204201.6	139045.6
Classification	164509	68391	39941	8704
Regression	151844 (-8%)	96520 (+41%)	70022 (+75%)	54988 (+532%)
Domain D_{BN}	**n = 79**	**n = 31**	**n = 16**	**n = 6**
Reference	146245.7	102029.7	75141.3	46690.3
Classification	50596	22164	10079	3542
Regression	67163.9 (+33%)	25764.2 (+16%)	15148.2 (+50%)	12595.2 (+256%)
Domain D_{BD}	**n = 19**	**n = 7**	**n = 4**	**n = 1**
Reference	46482.3	31957.3	22607.3	7555
Classification	16799	7472	5975	2691
Regression	14036 (-16%)	11370 (+52%)	8679 (+45%)	5484 (+104%)

Ranking quality from an n-sized campaign (NDCG)					
Domain D_{AN}	**NDCG**	**$NDCG_{528}$**	**$NDCG_{211}$**	**$NDCG_{106}$**	**$NDCG_{42}$**
Classification	0.52	0.25	0.16	0.11	0.07
Regression	0.57	0.32	0.26	0.23	0.18
Domain D_{AD}	**NDCG**	**$NDCG_{186}$**	**$NDCG_{74}$**	**$NDCG_{37}$**	**$NDCG_{15}$**
Classification	0.43	0.25	0.15	0.11	0.05
Regression	0.65	0.46	0.44	0.45	0.49
Domain D_{BN}	**NDCG**	**$NDCG_{79}$**	**$NDCG_{31}$**	**$NDCG_{16}$**	**$NDCG_{6}$**
Classification	0.45	0.22	0.15	0.11	0.08
Regression	0.47	0.29	0.19	0.16	0.17
Domain D_{BD}	**NDCG**	**$NDCG_{19}$**	**$NDCG_{7}$**	**$NDCG_{4}$**	**$NDCG_{1}$**
Classification	0.47	0.30	0.24	0.25	0.36
Regression	0.59	0.39	0.43	0.46	0.73

As we can see in the results, the regression approach can recover more energy than classification in most cases. In several cases, the amount of energy recovered is significantly greater, especially when the n threshold is small. This means more efficient campaigns in economic terms. The table at the bottom provides a similar analysis, comparing the campaigns in terms of $NDCG_n$. In this analysis, the regression results always outperform classification results in ranking performance (i.e. sorting the customers according to their NTL)

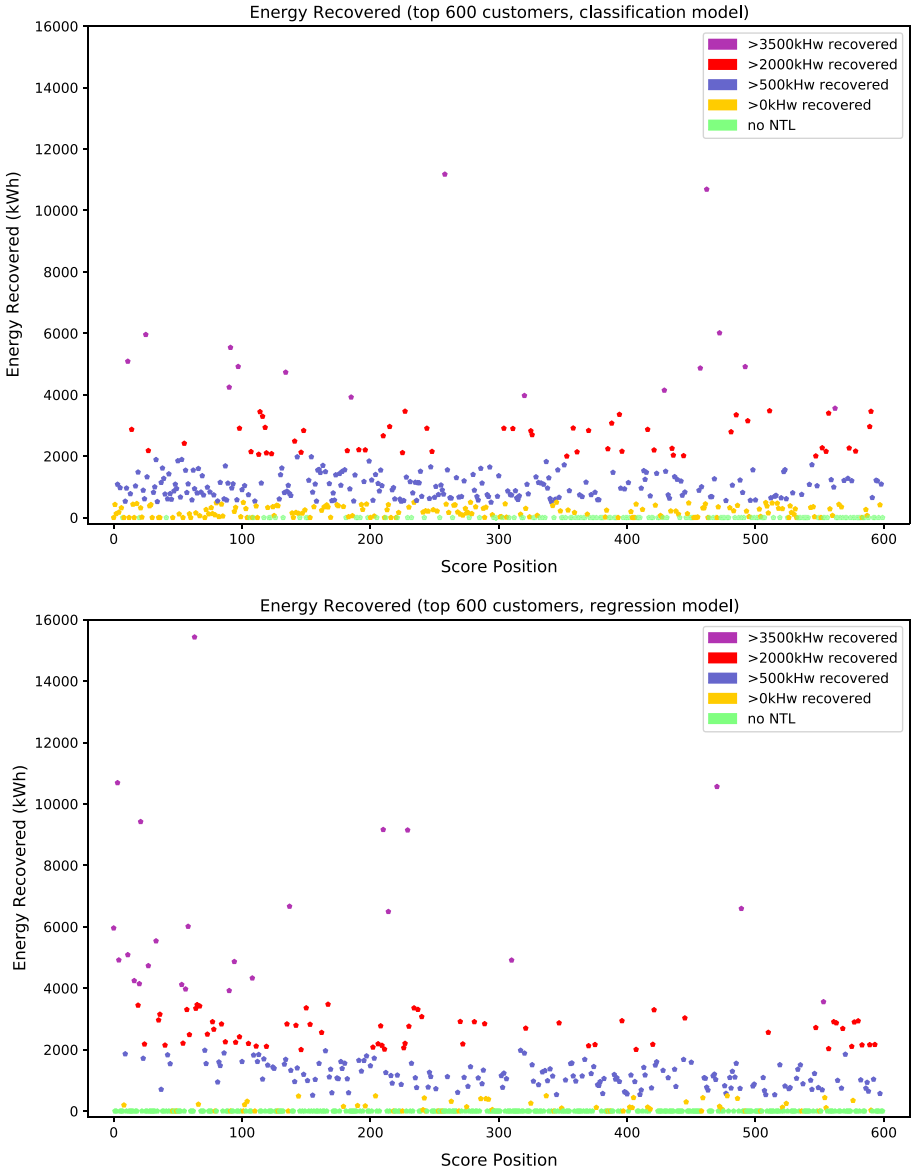


Fig. 4 The results obtained in Table 1 are confirmed in these images: the regression model recovers more energy at the very top of the test prediction list. More specifically, we can see how the purple cases (NTL cases with more than 3500 kWh, the average customer's energy consumption per year) in the regression model are recovered at the very top of the rank

5 Analysing NTL detection beyond benchmarking

5.1 Classification versus regression in terms of explainability

The results from Sect. 4.2 suggest that the regression models recover more energy than classification. However, as explained in Sect. 3.2.1, we are not confident with only benchmarking

our models: increasing the accuracy in validation sets that are subsamples of biased labelled instances does not guarantee that the system is fair (i.e. the system is unbiased against a particular type of customers, e.g. customers from poorer regions), and robust (the system will perform as expected in reality, learning causal patterns, with no Data Leakage Kaufman et al. 2012 nor Dataset Shift Quionero-Candela et al. 2009). The regression approach should be humanly validated as a better method (e.g. learn better patterns) than the classification approach. The purpose of this section is to illustrate this through explanatory algorithms.

The first explanatory algorithm tested in our system was the Feature Importance method. This approach was useful for us to detect biases (e.g. by detecting features that were not indicators of NTL but were too important in the model), but only provided a global vision of the model, with no possibility of analysing the importance of the features on specific customers with a high score. For this reason we explored the use of LIME to explain our predictions at instance level. As we explain in Coma-Puig and Carmona (2018), we were able to implement a rule-based double-checking method in campaigns to discard customers for whom, despite a high score, the explanation obtained from LIME was undesired (e.g. the patterns explained by the local model would not be validated by a human expert). Despite the good results we did not implement LIME as our explanatory algorithm due to the well-known problems of robustness (e.g. Alvarez-Melis and Jaakkola 2018) because of the random component of the algorithm but also the difficulty of having an optimal configuration.

After these two initial unsatisfactory approaches, we started to use SHAP (more specifically, the TreeSHAP implementation Lundberg et al. (2018) to obtain the Shapley Values from Tree Models). According to our experience, the TreeSHAP was the optimal approach to obtain an explanation from a Tree Model because of the following advantages summarised below:

- Consistent global and local explanations: SHAP provides like LIME local explanations but also a consistent global explanation like Feature Importance, since the Shapley values of each instance are the "atomic unit" of the global interpretations. Moreover, it maintains the feature dependence from the model trained.
- Robustness: SHAP always provides the same explanation for the same Tree Model, in contrast with LIME that includes randomness that makes the whole approach look unreliable.
- Reliability: The explanations obtained using SHAP are based on a solid theory and distribute the effects fairly based on the analysis of the original model trained. On the other hand, LIME surrogates the original model and, therefore, it can use features in the local interpretable model not used in the original model.
- Informativeness: The explanation from SHAP provides a very extensive explanation of how the model learnt, allowing the stakeholder and the scientist to be properly informed to support their decisions.
- Low computational cost: Although the computational cost of the Shapley Values are very high, the computational cost for the TreeSHAP is low (i.e. $O(TLD^2)$, T being the number of trees of the ensemble model, L the maximum number of leaves in any tree and D the maximal depth of any tree).

In the next section we will analyse both classification and regression from the Shapley Values' perspective for the case of NTL detection.

5.2 Experiments: classification versus regression explainability in real data

5.2.1 Preliminaries

Data, Classification and Regression Algorithms For the experiments of this section, we use the classification and regression model from Sect. 4.2 for the D_{A1} domain. Similar conclusions can be drawn for the rest of the domains.

Shapley Values and interpretability To analyse the goodness of our model, we use the *summary_plot* method from SHAP. This method provides two plots for our type of problem (i.e. tabular data): a bar chart that represents the mean of each Shapley Value of each feature, and a more complex plot that indicates how each value influenced (i.e. increased or decreased the prediction made from the base value). Both plots can be seen in Fig. 5, applied on the classification approach. Regarding the second plot, in red there are the higher values of the features and, in blue, the lower values. When the feature is categorical there is no colour scale and all the dots are grey. For example, in Fig. 5 we can see that, on average, *Current Reading Absences* is the variable that contributes the most to the prediction, increasing the prediction when the value is high (i.e. the customer has had reading absences). In contrast, when there is no reading absences (i.e. *Current Reading Absences* = 0, in blue), the Shapley Value is 0 or negative.

It is necessary to remark that when Shapley Values correspond to the regression model, they can be read directly as the apportion to the standard output. In contrast, in the binary classification the Shapley Value corresponds to the log odds ratio.[7] Moreover, it is necessary to clarify that the red/blue feature value representation is not valid for categorical features. In these cases, SHAP plots the dots in grey. Hence, Shapley Values on regression have the additional characteristic of being simpler to interpret.

Considerations regarding subjectivity in the analysis As it is widely analysed in the literature, the supervised methods only detect correlations, hence human supervision is necessary to validate them as reliable causal patterns (or, at least, reliable correlation in the company's context). For this reason, the following model comparison from Sect. 5.2.2 requires a human analysis of the Shapley Values and therefore includes subjective considerations.

In general, a reliable pattern would consist of a correlation between a feature value x_i and the prediction $\hat{y}$ that a stakeholder would trust. For instance, the stakeholders could easily validate patterns indicating that the customer is consuming less than expected based on their previous consumption or in comparison to other similar customers. A doubtful or questionable pattern would consist of those patterns that either cannot be easily validated by the stakeholders or whose interpretation is counter-intuitive (e.g. a correlation between a long period of average consumption and a high NTL score).

All these considerations are properly explained, in the following analysis, based on our experience in campaigns. In any case, we provide a fairly generic analysis that fits in most domains similar to the one used in this experiment. We try to avoid very complex analyses that could require information from the company (e.g. the historical NTL cases in specific towns) that cannot be disclosed.

Features referenced in the experiments The features referred to in this section are described in Table 2. For each model, we analyse in depth 8 features to ensure the

[7] That is, x being the sum of the base value and the Shapley Values from an instance, we would obtain the probability between 0 and 1 by doing $1/(1+\exp(-x))$.

Table 2 Features referred to in the experiments with their descriptions

Feature	Definition
Current reading absences	After the installation of smart meters the company can remotely communicate with the meters. The absence of the meter readings can indicate either an incident in the meter (e.g. that it stopped working) or fraudulent manipulation. This feature indicates how many months have passed since the last meter reading
Last visit: correct/fraud	Categorical information that indicates if the last visit done to that customer has been correct or an NTL has been detected. If the customer has not received any visit, the feature's value is empty
Town	The town where the customer lives
# Meters in property	How many meters the customer has in property. In general, the meter is owned by the company, and is rented/handed over to the customer
Date last reading	How many months have passed since the last meter reading
Last 'no fraud' visit	How many months have passed since the last time the customer was visited with a visit whose aim was not to detect fraud, i.e. the installation might not actually be checked during the installation
Min/max bill last 12 months	The ratio between the minimum and maximum bill during the last 12 months
Contracted Power	The contracted power by the customer. In general, it is expected that a customer with a high consumption needs a higher contracted power
Cons. zone/cons. last year	Ratio between the consumption of the customer and the average in the zone from the same type of customers. A zone is an internal reference that refers to a group of customers that receive the electricity from the same point of supply. Therefore all the customers in the zone are very similar and have similar energy needs
Last Bill	Last Bill in kWh
Diff consumption 6 months	The difference in terms of kWh between two equivalent periods of time (i.e. the same six consecutive months) in consecutive years. A higher value indicates that the customer has had a consumption reduction
# Months with no consumption	Consecutive months with no consumption until the present
Consumption penultimate year	Consumption of the customer in the penultimate year (i.e. from 24 to 12 months ago)

readability of the document. However, we also provide a more generic description of the model that includes more information beyond the 8 features at the end of the analysis.

5.2.2 Evaluation analysis through explainability

According to Fig. 5 and our interaction with the company's technicians, we cannot trust the classification model since there is only one consumption-related feature in the top eight most important features (the *Min/Max bill last 12 Months*, a feature that refers to the ratio between the minimum and maximum consumption bill in the last year). Instead, many of the features are visit related (features that, as exemplified in Fig. 3, can be useful but can also produce bias and other learning problems).

For a deeper analysis we can analyse the effect of each value on the output with the bottom plot from Fig. 5:

– *Reliable patterns*: In the classification model, several patterns can be easily confirmed as true indicators of NTL:

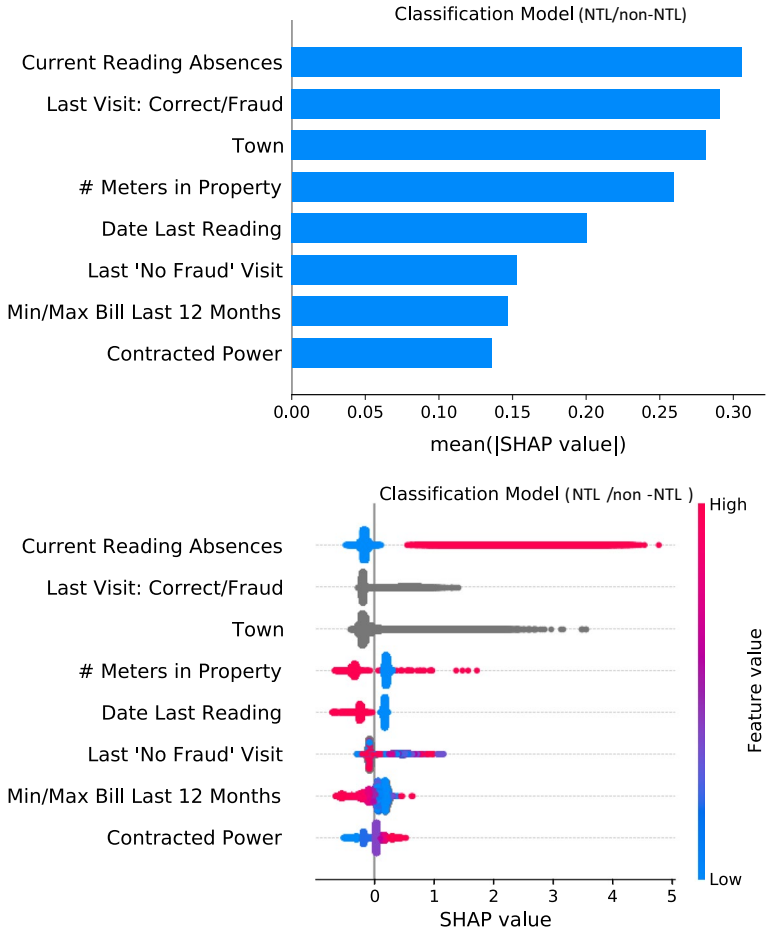


Fig. 5 SHAP explanation of the classification approach: there is only one consumption-related feature on the top 8 most important features. Moreover, how each feature influenced in the score assignation is not easy to interpret: only the *Current Reading Absences* can be fully trusted as a good pattern and, for this reason, we cannot validate the model as a good and robust model

1. *Current Reading Absences* This feature is the most important feature for the model (according to SHAP). This is a very reliable pattern learnt because the company expects to have, after the introduction of smart meters, information from the meter on an ongoing basis, including meter readings. The lack of meter readings is for sure a very suspicious behaviour since it may indicate meter manipulation.
2. *Contracted Power* According to the Shapley Values there is a correlation between a higher contracted power and a higher probability of committing NTL. This pattern can be a bias since the company usually tends to include customers with higher Contracted Power in the campaigns. However, the company validated this pattern based on their experience.
3. *Min/Max Bill Last 12 Months* We can see that, in general, the model considers a lower value more related to NTL behaviour. We consider this pattern valid because, in general,

we expect that monthly consumption will not vary in a very marked way during the year. If this occurs, it may be a consequence of meter tampering.

- *Categorical information* Two categorical features (with no colour scale in Fig. 5 bottom) are very relevant in our system, as we explain as follows:

1. *Last Visit: Correct/Fraud* This information is valuable since the patterns learnt should be contextualised to the visits carried out by the company. That is, a customer that committed Fraud in the past is, according to the company, very likely to commit fraud in the future.
2. *Town* The town where the customer lives can be a good indicator for the NTL detection system. Statistically, there are towns in which the company has always detected more NTL cases than in other towns.

- *Unknown interpretability* The interpretation of how a feature value influences the output can be hard to understand for the classification approach. Several examples are given below:

1. *# Meters in Property* When a customer owns a meter, it is more likely to be in an inaccessible location. Therefore, it would be easier for the customer to manipulate it. Moreover, having more than one meter increases the possibilities of having an NTL. Therefore, one would expect that a high feature value would correspond to a high Shapley Value. However, a high value in this feature influences unevenly on the output. With this information the stakeholder might not draw conclusions about the feature role in the prediction or its correctness.
2. *Last 'No Fraud' Visit* Several interpretations can be expected for this feature. For instance, a recent visit combined with a high electricity consumption can confirm that a customer is not committing NTL, but also a recent visit to a customer that is consuming less than expected can be suspicious. The lack of context harness the interpretation of the feature by the stakeholder.

- *Questionable pattern* Finally, there is a pattern learnt from a feature that the stakeholder cannot validate:

1. *Date Last Reading* According to the SHAP value, low values (i.e. the last meter reading is recent) is more related to the NTL behaviour. At first glance, this pattern is unintuitive since we would expect a similar pattern to the one learnt from the *Current Reading Absences* a recent reading would indicate that the meter is working as expected. A possible explanation for this unexpected output might be the correlation between the *Current Reading Absences* and the *Date Last Reading*: the model is already learning the expected pattern from the *Current Reading Absences*, and therefore the role of the *Date Last Reading* becomes unstable. Another option would be that the system is detecting

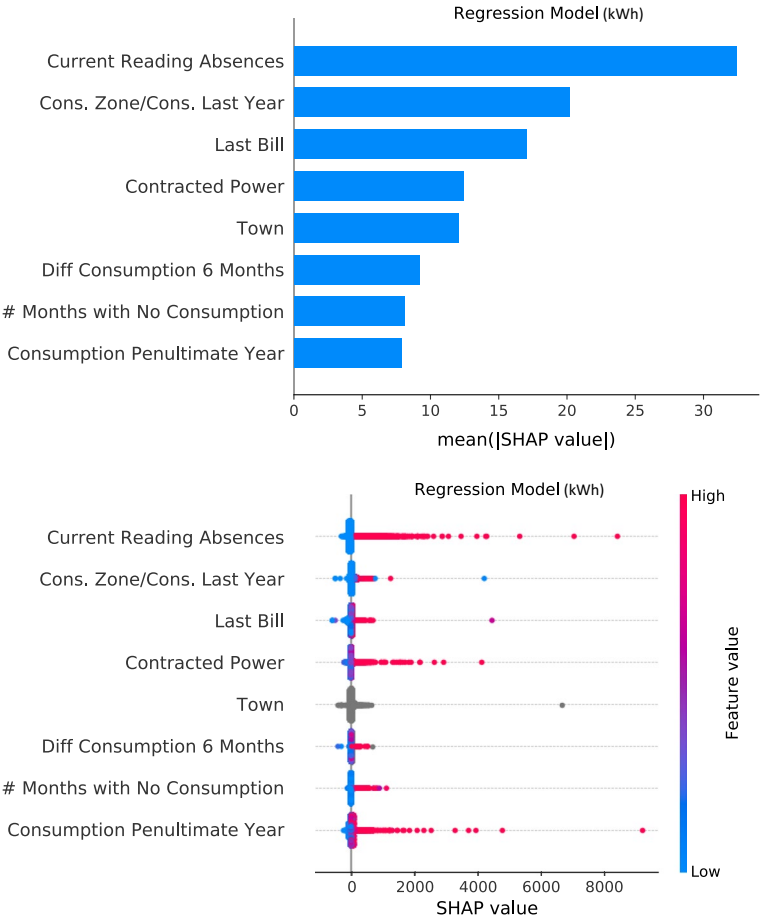


Fig. 6 The regression model relies on consumption features to learn patterns and, therefore, we can consider that this model is better than the binary approach. Moreover, the patterns learnt seem to be easier to understand by the stakeholder, since more abnormal behaviours (the absence of meter readings or the number of months with no consumption) are more clearly related to a higher prediction than in the classification model, where lesser patterns can be easily trusted as trustworthy indicators of NTL

an unexpected NTL pattern (e.g. a technician makes a manual meter read, detects an abnormal behaviour and informs the company that the meter should be checked, and therefore there exists in the next few days another technician visit that confirms the NTL case).

Despite several aspects of the model being reliable in terms of NTL detection, the model relies on very few consumption features in the prediction process. This can be problematic in terms of robustness and fairness since the consumption features are better NTL predictors.

Instead, the regression model shown in Fig. 6 is more robust, as it uses more consumption-related features, and it is easier to validate, as we explain as follows:

– *Reliable consumption patterns* In comparison to the classification model, the consumption features are the most relevant in the model:

1. *Cons. Zone/Cons. Last Year* Since we are comparing similar customers in terms of Tariff and region, we would expect that fraud corresponds to low consumption. This feature has learnt this pattern and, therefore, we consider it correct.
2. *Diff Consumption 6 Months* A high value indicates that in the past the customer consumed more than in the present. Therefore, the pattern learnt that a high value increases the output of the prediction and therefore should be considered reliable and correct.
3. *# Months with No Consumption* if the customer has several months with 0 kWh of consumption, it should be considered as a probable case of NTL, especially in populated regions and cities where there are not as many empty homes as in rural regions (at least in Spain).
4. *Consumption Penultimate Year* A high electricity consumption two years ago is not in itself a clear pattern of fraud. Nevertheless, it can be a very good complementary feature that indicates a change in consumption behaviour. For instance, a customer who has always had low consumption is not the same as a customer who consumed in the past a lot and has recently changed their consumption behaviour.

– *Reliable patterns from the binary model* Two important features in the classification model remain important in the regression model:

1. *Current Reading Absences* As explained in the previous analysis, the absence of meter readings is a likely indicator of NTL.
2. *Contracted Power* The contracted power was also considered a very important feature in the classification approach. However, in the regression approach, the use of this feature makes more sense: in the regression model we are trying to maximise the amount of energy to recover and, in general, the customer with a higher contracted power consumes more energy.

– *Categorical information* Only one categorical feature is in the top important features in the regression model:

1. *The Town feature* In comparison to the binary approach, the Town feature seems to have less relevance. However, we can see one specific Town value whose Shapley Value is much higher than the other towns. This town corresponds to a small municipality where the company recovered a lot of energy in the past, and therefore it can be trusted.

– *Doubtful/Questionable pattern* Finally, we consider that there is one pattern in Fig. 6 that the stakeholder cannot fully understand:

1. *The Last Bill* According to SHAP, a high value is learnt by the model as an indicator on NTL. The classical NTL behaviour consists of manipulating the meter to avoid high bills and, therefore, we would expect the opposite behaviour regarding this feature. However, there are circumstances in which a high last bill can be correlated with an NTL case:

– A recidivist fraudulent customer that has been visited twice in a short period of time. The high bill corresponds to the back-payment of the previous fraud detected.
– A customer with very high consumption that is not normal (e.g. illegal drug cultivation) that combines a correct installation of electricity with an illegal junction to get enough power.

In any case, these cases are more exceptional than the classic examples of reduced consumption and should therefore not be a pattern that is so prominent in the system.

This in-depth analysis of each model through their most important variables faithfully represents each model. For instance, the classification model only has 3 consumption features in the top 15 most important features, and 7 consumption features in the top 25 most important features according to Shapley Values, while the regression approach has 10 and 19, respectively. In addition to that, it is tangible (as we have explained for each variable) that the patterns from the regression model are easier to analyse and corroborate by the stakeholder. This is true because as we have analysed variable by variable, in the regression model, we can interpret what NTL patterns have been detected in that variable in a much simpler way. In classification, such analysis requires much more effort (the stakeholder cannot easily interpret what the pattern learnt by the model is), and the conclusions are often nuanced or unclear.

5.3 Customer selection through local explainability

5.3.1 Preliminaries: local explanation as sanity check

In Sect. 5.2.2 we have seen that the increase in energy recovered in Sect. 4.2 is justified because the regression model learns better patterns from the Stakeholder's perspective than the classification model. The resulting system is more robust since it learns less circumstantial patterns (e.g. fewer patterns related to the company's decision that highly influence the observational data). Thus, the challenges regarding the lack of robustness and the low energy recovered per campaign generated are mitigated. Nonetheless, we can see in Table 1 that the system has room for improvement. That is, the system does not provide a perfect ordering of the customers according to NTL. Moreover, in Fig. 4, we can detect that still, some non-NTL cases (or NTL cases with a very low amount of energy to recover) have a high score. In Coma-Puig and Carmona (2018) we propose a solution to reduce the number of these undesired high-scoring customers with low or no NTL: to analyse through LIME the local explanation of each high-scoring customer included in the campaign, discarding those that, according to human knowledge, the explanation obtained is not reliable. Therefore, the final selection is a subset of the original sample.

In this section we propose an updated version using the local explanations of the Shapley values instead of LIME. This change of explanatory algorithm has two significant advantages. On the one hand, the Shapley Values provide local explanations consistent

Table 3 The post-processing at instance level (by not including those customers whose most important fraudulent feature according to the Shapley Values is not a consumption-related feature), referred to in the table as *Regression + Rule*) reduces the size of the selection but increases the amount of energy to recover on average for each visit

Domain D_{AN}	$n = 528$	$n = 211$	$n = 106$	$n = 42$
Average energy recovered per customer in an n-sized campaign (kWh)				
Reference	2107	3782.9	5495.1	8716.4
Classification	823	930.8	921.3	900.9
Regression	887.3	1266	1554.8	1740.3
Regression + Rule	944 (+ 6%)	1398.4 (+ 10%)	1741.5 (+ 12%)	2328.7 (+ 34%)

More specifically, 31 out of 42 customers, 84 out of 106, 173 out of 211 and 469 out of 528 customers would be included in the final campaigns, but in each case, we would increase the amount of energy recovered per customer visited, a clear indicator that this post-process would discard more non-NTL cases (or NTL cases with low energy recovered) than otherwise. That is, we increase the economic efficiency of our campaign, recovering more energy per visit carried out by the technician

with the global explanation of the model since the global explanation is constructed as the sum of the local explanations. On the other hand, the solid theory behind Shapley Values (particularly the implementation for TreeSHAP trees) provides us with robust explanations (i.e. the explanations obtained for a model and prediction are always the same).

This sanity check has points in common with the analysis proposed in Sect. 5.2.2, where we analyse the correctness of the modular explanations. However, a good modular explanation does not guarantee that all the explanations at instance level of the top-scored customers are also reliable. Similarly, just because the model has learned a reliable and important fraudulent pattern at the modular level (e.g. a feature that, on average, greatly increases the prediction score) does not guarantee that all high-scoring customers have learned that pattern. Having said that, a good modular explanation, as it is built as the sum of the local explanations, should be an indicator of good explanations at instance level.

5.3.2 Post-process example

By way of illustration of this method, this example implements a simple rule system that automatically discards all the high-scored instances in which the most important fraudulent pattern (i.e. the feature value that increases more the prediction according to the Shapley Values) is not consumption-related. This is in line with the modular analysis from Sect. 5.2 in which we regard the regression model as a better predictor because the most important features are consumption-related.

This post-process approach aims to increase the campaign's economic efficiency by increasing the amount of energy recovered per customer visited. Therefore, we compare in Table 3 the amount of energy recovered for each customer on average in an n-sized campaign,[8] for the Domain D_{AN}. As expected, we can see in Table 3 that the regression approach outperforms the classification approach in terms of energy recovered per customer visited. However, our post-processing at instance level implemented in the regression approach outperforms the regression approach by up to 34%.

[8] n corresponds to the customers preselected for the campaign, as explained in Table 1.

In this example, we have used a straightforward rule to provide a rather generic example. However, this approach is very useful to nuance the campaign based on the Stakeholder's knowledge. For instance, as we explained in Sect. 4, one of the existing biases is related to the fact that the company generates campaign to over-control historically fraudulent customers. From our perspective, this pattern is valuable and trustworthy since many fraudulent customers are recidivists. However, we would like to avoid high-scoring customers with only this pattern as an indicator of NTL. Therefore, this post-process method would be helpful to discard these specific high-scoring customers that would not be humanly validated.

6 Conclusions

6.1 Positioning of our work in the literature

This work introduces an NTL detection system grounded on regression as a valid alternative to using classification. Moreover, we illustrate the use of explanatory algorithms to understand the predictions of the system. Experiments performed indicate that using the energy recovered as the priority setter helps the system be more successful, mitigating the biases problems regarding the use of observational data. The patterns learnt are easier to validate from a human perspective, and therefore the models generalise better. Surprisingly, the use of regression in the NTL literature is scarce. For instance, (Krishna et al. 2015) describes an outlier detection system, where the amount of energy to be spent by a customer is forecast. We believe our approach can be enhanced by using the techniques in the aforementioned work.

On the other hand, this work is one of the few examples in the literature that implements explanatory algorithms for NTL detection. Our experiences and lessons learnt can be useful not only for any initiative that aims at increasing interpretability but also for any data-oriented industrial project.

6.2 Future work

For our research project this work is the starting point from which to develop different improvements in our system, explained below.

Possibilities regarding the Regression and Classification approach This work proposes and evidences that the use of a regression approach (RMSE loss function) to detect NTL has benefits in terms of robustness and economic efficiency. This is an initial approach that has been satisfactory. However, we are considering the use of other loss functions that could also fit this problem, e.g. the use of a ranking loss function (e.g. LambdaMart Burges 2010) or a more complex regression function where the over-representation of the non-NTL cases are considered (e.g. Tweedie regression Zhou et al. 2019).

Similarly, we would explore the possibility of exploiting the information from the classification models, not so much as the basis of the predictive system, but as a complementary method for the regression approach. Currently, we are making a smooth transition from the classification to the regression method, including in the campaigns both scores and including in campaigns customers that have both a high regression and classification score. Our future effort consists of building a smart meta-scorer based on the combination of both pieces of information.

Exploiting the information from the Shapley Values The use of Shapley Values in this work focuses on analysing two different approaches to detect NTL cases beyond benchmarking, allowing us to confirm what the results in the table seemed to indicate: that the regression approach provides better models than classification. Moreover, in Sect. 5.3, we explained how we could use the Shapley Values to post-process a campaign to improve its accuracy and energy recovered per customer visit. However, the SHAP library provides many tools to analyse the models that we have not yet used in our system.

One of the methods that could be useful is the *interaction values* from SHAP that provide a plot representing the pairwise interaction between two features. This plot could be extremely useful, for instance, analysing the relationship between the *Date Last Reading* and the *Current Reading Absences* from the classification model, two features that, as we explained in Sect. 5.2.2 are correlated and can justify the abnormal pattern learnt from the *Date Last Reading* feature.

Using the Shapley Values as a pre-processing technique The process of building a predictive machine learning model usually includes a feature selection process (to avoid, for instance, overfitting). In our case, the Gradient Boosting models were trained with the features explained in Sect. 3, and internally the training process would select those features that would minimise the loss function, discarding the non-informative features in the splitting process. In each model, the features used vary, i.e. we automatically learn patterns from the non-static domains at that moment.

However, as previously explained in Sect. 3.2.1, this process relies on benchmarking of biased data and, therefore, might learn undesired patterns that exploit biased information. For instance, in Sect. 5.2.2 the pattern learnt from the *Last Bill* feature for that specific model is doubtful (even though the patterns usually learnt by the system using this feature are reliable). For this reason, we propose the Shapley Values as a method to implement feature selection since it would allow us to determine the goodness of the feature in terms of patterns learnt in every model trained beyond benchmarking. A first approach to automatise the exploitation of the Shapley Values as a feature engineering tool can be seen in Coma-Puig and Carmona (2020). In this work we exploit the stakeholder's knowledge and the information provided by explanatory methods to implement online smart feature engineering (similar to a query learning process). This initial approach has been useful to build more robust models and will be improved based on our experience after using it in different campaigns.

Similarly, the Shapley Values can also be used as a search method to detect the optimal tuning process since it is a process that relies on benchmarking different models in a validation dataset. For instance, in Sect. 4.2.1 we explain that we use the validation dataset as an early stopping tuning process. However, by analysing the patterns learnt, we might detect that the optimal number of tree iterations might differ from the number of iterations established by the benchmark analysis.

Improve the interpretability of the Shapley Values Finally, it is necessary to remark that the interpretation of the Shapley Values might not be straightforward for the stakeholder. In this work we only analyse the top 8 most important features, but we consider it necessary to increase the number of features analysed. This, altogether that we aim to explore techniques to exploit the Shapley Values makes us pose the need for implementing an ad-hoc method that simplifies the information provided to the stakeholder.

Author Contributions Conceptualisation, B.C.-P. and J.C.; Methodology, B.C.-P.; Software, B.C.-P.; Validation, B.C.-P.; Formal Analysis, B.C.-P.; Investigation, B.C.-P.; Resources, J.C.; Data Curation, B.C.-P.;

Writing–Original Draft Preparation, B.C.-P., Writing–Review and Editing, B.C.-P. and J.C., Visualization, B.C.-P., Supervision, J.C., Project Administration, J.C.; Funding Acquisition, J.C.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This work has been supported by the Ministry of Science and Innovation under the Grant PID2020-112581GB-C21, and a collaboration with Naturgy.

Availability of data and material Due to the nature of this research, the company funding this study (Naturgy) did not agree to the data being shared publicly for legal reasons, so no data is available.

Code availability Due to the nature of this research, the company funding this study (Naturgy) did not agree to the code being shared publicly for contractual reasons, so the code is not available.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Consent to participate Not applicable.

Consent for publication Not applicable.

Ethics approval Not applicable.

References

Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. arXiv preprint arXiv:1806.08049.

Angelos, E. W. S., Saavedra, O. R., Cortés, O. A. C., & de Souza, A. N. (2011). Detection and identification of abnormalities in customer consumptions in power distribution systems. *IEEE Transactions on Power Delivery, 26*(4), 2436–2442. https://doi.org/10.1109/TPWRD.2011.2161621

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

Badrinath Krishna, V., Weaver, G. A., & Sanders, W. H. (2015). Pca-based method for detecting integrity attacks on advanced metering infrastructure. In J. Campos & B. R. Haverkort (Eds.), *Quantitative Evaluation of Systems* (pp. 70–85). Springer.

Burges, C. J. (2010). From ranknet to lambdarank to lambdamart: An overview. *Learning, 11*(23–581), 81.

Buzau, M. M., Tejedor-Aguilera, J., Cruz-Romero, P., & Gómez-Expósito, A. (2018). Detection of non-technical losses using smart meter data and supervised learning. *IEEE Transactions on Smart Grid*. https://doi.org/10.1109/TSG.2018.2807925

Cabral, J. E., Pinto, J. O., Martins, E. M., & Pinto, A. M. (2008). Fraud detection in high voltage electricity consumers using data mining. In *2008 IEEE/PES transmission and distribution conference and exposition* (pp. 1–5). IEEE.

Coma-Puig, B., & Carmona, J. (2018). A quality control method for fraud detection on utility customers without an active contract. In *Proceedings of the 33rd annual ACM symposium on applied computing, SAC '18* (pp. 495–498). ACM, New York, NY, USA. https://doi.org/10.1145/3167132.3167384.

Coma-Puig, B., & Carmona, J. (2019). Bridging the gap between energy consumption and distribution through non-technical loss detection. *Energies*. https://doi.org/10.3390/en12091748

Coma-Puig, B., & Carmona, J. (2020). An iterative approach based on explainability to improve the learning of fraud detection models.

Coma-Puig, B., Carmona, J., Gavalda, R., Alcoverro, S., & Martin, V. (2016). Fraud detection in energy consumption: A supervised approach. In *2016 IEEE international conference on data science and advanced analytics (DSAA)* (pp. 120–129). IEEE.

Costa, B. C., Alberto, B. L., Portela, A. M., Maduro, W., & Eler, E. O. (2013). Fraud detection in electric power distribution networks using an ann-based knowledge-discovery process. *International Journal of Artificial Intelligence & Applications, 4*(6), 17.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning* (pp. 233–240).

Depuru, S. S. S. R., Wang, L., Devabhaktuni, V., & Green, R. C. (2013). High performance computing for detection of electricity theft. *International Journal of Electrical Power & Energy Systems, 47*, 21–30.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning.

Drummond, C., & Japkowicz, N. (2010). Warning: statistical benchmarking is addictive. Kicking the habit in machine learning. *Journal of Experimental & Theoretical Artificial Intelligence, 22*(1), 67–80.

Ford, V., Siraj, A., & Eberle, W. (2014). Smart grid energy fraud detection using artificial neural networks. In *2014 IEEE symposium on computational intelligence applications in smart grid (CIASG)* (pp. 1–6). https://doi.org/10.1109/CIASG.2014.7011557.

Galanti, R., Coma-Puig, B., de Leoni, M., Carmona, J., & Navarin, N. (2020). Explainable predictive process monitoring.

Glauner, P., Meira, J. A., Valtchev, P., State, R., & Bettinger, F. (2017/01). The challenge of non-technical loss detection using artificial intelligence: A survey. *International Journal of Computational Intelligence Systems, 10*, 760–775. https://doi.org/10.2991/ijcis.2017.10.1.51.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS), 20*(4), 422–446.

Kadurek, P., Blom, J., Cobben, J., & Kling, W. L. (2010). Theft detection and smart metering practices and expectations in the netherlands. In *2010 IEEE PES innovative smart grid technologies conference Europe (ISGT Europe)* (pp. 1–6). IEEE.

Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD), 6*(4), 15.

Krishna, V. B., Iyer, R. K., & Sanders, W. H. (2015). Arima-based modeling and validation of consumption readings in power grids. In *International conference on critical information infrastructures security* (pp. 199–210). Springer.

León, C., Biscarri, F., Monedero, I., Guerrero, J. I., Biscarri, J., & Millán, R. (2011). Integrated expert system applied to the analysis of non-technical losses in power utilities. *Expert Systems with Applications, 38*(8), 10274–10285.

Liu, Y., & Hu, S. (2015). Cyberthreat analysis and detection for energy theft in social networking of smart homes. *IEEE Transactions on Computational Social Systems, 2*(4), 148–158.

Lundberg, S., & Lee, S.I. (2017) A unified approach to interpreting model predictions.

Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.

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

Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., & Burke, R. (2020). Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management* (pp. 2145–2148).

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019) A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635.

Messinis, G. M., & Hatziargyriou, N. D. (2018). Review of non-technical loss detection methods. *Electric Power Systems Research, 158*, 250–266.

Molnar, C. (2019). *Interpretable machine learning*. Retrieved May, 2020 from https://christophm.github.io/interpretable-ml-book/.

Monedero, I., Biscarri, F., León, C., Guerrero, J. I., Biscarri, J., & Millán, R. (2012). Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees. *International Journal of Electrical Power & Energy Systems, 34*(1), 90–98.

Nagi, J., Yap, K. S., Tiong, S. K., Ahmed, S. K., & Mohamad, M. (2009). Nontechnical loss detection for metered customers in power utility using support vector machines. *IEEE Transactions on Power Delivery, 25*(2), 1162–1171.

Nagi, J., Yap, K. S., Tiong, S. K., Ahmed, S. K., & Nagi, F. (2011). Improving svm-based nontechnical loss detection in power utility using the fuzzy inference system. *IEEE Transactions on Power Delivery, 26*(2), 1284–1285. https://doi.org/10.1109/TPWRD.2010.2055670

Nizar, A., Dong, Z., & Wang, Y. (2008). Power utility nontechnical loss analysis with extreme learning machine method. *IEEE Transactions on Power Systems, 23*(3), 946–955.

Papa, J. P., Falcao, A. X., & Suzuki, C. T. (2009). Supervised pattern classification based on optimum-path forest. *International Journal of Imaging Systems and Technology, 19*(2), 120–131.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pereira, L. A. M., Afonso, L. C. S., Papa, J. P., Vale, Z. A., Ramos, C. C. O., Gastaldello, D. S., & Souza, A. N. (2013). Multilayer perceptron neural networks training through charged system search and its application for non-technical losses detection. (pp. 1–6). https://doi.org/10.1109/ISGT-LA.2013.6554383.

Posada-Quintero, H. F., Molano-Vergara, P. N., Parra-Hernández, R. M., & Posada-Quintero, J. I. (2020). Analysis of risk factors and symptoms of burnout syndrome in Colombian school teachers under statutes 2277 and 1278 using machine learning interpretation. *Social Sciences, 9*(3), 30.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2017). Catboost: Unbiased boosting with categorical features.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2009). *Dataset shift in machine learning*. The MIT Press.

Ramos, C. C., Souza, A. N., Chiachia, G., Falcão, A. X., & Papa, J. P. (2011). A novel algorithm for feature selection using harmony search and its application for non-technical losses detection. *Computers & Electrical Engineering, 37*(6), 886–894.

Ramos, C. C. O., de Sousa, A. N., Papa, J. P., & Falcao, A. X. (2011). A new approach for nontechnical losses detection based on optimum-path forest. *IEEE Transactions on Power Systems, 26*(1), 181–189. https://doi.org/10.1109/TPWRS.2010.2051823

Ramos, C. C. O., Rodrigues, D., de Souza, A. N., & Papa, J. P. (2018). On the study of commercial losses in brazil: A binary black hole algorithm for theft characterization. *IEEE Transactions on Smart Grid, 9*(2), 676–683. https://doi.org/10.1109/TSG.2016.2560801

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.

Saria, S., & Subbaswamy, A. (2019). Tutorial: Safe and reliable machine learning. arXiv preprint arXiv: 1904.07204

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games, 2*(28), 307–317.

Souza, R., Rittner, L., & Lotufo, R. (2014). A comparison between k-optimum path forest and k-nearest neighbors supervised classifiers. *Pattern Recognition Letters, 39*, 2–10.

Spirić, J. V., Dočić, M. B., & Stanković, S. S. (2015). Fraud detection in registered electricity time series. *International Journal of Electrical Power & Energy Systems, 71*, 42–50.

Tsymbal, A. (2004). The problem of concept drift: Definitions and related work. *Computer Science Department, Trinity College Dublin, 106*(2), 58.

Xiao, Z., Xiao, Y., & Du, D. H. C. (2013). Exploring malicious meter inspection in neighborhood area smart grids. *IEEE Transactions on Smart Grid, 4*(1), 214–226.

Yapo, A., & Weiss, J. (2018). Ethical implications of bias in machine learning. In *Proceedings of the 51st Hawaii international conference on system sciences*.

Zhou, H., Qian, W., & Yang, Y. (2019). Tweedie gradient boosting for extremely unbalanced zero-inflated data. arXiv preprint arXiv: 1811.10192