Fraud Detection in Energy Consumption: A Supervised Approach

Bernat Coma-Puig*, Josep Carmona*, Ricard Gavaldà*, Santiago Alcoverro† and Victor Martin†
*Universitat Politècnica de Catalunya
Barcelona, Spain
{bcoma,jcarmona,gavalda}@cs.upc.edu
† Gas Natural Fenosa Distribución
Barcelona, Spain
{salcoverro,vmartinf}@gasnatural.com

***Abstract*—Data from utility meters (gas, electricity, water) is a rich source of information for distribution companies, beyond billing. In this paper we present a supervised technique, which primarily but not only feeds on meter information, to detect meter anomalies and customer fraudulent behavior (meter tampering). Our system detects anomalous meter readings on the basis of models built using machine learning techniques on past data. Unlike most previous work, it can incrementally incorporate the result of field checks to grow the database of fraud and non-fraud patterns, therefore increasing model precision over time and potentially adapting to emerging fraud patterns. The full system has been developed with a company providing electricity and gas and already used to carry out several field checks, with large improvements in fraud detection over the previous checks which used simpler techniques.**

I. Introduction

Utility companies provide an essential service to developed societies, supplying electricity, gas and water to homes, businesses and factories. The necessary infrastructure that guarantees services includes from the kilometers of pipes or lines that transport the energy to the millions of meters that monitors consumption of individual customers. An important problem that these companies face is the imbalance between the energy billed with respect to the energy provided, called *energy losses*. *Non-technical losses* is a widely used, somewhat euphemistic name including fraud and meter malfunctions among other.

Methods for committing fraud include splicing the pipes to bypass the meter, tampering with the meter to stop it or to slow it down, and simply connecting to the distribution network without even having a contract with the company or a meter. On the other hand, an accidental malfunction of the meter also results in net energy loss for the company. For brevity, along this paper, we will use the term fraud to refer to both forms of non-technical loss, intentional fraud and meter malfunction: The company actions and revenue after the detection will be very different[1], but both are issues that the company wants to detect and fix as early as possible, and the detection schemes are essentially the same.

Most cases of fraud involving meter tampering or malfunctioning can be detected by direct inspection by a trained technician. However, it is extremely expensive to send technicians to inspect a large number of meters. Therefore, companies usually perform a pre-selection of a subset of meters to be directly verified by technicians in a given period of time and area, a concept that we call a *campaign*. Every customer visit has a cost, so in order to be worthwhile campaigns need to have relatively high precision (i.e., percentage of problems detected with respect to the number of meters verified). Company gains, on the other hand, are directly proportionally to campaign recall (i.e., fraction of the existing fraud that is detected), so campaign design is all about the classical precision-recall tradeoff.

Traditionally, campaigns are based on simple sets of rules indicating fraud (e.g., abrupt decrease of consumption, or no consumption during a long period of time). These rules can be used to detect the fraudulent/irregular customers, but achieve a low success rate, not much higher than selecting customers for the campaign at random. This can be explained by many other reasons besides the fraud (e.g. customer spends a long convalescence in the hospital, or the house is a second residence that does not follow the consumption patterns of a all-year home). It is very natural to try to exploit data from the past to design better campaigns, including statistical and machine-learning based techniques.

This paper introduces a dynamic supervised approach to detect fraud and irregularities in a energy utility company. Overall, we encode the detection of fraud as a classification problem, where supervised techniques over the set of historic cases of fraud are applied. Remarkably, the system continuously evolves its knowledge by incorporating the results of each campaign in order to improve model quality and obtain more efficient campaigns. Our development has been based on real datasets from a utility company and has already served several campaigns of different kinds (city and country, electricity and gas). As we will argue, the approach proposed is a significant step towards automating the detection of fraud and irregularities in the consumption of energy, when compared to the traditional rule-based approach that nowadays utility companies apply. But it also improves on most machine

[1] In case of detecting a fraud, the current regulation in Spain establishes billing the consumption corresponding to six hours per day during a year, if the real energy consumed during the fraud period cannot be determined.

learning approaches described in the literature, which build a classifier offline on a single dataset to be employed for a long time.

The paper is organized as follows: Section II highlights some related work and the differences with ours. In Section III we describe the gathering and processing of the data. In Section IV we provide a detailed description of the architecture of the system. Then in Section V we report the results on actual campaigns. Finally, Section VI presents some conclusions and sketches future work.

II. Related Work

Detection of fraud or irregularities (the *non-technical losses* or NTL) has received considerable attention in the last decades; see for instance [10], [4] for the particular case of electricity. In this section we focus on recent approaches [13], [11], [12], [3], [5], [2], [9], [6], [7].

Several perspectives have been taken when facing the problem of NTL in electricy. They range from game-based computational models [2], combinations of statistical techniques [3], supervised machine learning [12], [5], [6], [7], and time series [13], [11]. Works such as [9] focus specifically on the important process of feature selection. Differences with our work include:

- Existing work focuses on a single source of energy (electricity). Since the company providing our study case distributes both electricity and gas, we exploit the correlations among consumptions to improve detection in either, when possible. This feature has been incorporated recently, but preliminary studies on correlation indicate that it helps in detecting fraud.
- Most related work considers a single source of consumption information, often the one obtained from smart meters. In this work we combine consumption data from different sources, in particular reliable information from smart meters and nonreliable information (estimated or self-reported) for users having older, non-smart meters – which will be prevalent in large parts of the world for a while.
- In most related work, a classifier or combination of classifiers is created after analysis of a single historical dataset, and built into the detection system. Improving the models requires a human analyst analyzing more or newer data. We implement a continuously running system that "closes the loop": The results of campaigns are given as feedback to the system, in order to incrementally retrain the models. The feedback also suggests specific campaigns, not imagined a priori, upon discovering interesting niches of customers with higher-than-average fraud. The system might use drift management methods to adapt to changes in customer behavior and fraud methods.

III. Data Processing

Gas Natural Fenosa, from now on "the company" is a utility company distributing both electricity and gas in Spain and 26 countries of 5 continents. The project started in mid 2013 when the company approached UPC researchers looking to improve the rate of fraud and anomaly detection of their current campaigns, which were based on a set of simple heuristic rules. The information potentially available was consumption records for any subset of their customers for several years, some static information of each customer, and a limited number of reports corresponding to verified fraud cases in the past. This information, its strengths and limitations is described in this section.

The project gradually evolved from the original goal of "mining" one specific dataset to create a one-time campaign to the development of a software system that connects to the company's operational system for generating both routine campaigns and on-demand parameterized campaigns, managing feedback, and investigating the usefulness of new features, among other functionalities.

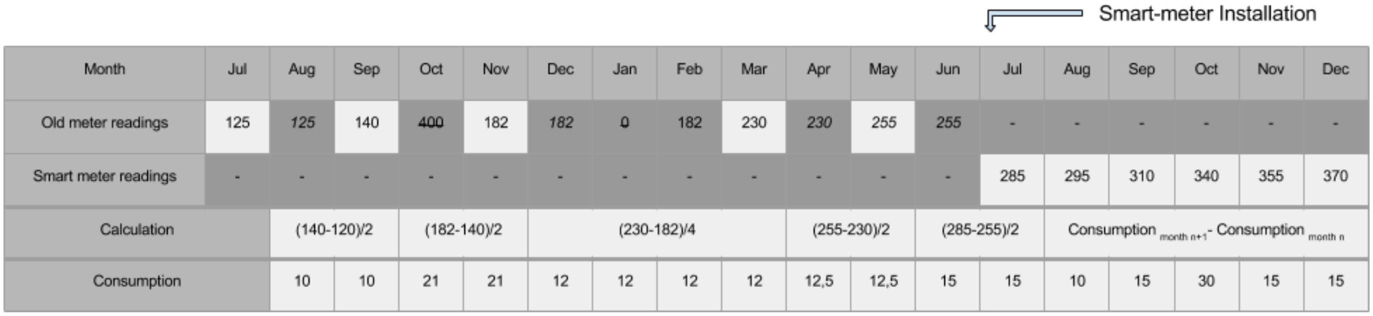
A. The Datasets

The first dataset received from the company was to be used to generate a static campaign (no software system). It came from a medium-size city with a few thousand electricity customers and another few thousand gas customers, and contained contract information (e.g., the tariff or the age of the meter) and about two years of consumption information. After this initial dataset, we received four more datasets with similar characteristics (i.e. cities with thousands of customers), either of gas or electricity or both. From the sixth dataset onwards, and in view of the good results of the initial campaigns, we were given access to data at the country scale, with information of several million customers in gas and another several million of customers in electricity; around a million customers are customers of both utilities. The last dataset used comprises the time period from June 2012 in electricity and from November 2012 in gas to today.

The sources of data used so far in the project include consumption and profile data (with a total of almost 200 millions of monthly information in electricity, and around 225 millions of monthly information in gas), historical fraud cases, and some external information.

a) Consumption data: This is the data reflecting the energy used by the customers. It includes meter readings as well as billing extractions - the invoices the company charges to the customer based on the meter readings or, when not available, an estimate by the company based on historical information. In this paper, we call a *consumption reading* the difference between consecutive meter readings of the customer; they will be our main source to extract the consumption data; Figure 1 is an example. Note that from this information one cannot detect the fraud cases in which someone connects directly to the distribution network rather than e.g. manipulating the meter; this type of fraud can only be detected further upstream.

The origin and reliability of consumption data is varied. In electricity, about half the customers have smart meters that send customer consumption to the utility company monthly and reliably. The other customers have old meters that require



Smart-meter Installation

Month	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Old meter readings	125	*125*	140	~~400~~	182	*182*	~~0~~	182	230	*230*	*255*	*255*	-	-	-	-	-	-
Smart meter readings	-	-	-	-	-	-	-	-	-	-	-	-	285	295	310	340	355	370
Calculation		(140-120)/2		(182-140)/2		(230-182)/4				(255-230)/2		(285-255)/2		Consumption $_{\text{month n+1}}$- Consumption $_{\text{month n}}$				
Consumption		10	10	21	21	12	12	12	12	12,5	12,5	15	15	10	15	30	15	15

Fig. 1: Example customer consumption record. In light gray, useful information (new readings and monthly consumption). In dark gray, non-useful information (i.e. erroneous information is crossed out, in italic repated readings, with a dash months without readings). Non-smart meters give less acurate monthly consumptions; in this case, the first October reading is obviously wrong, so apportioned consumption because we have new readings every two months. Since the second July, consumption readings are reliable because a smart meter was installed.

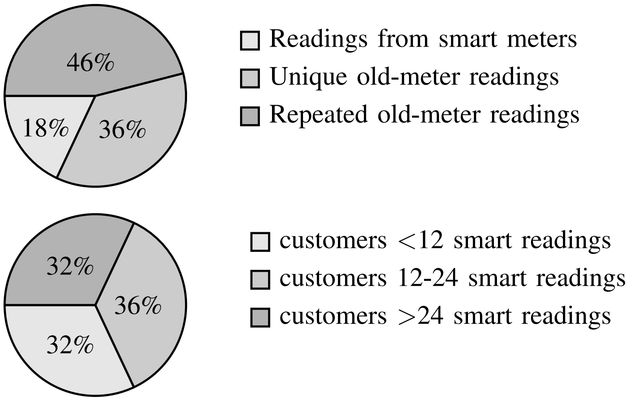


Fig. 2: Half the customers for electricity have smart meters. From all the readings we have, 18% come from the smart meters, 36% are unique readings from old meters and 46% are repeated. Of the customers that have smart meters, around 32% have less than 12 smart-meter readings, 36% have between 12 and 24 readings and the rest have more than 24 readings.

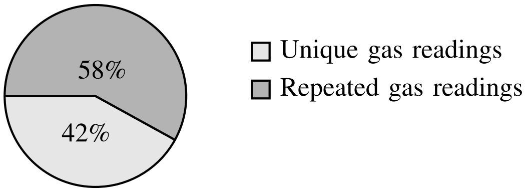


Fig. 3: 42% of the readings from gas are unique, and 58% are repeated. In gas, the absence of smart meters makes the information available less reliable.

manual readings: Customers are expected to use any of a number of options to send the reading (calling the company, sending the reading via mail or a mobile app, or writing it down on a paper available in the building). In gas there are no smart meters yet, so all the information is sent manually.

When there are no smart meters and the customer has not provided the reading, the company needs to estimate the consumption of the customer using reference values from similar historical periods. Customer-generated readings are notoriously unreliable and error-prone, and many customers simply do not send any for many months in a row. Eventually a technician will be sent to read the meter, but only after several months of estimated readings. Fortunately, each consumption record is labeled as company verified, customer provided or estimated, so we can assign them different reliabilities.

Erroneous, absurd or missing readings for non-smart meters, the co-existence of two metering systems with different reading periodicity (1 vs. 2 month between readings), and the fact that some customers changed from one system to another in the process (a small proportion of clients have more than two years of smart-readings) were complications that we had to deal when reading, parsing, and standarizing consumption data.

b) Static Profile Data: Within this concept we include information related to the contract of the customer with the company (e.g. the tariff), information of the customer (e.g. their address) and characteristics of the equipment (e.g. the age and model of the meter, whether it is inside or outside the house). This information can be used to categorize the customer; for example, the tariff indicates whether it is a home, a shop or a restaurant, and whether a gas cooker or central heating is present.

c) Historical Fraud Cases: For the first campaign, we were also provided a list of customers who had commited fraud and were discovered in baseline campaigns carried out by the company in the recent past years. The list included several tens of thousands of verified fraud cases which could be joined with their consumption data to investigate fraud patterns. No “negative cases” (i.e., clients that were known to not commit fraud) were received.

d) External Information: We used as well the Koppen climate classification data of the different regions under study, as climate obviously affects energy consumption patterns, as well as census data about socioeconomic classification of cities and regions.

B. Creating a Classification Problem

We now comment on the three main issues we had to tackle when transforming the data into a classification problem.

a) Unreliability in Consumption Data: As discussed before, consumption data as given by the readings may not be an exact picture of the consumption of the customer but an approximation, because the heterogeneous channels used to obtain the readings. For instance, in Figure 1 one can see that the reading in February is wrong, according to the readings in January and March. A small number of heuristics were designed to correct or discard suspicious or inconsistent data, both in gas and electricity (e.g. discard a reading smaller than the previous and the following reading, or discard an absurdly high reading for a month, replacing them with interpolations). Billing information is also used, but as a secondary source compared to actual or estimated readings.

b) Statistical evaluation of the features: The main metric used to evaluate the features is the odds ratio (the odds that an outcome will occur given a particular case, in comparison to the odds that the outcome will occur otherwise), from now on denoted as *OR*. For each feature, we analyze the OR between the results from the feedback (e.g. the OR between the fraudulent clients against the non-fraudulent clients), denoted as *ORPN*, as well as the OR between the fraudulent clients against all the clients not included in any campaign, denoted as *ORPG*. Table I shows the odds ratios of some features used.

c) Feature construction: Consumption, profile, and external information was combined to create a number of features. These features are the result of an evolution of the initial rules used by the company for their baseline campaigns (e.g. the detection of customers with a long period of time with no consumption). Yet, each rule used in the baseline campaigns used one or at most two of the features, while we expected machine-learning-based algorithms to be more accurate by taking into account hundreds of features as well as their combinations.

Variables	*ORPG*	*ORPN*
Abrupt decrease of consumption	18.6	3.4
Long period of low consumption	6.2	3.0
High consumption discrepancy	10.4	2.4
High range higher lower consumption	12.2	2.0
Gas consumption without electricity	11	2.9

TABLE I: Significant odds-ratio of some features from the electricity campaigns. Both the odds-ratio between the fraudulent and non-fraudulent customers (ORPN) and the odds-ratio between the fraudulent customers and the customers not included in any campaign (ORPG) are included.

Some features focus on the behavior of the customers in comparison to themselves: An abrupt or gradual decrease in their consumption exemplified in Figure 4, a repeated lack of reported readings, a substantially different pattern from previous years, consumption peaks (as seen in Figure 7), the difference between the minimum and the maximum consumption of the customer, etc..

Feature	*Definition*
Abrupt decrease of consumption	A reduction of $x\%$ in consumption during n months in comparison to the previous n months.
Abrupt decrease of equivalent consumption	A reduction of $x\%$ in consumption during n months in comparison to the same months from the previous year.
Long period of low consumption	A reduction of $x\%$ in consumption during n months in comparison to the average.
Consumption discrepancy	High/medium/low discrepancy between a consumption in comparison to the average.
Decrease of consumption (correlation)	A consumption reduction during n months using Pearson correlation.
Tariff	Tariff of the customer.
Location of the meter	Location of the meter (e.g. inside/outside the house)
Contracted power	Contracted power (only in electricity).
Consumption peak	Consumption in a month $x\%$ times higher/lower than the previous and the following month.
Electric tension	Electric tension (only in electricity).
Abnormal contractual status	The client has abnormal contractual status (e.g. has canceled the contract with the company).
Regional income	Whether the customer lives in a region with an average income above, similar or below the Spainish average.
Climatology	Köppen climatology classification of the region where the customer lives.
Billing/consumption similarity	Similarity between the consumption computed and the billing (only in gas).
Reading periodicity	Reading periodicity (1 or 2 months, only in electricity).
Unknown consumption	Number of consecutive estimated readings, consecutive 0 consumption (in the present and historically),
Difference between years	Highest difference between the consumption of two consecutive years, or if the customer has been consuming less year after year (only in gas).
Gas consumption without electricity	Gas consumption without electricity.
Number of readings	High/medium/low number of different readings from the customer.
Province Capital province	The province where the customer lives. Whether the customer lives in a capital of a province.
Date information	Age of the meter, date of installation and contract.
Difference of consumption	High/medium/low difference between the higher and the lower consumption of the customer.
Ratio difference of consumption	High/medium/low difference between the average consumption of the customer and the average consumption.
Negative Consumption	Whether the customer has a negative consumption (e.g. count was reset at the installation of a new meter).
Reading correction	Whether the consumption obtained required a correction.
Smart meters	Whether the customer has a smart meter.
Old fraud	Whether the customer was detected as fraudulent by the company in gas or electricity in the past.

TABLE II: Types of features extracted from the data available. In some cases several versions of the feature are included in the system (e.g. abrupt decrease of consumption, where different $x\%$ and n lead to different features, or using Spanish and regional averages to compute different features).

Other variables measure the inconsistencies of the customer consumption in comparison to other similar customers. For example, long periods of time where the customer consumes much less than the average of the customers with the same tariff (exemplified in Figure 6), high discrepancy between the consumption curve of the customer and the typical consumption curve of the customers with the same tariff (a flat consumption line showing no seasonality pattern may indicate that the metering has been tampered to not exceed a certain metering speed, as seen in 5), nominal difference between the consumption of the customer and the average consumption of the clients with the same tariff, etc.

It is worth mentioning the features that combine information from gas and electricity consumption (e.g. the behavior comparison between the gas and electricity consumption, or the consumption of gas without electricity consumption, exemplified in 8). These features, tested statistically, were only included in the most recent campaigns, which are the more successful ones.

From the static data we also extract some features; the province where the customer lives as well as its climatology, the location of the meter, the date of installation of the meter, etc. Table II contains a list with the feature types included in our system.

These variables were all binarized, for uniformity and simplicity; in particular, this avoids problems with algorithms that are too sensitive to outliers or extreme values. For example, one binary variable was introduced for each tariff type that a customer may have. In some cases, several variants of the same variable were introduced, corresponding to different horizons or thresholds; for example, whether there has been a reduction of $x\%$ in the last n months generates many variables for varying x and n. For each candidate variable, the ORPG was checked, and those variables with values near 1 that were not useful (e.g. did not have profiling information) were removed.

All in all, the number of variables included in the campaigns has been growing over time reaching 250 features in the latest electricity campaign and 150 features in the latest gas campaign.

At the end of this process, a customer is represented (with the information available at the time of generating a campaign) as a vector of binary variables, which we call the *customer profile.*

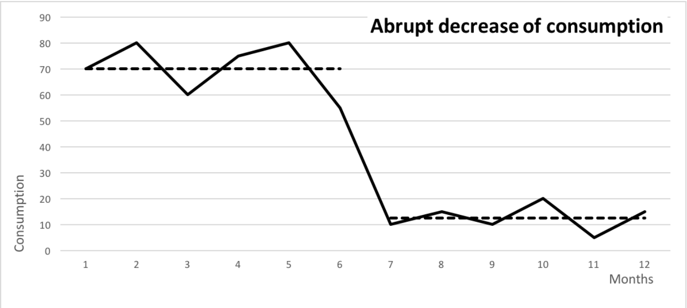


Fig. 4: Example variable: Consumption drop.

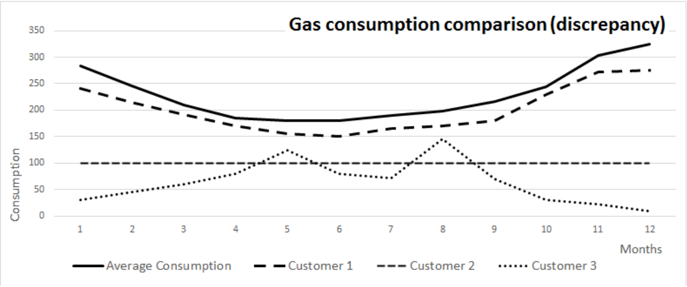


Fig. 5: Example variable: Consumption discrepancy. Customer 1 has a similar consumption curve than the average consumption. On the other hand, both customer 2 and 3 have an abnormal curve that might be an indicator of fraud.

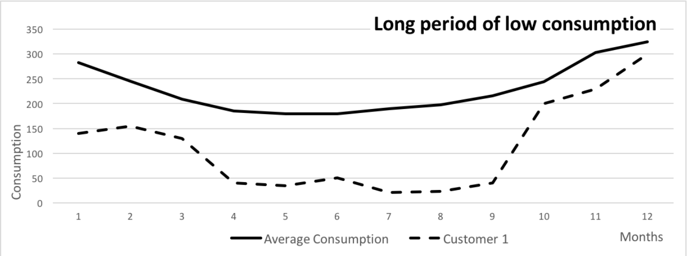


Fig. 6: Example variable: Long period of low consumption. As we can see, we have a period of time (from month 4 to 9) where the customer consumes much less than the average.

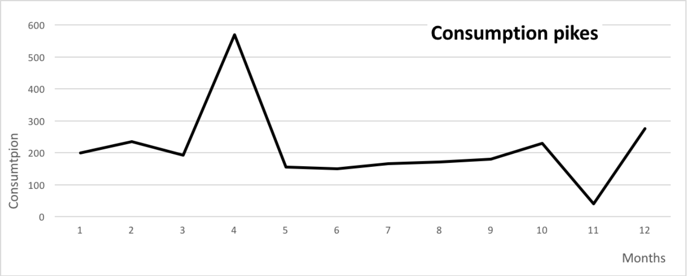


Fig. 7: Example variable: Consumption peaks. We can see that the customer has both a positive peak (4th month) and a negative peak (11th consumption) in its consumption curve.

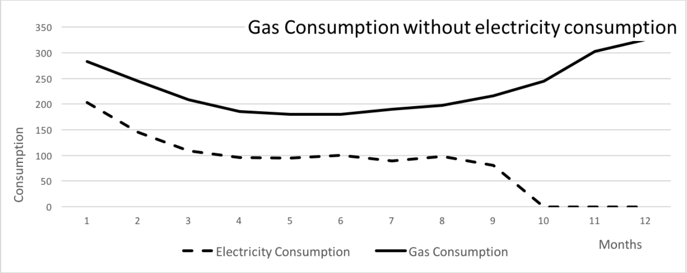


Fig. 8: Example variable: Gas Consumption without electricity consumption. Gas heater needs electricity to work.

d) Imbalanced Classification Problem: We created a classification problem by labeling each customer profile in a potential training dataset either with the positive class (P), representing fraudulent behaviors, or the negative class (N) representing non-fraudulent behavior. For populating P, we considered customers from the historical fraud cases (see Section III-A) and those detected as fraud in previous campaigns. Populating N was a problem in the initial campaign, as we did not have certified non-fraud cases; we simply took a random sample of all customers, which should be approximately correct under the assumption that fraud prevalence is low enough. As we started receiving feedback from the first campaigns, we did have certified negatives.

The prediction desired from the system could be a P/N value. In this case the campaign is simply an unordered set of suspicious customers (predicted to be P). The company, however, preferred to have a *fraud scoring*, or probability of being fraud, for each customer, which makes the campaign an ordered list; this allows to detect in-place that a campaign has entered a point of diminshing returns.

IV. The Process

Our approach to detect fraud is based on supervised learning algorithms that create models based on labeled data. This can be explained as follows (see Figure 9):

1) From the data sources we extract all the necessary data to create the variables.
2) We create a profile of each user, a vector of variable values that defines their behavior up to a certain date.
3) Based on feedback from older campaigns, we run a number of algorithms in a number of configurations to determine which one could be best for this campaign.
4) We use the chosen model to compute a fraud score (a prediction or probability) for every customer in the target area for a new campaign. We exclude customers that have been checked already in recent campaigns.
5) We create a campaign of a desired size N by selecting the N customers with highest fraud score.
6) When the campaign results returns from the field, the feedback (verified fraud and non-fraud cases) is added to the system automatically[2].

This methodology allows iterative learning from the feedback from previous campaigns, as well as the addition of new algorithms and variables. A detailed description of the procedure is given below.

A. Initial Campaign

The first step in our approach is to create a model able to learn the pattern frauds. Figure 10 illustrates this stage.

This step needs to be taken the first time we create a campaign for a new population for which we have no feedback. It is assumed however that some set of verified fraud cases is available from the baseline campaigns by the company, which are labeled P. A set of randomly chosen customers is selected and labeled N, with the understanding that some labels may be wrong. In our case, we needed to do this learning phase on the first campaign for a small city and for the first country-wide campaign.

In all campaigns from the initial one in one population, the training and test sets will contain verified positives and negatives from previous campaigns.

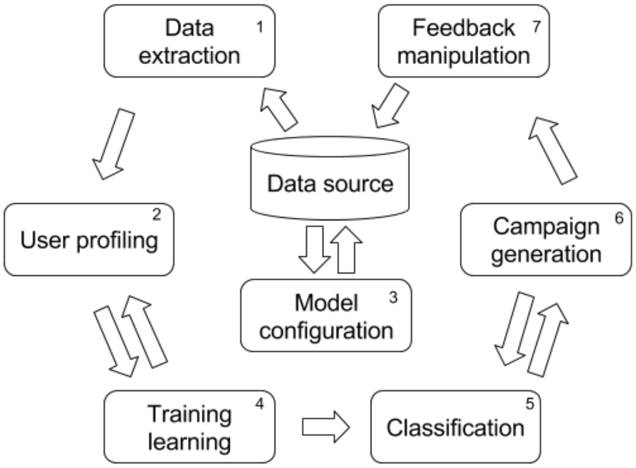


Fig. 9: System architecture. Campaigns are generated iteratively, allowing classifiers to improve based on their performance.

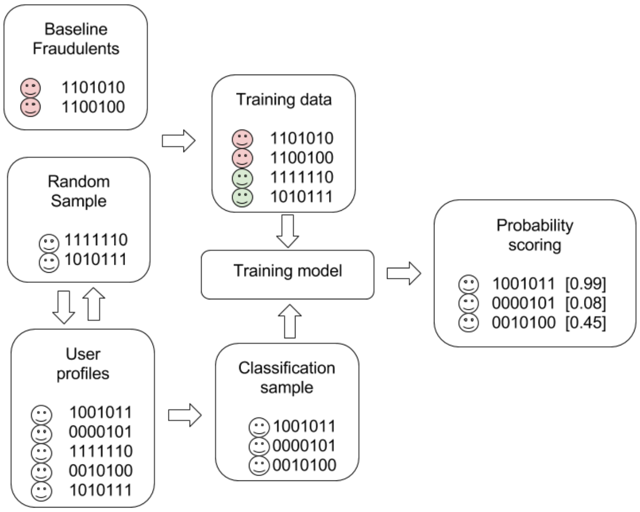


Fig. 10: Generation of the first campaign. A model is created using as positive examples the historical fraudulent customers, and as negative examples a small random sample of the population.

B. The Campaign Phase

After the customers have been selected for a campaign, each customer will have a fraudulent score, and the utility company will select those with higher score that have not been verified

[2] The company centralizes all the campaigns results introduced manually by each technician in a database, and our system reads this information automatically to update the feedback.

recently, and send technicians to check the corresponding meters.

The results for each customer can be summarized as *Fraudulent*, *Non-fraudulent* and *Absent*. Fraudulent are those customers who have committed fraud or whose meter does not mark (i.e. it does not correctly measure consumption). Non-fraudulent customers are those whose meter could be checked and showed no signs of tampering or malfunction. Finally, absent customers are those for which the technician could not have access to the meter. Absent customers are a significant fraction of the campaign feedback. We do not include them in our performance calculation or in the feedback to our system (i.e., are labeled neither P nor N), although it is believed that a fraud among these may be higher than average, because fraudulent customers will try to avoid being inspected. The field reports for the campaign contains a number of distinct codes, some corresponding to malfunctioning meters and some to true fraud; as mentioned, we do not differentiate them in our system and label them all as positive for the feedback.

Finally, the system needs to process the results of the campaign. The profiles of the users from the campaign are stored with its corresponding P/N label. This labeled data will be used as training data in the following campaigns.

C. Learning from Feedback

After feedback has been incorporated in the system, we have additional verified fraud cases (in addition to those coming from the baseline campaigns) and verified non-fraud cases. These can be used now to train new models for further campaigns. Note that we decided *not* to include the baseline fraud cases in the training sets of further campaigns because there was no guarantee that the profiles *at the time the campaign was performed* were indicative of the fraud. The company could have chosen them for inspection based on behavior previous to the records we were given, or based on side information not included in the records. Direct feedback information was considered more reliable.

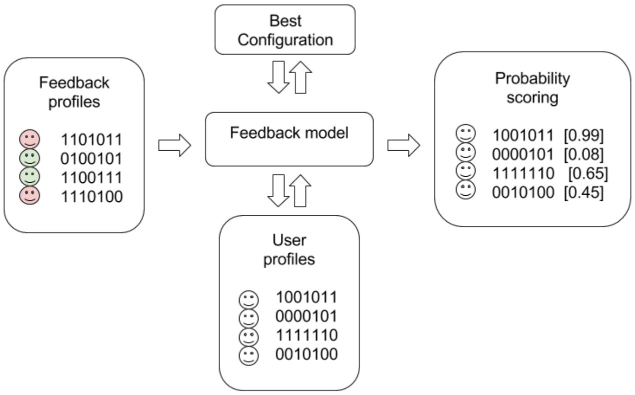


Fig. 11: In the Feedback phase, we study several models using the feedback profiles and choose the best one (or best combination) to assign to each customer a fraud score.

D. Algorithmic Details

Many classifier-building algorithms with different configurations have been tested so far. Those that have up to now contributed to real campaigns include:

- Naive Bayes.
- K-nearest neighbors: Different number of neighbors and distance weight have been tested.
- Decision Tree inducers, including C4.5 and CART: Both Gini and Entropy split criteria as well as the number of features used were tuned.
- Neural Networks with backpropagation training: The learning rate, momentum, epochs, the number of hidden layers and the number of errors allowed have been parameterized.
- Support Vector Machines: both linear and radial basis kernel functions have been tested. We have also tuned the cost for misclassification as well as the gamma (for the RBF kernel).
- Random Forests: the number of iterations, as well as the parameters tuned in the Decision Tree have been modified.
- Gradient descent Decision Tree with CART: besides the optimization applied in the Random Forest, we have also seen how the loss function (deviance or exponential) modifies the performance.
- AdaBoost with C4.5 decision trees, naive Bayes, and PART: The number of iterations have been optimized.

The tools used to implement these algorithms are the Knime Analytics Platform [1] and the scikit-learn Python library [8].

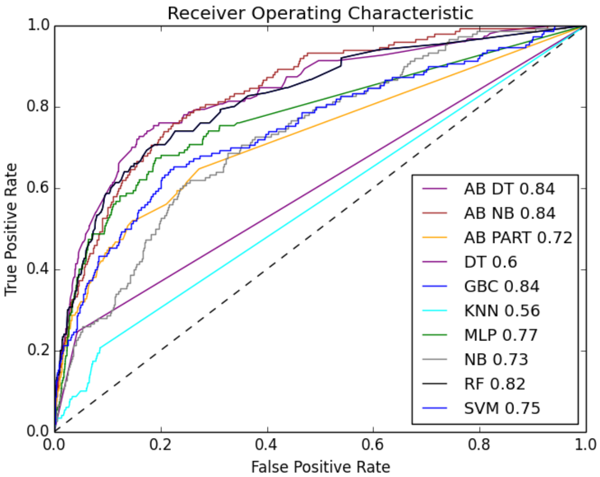


Fig. 12: Area-under-curve values of the algorithms applied to the 4.5x feedback campaign in 4-fold cross validation. The meta-algorithms (Gradient Boosting Decision Tree, Random Forests, and AdaBoost with Naive Bayes) were the top performers.

Initially we created fixed-size campaigns and the scoring of each customer was boolean (predicted fraud or non-fraud), so we optimized f-measure to balance precision and recall.

After the first campaign we wanted to assign a numerical fraud score to each and every customer in order to create a sorted list of all customers; the company could then choose the size of the campaign going down the list as far as desired. The metric to be optimized among configurations was then the Area under Curve, to maximize the position in the list of fraudulent customers rather than P/N hard classifications.

In the initial campaigns we combined several classifiers, with the final scoring being the voting or average among their individual scores. Furthermore, most algorithms can be run in several configurations (e.g. parameter settings). Our system semi-autonomously[3] explores several configurations of each algorithm and several candidate combinations (e.g., including or not each algorithm and assigning voting weights). This model choosing process is not fully automatized, in spite of that it can be automatized using a classification-validation cross-validation process, to facilitate our understanding and post analysis of the results. In our most recent campaigns we have opted for a single Gradient Boosting Model, because we observed that it gave better AUC than any ensemble including other algorithms (see Figure 10).

The campaigns are generated once a month, executed in a commercial computer (i.e. not a cluster). For this reason, we have prioritized the scalability of the software instead of the performance speed. Depending on the input data (i.e. the population or the number of months to compute) the system can last from hardly an our to a day. With a fairly optimization the software speed can be boost easily.

V. Some Results

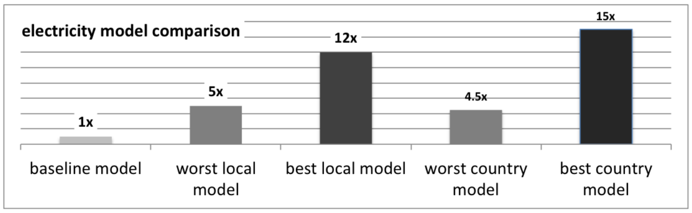


Fig. 13: Comparison between the precision of the baseline model and our methodology in a local population and in all the country in our electricity campaigns.

Our first campaigns were conducted on three medium-size locations (population between 50,000 and 100,000) to experimentally test the efficiency of our methods to detect electricity fraud without investing on large, costly campaigns. It was soon clear that they achieved precision notably better than the baseline.

To be precise, let us take precision as our main criteria, i.e., fraction of fraudulent users discovered among those inspected. The size of the campaigns was equal to the baseline ones, so we did not increase precision by simply inspecting less customers. Then campaigns consisting of randomly chosen customers had precision around $z\%$[4], and company baseline campaigns had essentially no better results. Our three initial campaigns had precision 12x, 5x, and 5x the baseline.

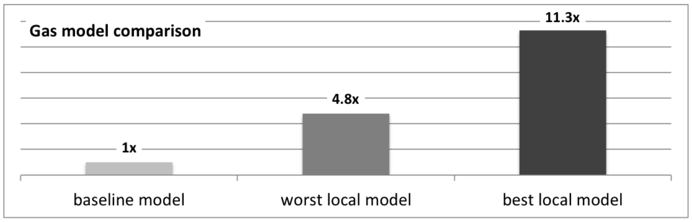


Fig. 14: Comparison between the precision of the baseline model and our methodology in all the country in our gas campaigns.

Encouraged by these results, the company proposed to conduct a test at the country level (Spain) with several million users, also to test the scalability of the approach to detect electricity. We were provided with historical fraud cases from the whole country, and returned the company a list of several thousand of customers sorted by fraud score. The company ran a campaign consisting of the top 10,000, as that was the standard size of their baseline country-wide campaigns. The campaign had precision 4.5x that of the baseline. That is slightly less than the worst score achieved in small populations, but of course much more than the baselines. More interestingly, months later the company has conducted a second campaign taking the next 10,000 customers from the same list, which had lower scores, so a priori lower precision was expected. Surprisingly, precision was again very close to 4.5x the baseline. As a side-effect, the campaigns have provided a dataset of 16,000 customers with reliable fraud/non-fraud labels for further campaigns, the remaining 4,000 being “absent”.

The lower performance of the first country-wide campaign (the one that achieved a 4.5x performance) with respect to the best city-wide ones (12x) merited consideration. We attributed this fact to the higher diversity of customer behavior, energy rate usages, climates, and fraud patterns at the country level. More generally, considering a large user base may blur the patterns that affect only some subsets of customers.

If we break down the data and analyze the results of our campaigns as partial results by the tariff (Figure 15), the performances varies notably depending on the tariff; for example, in the 4.5x campaign, the two most common tariffs were those that achieved best performance). This can be read as:

- The information of the less common customers is blurred by that of the most common customers.
- We have less information from these customers, being more difficult to profile and detect.

To fix all these problems we considered to:

[3] The choice of configuration is not totally automatized to facilitate our understanding and post analysis of the results. This process can be fully automatized using the validation-test case with the feedback information easily.

[4] The exact figure is withheld at the request of the company but it is a small 1-digit percentage.

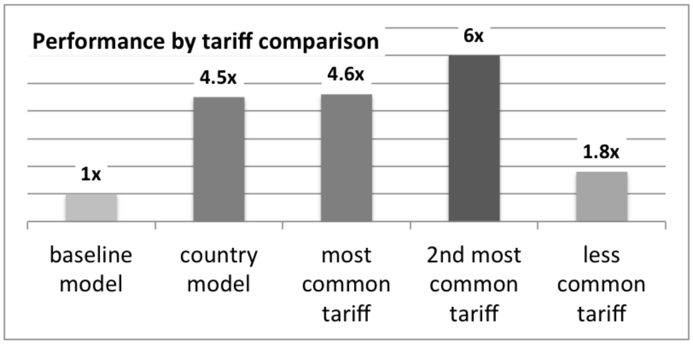


Fig. 15: Reading of the results of the country-wise campaigns. Considering only the tariffs with at least 100 customers in the campaign, the most common tariffs are those easier to detect by our methodology.

- Take geographical proximity into account when computing variables that indicate anomaly with respect to other customers. That is, "typicality" is considered with respect to consumers in the same region (in particular, the same climate).
- Also perhaps diversify the models taking into account tariff, type of contract (home, restaurant, shop…) and other such factors.
- Therefore, develop a complex predictive model structure. Instead of creating a single model for a campaign at the country level, create many local models and patch them together to create a campaign. Of course, controlling overfitting may be a challenge in this approach.

We have started to introduce these solutions to the system. Specifically, we have already introduced segmented features (e.g. comparison consumption between a customer and the customers from the same province). The following country-wide campaign, that included this new information, boosted the global performance and achieved a precision 15x the baseline.

Another issue that required our attention was the evolution of the software over time. It is expected that new fraudulent behavior appears periodically, and in this context the infrastructure proposed in this paper is able to capture them. On the other hand, two factors may prevent our (and similar) techniques to continuously be more accurate:

- the appearance of new types of fraud/irregularities, not captured by the learned models.
- The periodicity and size of the campaigns. The system may tend to degrade the quality over time: when campaigns are applied too often without new information or the size of campaigns is too big.

To solve this problems, we consider to:

- Keep us a small % of the campaigns to explore, using alternative methods (e.g. new features not included in the system) to detect new fraudulent behaviors.
- In this case, the complex predictive model structure explained to improve the results will be useful to structure the periodicity of the campaigns to maximize the results.

We have seen that in spite of the smart meters can detect autonomously the detection of fraud (and therefore part of the fraud is already detected by the meter) our system performs better (around 15%) when the customer has an smart meter.

In parallel to the country level campaign to detect electricity fraud, we started to generate country level gas campaigns, using the same algorithm (i.e. Gradient Boosting Algorithm) The results of these campaigns, as seen in Figure 14, are also promising, achieving 4.8x and 11.3x the baseline model.

VI. Conclusions and Future Work

We presented a supervised approach to detect fraud in utility companies. Machine learning algorithms are used to detect patterns involving several indicators that even specialized employees have difficulty detecting. Our campaigns outperform both the gas and electricity baseline methodology that was being applied in both electricity and gas by an important factor. Moreover, it uses campaign feedback to learn continued and be aligned with not only historic but also novel fraud techniques that may arise. It can be deployed on new populations with relatively modest requirements (essentially, access to an initial set of detected fraud cases).

There is of course ample room for improving on the technical, machine learning aspect of the system (testing new classifiers, making the system more autonomous, using different representation of the current features as well as discovering new useful features automatically, …). As mentioned, we believe that in the future models should be more localized (i.e., normalize consumption with respect to geographical areas and similar climate, consider usage by users with the same tariff only, etc.), and that campaigns for large populations should be built by patching together a number of localized or specialized models.

Another possibility to explore is provided by smart meters. Although only monthly aggregates are being used right now, these meters register consumption at a much higher granularity, potentially providing extremely useful information for fraud detection.

One mid-term concern in our system has to do with an exploitation vs. exploration tradeoff. Since the system is driven by campaign feedback, there is a risk that it tends to focus on some niches that have given good reward in the past but become overexploited (all customers with those patterns have been checked), yet fails to explores other promising niches. We are considering ways of ensuring diversity in campaigns, perhaps introducing a randomization component.

Acknowledgements

This work has been partially supported by funds from the Spanish Ministry for Economy and Competitiveness (MINECO) and the European Union (FEDER funds) via grants COMMAS (ref. TIN2013-46181-C2-1-R) and APCOM (ref. TIN2014-57226-P), AGAUR (ref. MACDA SGR2014-890) and Gas Natural Fenosa Distribución.

References

[1] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer, 2007.

[2] A. A. Cárdenas, S. Amin, G. Schwartz, R. Dong, and S. Sastry. A game theory model for electricity theft detection and privacy-aware control in ami systems. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pages 1830–1837, Oct 2012.

[3] J. I. Guerrero, C. León, I. Monedero, F. Biscarri, and J. Biscarri. Improving knowledge-based systems with statistical techniques, text mining, and neural networks for non-technical loss detection. *Knowledge-Based Systems*, 71:376 – 388, 2014.

[4] R. Jiang, H. Tagaris, A. Lachsz, and M. Jeffrey. Wavelet based feature extraction and multiple classifiers for electricity fraud detection. In *Transmission and Distribution Conference and Exhibition 2002: Asia Pacific. IEEE/PES*, volume 3, pages 2251–2256 vol.3, Oct 2002.

[5] I. Monedero, F. Biscarri, C. León, J. I. Guerrero, J. Biscarri, and R. Millán. 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, 2012.

[6] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and M. Mohamad. Nontechnical loss detection for metered customers in power utility using support vector machines. *IEEE Transactions on Power Delivery*, 25(2):1162–1171, April 2010.

[7] A. H. Nizar, Z. Y. Dong, and Y. Wang. Power utility nontechnical loss analysis with extreme learning machine method. *IEEE Transactions on Power Systems*, 23(3):946–955, Aug 2008.

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

[9] C. C. O. Ramos, A. N. de Souza, A. X. Falcao, and J. P. Papa. New insights on nontechnical losses characterization through evolutionary-based feature selection. *IEEE Transactions on Power Delivery*, 27(1):140–146, Jan 2012.

[10] M. Sforna. Data mining in a power company customer database. *Electric Power Systems Research*, 55(3):201 – 209, 2000.

[11] J. V. Spirić, M. B. Doi, and S. S. Stankovi. Fraud detection in registered electricity time series. *International Journal of Electrical Power & Energy Systems*, 71:42 – 50, 2015.

[12] J. V. Spirić, S. S. Stanković, M. B. Dočić, and T. D. Popović. Using the rough set theory to detect fraud committed by electricity customers. *International Journal of Electrical Power & Energy Systems*, 62:727 – 734, 2014.

[13] R. D. Trevizan, A. S. Bretas, and A. Rossoni. Nontechnical losses detection: A discrete cosine transform and optimum-path forest based approach. In *North American Power Symposium (NAPS), 2015*, pages 1–6, Oct 2015.