SmartFD: A Real Big Data Application for Electrical Fraud Detection

D. Gutiérrez-Avilés[1], J. A. Fábregas[2], J. Tejedor[3],
F. Martínez-Álvarez[1], A. Troncoso[1], A. Arcos[4], and J. C. Riquelme[2]

[1] Division of Computer Science, Pablo de Olavide University, Seville, Spain
dgutavi@upo.es,fmaralv@upo.es,atrolor@upo.es
[2] Department of Computer Science, University of Seville, Spain
jfabregas@us.es,riquelme@us.es
[3] Endesa SA, Madrid, Spain
javier.tejedor@enel.com
[4] Department of Industrial Organization and Business Management, University of Seville, Spain
aarcos@us.es

Abstract. The main objective of this paper is the application of big data analytics to a real case in the field of smart electric networks. Smart meters are not only elements to measure consumption, but they also constitute a network of millions of sensors in the electricity network. These sensors provide a huge amount of data that, once analyzed, can lead to significant advances for the society. In this way, tools are being developed in order to reach certain goals, such as obtaining a better consumption estimation (which would imply a better production planning), finding better rates based on the time discrimination or the contracted power, or minimizing the non-technical losses in the network, whose actual costs are eventually paid by end-consumers, among others. In this work, real data from Spanish consumers have been analyzed to detect fraud in consumption. First, 1 TB of raw data was preprocessed in a HDFS-Spark infrastructure. Second, data duplication and outliers were removed, and missing values handled with specific big data algorithms. Third, customers were characterized by means of clustering techniques in different scenarios. Finally, several key factors in fraud consumption were found. Very promising results were achieved, verging on 80% accuracy.

Keywords: Big data, sensors, classification, fraud detection

1 Introduction

During most of the 20^{th} century, the interrelationship between electricity users and distribution companies remained unchanged. Suppliers were not chosen and, therefore, there was no need to treat consumers as customers. However, deregulation, the green agenda and the continuous technological leap have changed this relationship. New constraints such as security of supply, competitiveness and sustainability are the three priority axes towards changing the energy model that is currently demanded, which is materialized in objectives such as reducing emissions, renewable energy generation and improving energy efficiency.

An essential tool in this new model is the so-called smart meters that should not be understood only as devices that measure consumption but act as true sensors of the electrical network. These sensors configure a highly flexible and adaptable network that intelligently integrates the actions of the users that are connected to it, in order to achieve an efficient, safe and sustainable supply. The volume of information available from these networks is so huge that it can only be handled with Big Data techniques.

This proposal aims to provide a pioneering solution in the field of electrical distribution oriented to the analysis of the data provided by smart meters using big data techniques. The main objective of the paper is to develop a methodology aimed at the intelligent detection of non-technical losses in the field of electrical distribution, but it is not the only possibility. The data infrastructure and algorithms resulting from this paper may serve for a better understanding of the consumption patterns of customers. This study has the endorsement of the Endesa company to be able to access the data of its network.

With the aim of building a complete big data system for effective fraud detection, the authors have been accomplishing several tasks: A big data infrastructure based on HDFS and Spark has been built. Then, a knowledge discovery in databases (KDD) strategy has been followed. Raw data, which consisted of many duplicates, missing values and, even outliers, needed intensive preprocessing. Later, minable views were created, identifying labels and fraud targets. Finally, classification algorithms have been applied to different scenarios, reporting accuracies higher than 80%. Currently, site visits are being carried out, confirming the effectiveness of the proposed approach.

The rest of the paper is structured as follows. Section 2 reviews the most relevant papers related to this work. The applied methodology is described in Section 3. Reported results can be found in Section 4. Finally, the conclusions drawn from this study are summarized in Section 5.

2 State of the art

This section reviews the most relevant works published in the field of fraud detection during the last decade. A novel method for fraud detection in high voltage electrical energy consumers using data mining was introduced in [3]. The use of Self-Organizing Maps was proposed in order to identify consumption profiles. The authors mainly compared usage patterns in historical data with current behaviors, detecting anomalies in cases of fraud.

Monedero et al. [14] studied users with anomalous drops in energy in 2012. For this purpose, they used Bayesian networks and decision trees, also finding other types of non-technical loss patterns. The proposed methodology was tested with real customers, also from the Endesa company (hereon the partner).

The authors in [7] addressed the fraud detection problem in electric power distribution networks (low-voltage consumers). Namely, artificial neural networks were applied to discover fraud in Brazilian costumers. The authors claimed an improvement of over 50% when compared to other existing approaches. One year

later, in 2014, the use of artificial neural networks was proposed again for smart grid energy fraud detection [9].

The work in [16] analyzed time series without the seasonal component of consumers' power consumption at low voltage for the purpose of detect fraud and illogical consumption by customers. The authors drawn two main conclusions: energy drop in the last series is dominant sign in suspicious customer's detection and in series of suspicious customers it is noticed alternating positive autocorrelation.

Decision tree learning for fraud detection in consumer energy consumption can be found in [5]. In fact, this work proposed this kind of learning to profile normal energy consumption behavior, thus allowing for the detection of potentially fraudulent activity.

In 2016, a supervised approach for fraud detection in energy consumption was introduced in [6]. The model found anomalies in meter readings thanks to the application of machine learning techniques to historical data. Furthermore, the model is updated with incremental learning strategies and was tested on real Spanish data.

Finally, an approach based on machine learning to detect abnormalities is customer behaviors was proposed in [12]. They assessed linear discriminant analysis and logistic regression performances. Reported results by logistic regression reached higher accuracy since it was able to forecast irregularities accurately.

3 Methodology

The main objective of *SmartFD* methodology is the application of big data analytics to smart electric networks and the construction of classification models in order to determine the customers with high probability of fraud in their electricity consumption.

The global process to achieve this objective is shown in Figure 1. We can see how the methodology has five phases framed in the *KDD* process [8] that are described in the following sections.

3.1 Data retrieval

The first phase consists in the data retrieval processing wherein the raw files and the data model documentation is provided by the Partner. Data from the smart electric network of the Partner is contained in the raw files. In addition to being elements to measure consumption, smart meters also constitute a network of millions of sensors in a power network. These sensors provide very large amounts of data that, once analyzed, can lead to significant advances in society.

These raw files had to be unzipped and stored our storage system. This task would be trivial in common *KDD* environment, in contrast, it is hard and critical in big data environments since issues related to data transport, network bandwidth and computational cost emerge.

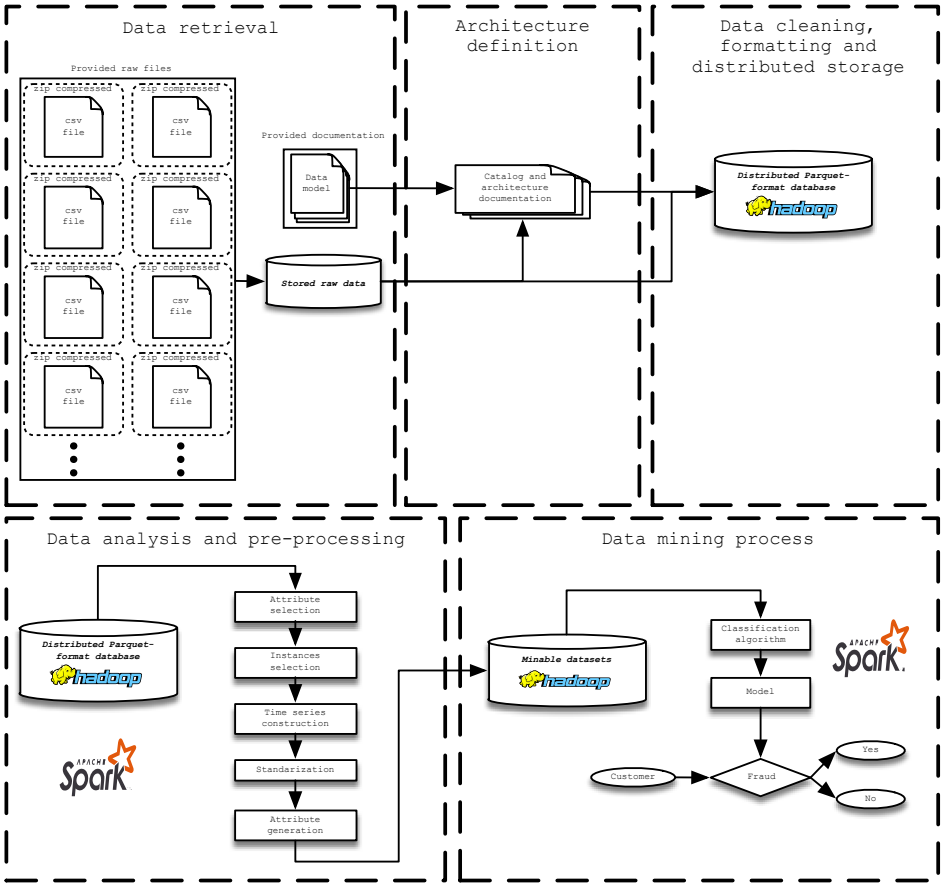

Fig. 1. *SmartDF* methodology

3.2 Architecture definition

The second phase consists of the study of data model study, the inference of the database architecture and the inventory of the schemes, tables, and relations that contains. The importance of this process lies in the necessity of produce a consistent, precise and self-contained documentation of the catalog and architecture of the database from a disorganized and inaccurate documentation and unstructured source data. Again, the effort and time invested, as well as the difficulty of this task, increases considerably in our big data environment.

3.3 Data cleaning, formatting and distributed storage

This achievement of this third phase is essential and critical in big data environments. These task enable us to process and handle large amount of data. With the support of the catalog and architecture documentation, the stored raw data

is processed in order to drop the inconsistencies, wrong formats and duplicate values for the purpose of obtain database with consistency and integrity.

Then, to be able to work in a big data context, the tables of the database have to be converted form CSV format to Parquet format as well as must be stored in a distributed file system implemented in a cluster of computer machines.

Apache Parquet [1] is an open-source column-oriented data store designed to support highly efficient compression and coding schemes, which allows for lower data storage costs and greater efficiency of one's query. In a column-oriented database, the information is stored in order of registration, so that a particular entry for different columns belongs to the same entry record. This means that one can access individual data elements (the name of a customer, a consumption date, a postal code, etc.) through columns as a group rather than reading row by row. In addition, this compression makes it easier for queries with column operations such as sum, average, minimum, etc. to be carried out much faster, so that when a query is made, it is only executed on the necessary columns. It should be added that, in addition to the considerable size reduction of the tables, they can be compressed in Snappy format [11], compatible with HDFS and Spark. If we also consider the option of using machine learning services in the cloud, such as those provided by Amazon Web Service (where services are billed by runtime and/or size of the scanned data), the savings in time and money would be enormous compared to using the CSV format.

For this purpose of distributed storage, a Hadoop environment has to be installed and configured, which will allow us to have an HDFS architecture [15]; thus, in addition to ensuring greater speed of access to data, we also implement greater tolerance to failures and crashes of cluster nodes due to the replicas that this file system generates.

3.4 Data analysis and pre-processing

The fourth stage aims to produce minable datasets to apply classification algorithms to them. From our target distributed Parquet-format database, we use Spark SQL [2] framework to carry out the following five steps:

1. Attribute selection: which attributes may have the greatest influence in identifying possible fraud have been studied. Along with the consumption of each customer, attributes such as postal code, business activity (in the case of non-residential customers), the model and status of smart meters, the power contracted by customers, or certain events that have occurred over the life of a contract have been some of the most interesting when constructing a first set of data.
2. Instances selection: in a second step, it has been necessary to check which of the instances in our set are optimal, so that in the classification stages we can obtain more interpretable sets of rules providing more information. Aspects such as the number of null values or number of correct readings have been some of the most important aspects to take into account in this stage of data pre-preparation. In addition to this, the data set have to be

balanced, a fundamental requirement for a dataset with which to generate a classification model.
3. Time series construction: once a set of quality data is generated. Our objective have been to build time series based on customer consumption. To do this, we have consumption readings with a quarterly frequency, so these time series encompass a wide range of frequencies, from low (quarterly, monthly) to very high (hourly). As a result, different datasets have been built with which to generate different classification models in later stages of the project.
4. Standardization: due to the difference in consumption amongst customers, the possibility of carrying out different standardizations of consumption data has been assessed. Calculating all consumptions with respect to a customer's average or maximum consumption are two possible options when generating new datasets.
5. Attribute generation: in addition to the different time series (standardized or not) and the attributes that we previously considered most relevant, we have also generated attributes that can provide additional information such as the number of estimated readings associated to a customer.

3.5 Data mining process

Finally, the fifth stage uses the minable datasets, that were produced in the previous stage, to extract hidden, useful, and valid information from them by means of machine learning algorithms. Specifically, we have used Spark MLlib framework [13] to apply classification algorithms which produce models that detect fraudulent behaviors of the Partner's customers.

4 Results

The experimental design and the preliminary classification results related to the global process presented in this paper (section 3) are described hereunder. This section is structured as follows: a brief description of the utilized data and the infrastructure is outlined in sections 4.1 and 4.2, next, the experimental setup is explained in 4.3 and, finally, the preliminary classification results are presented in 4.4.

4.1 Data

The analyzed database contains all data related to the Partner's customers of the Spanish region of Catalonia. These data have been retrieved in form of 251 csv files, likewise, this files compound 35 tables of 832 attributes in total. The size of the database is 1.48707619 TB (1487.076192 GB), it implies a real big data problem.

The most important characteristic of this database is that its content and relationships are divided in two: a scheme related to customer contracts and another with smart meters that collect consumption data. In each of the schemes,

called *stars*, there is a central table that relates to the tables belonging to the same scheme. This table is also connected to the other central table. In this manner, both stars are linked by their centres and are joined to their corresponding tables (tips of the star). These stars are:

- Contract star: it is formed by the tables that contain all the information related to contracting, geolocation, invoicing, customers, files, consumption types, campaigns, and technicians' work in the field.
- Device star: it is made up of tables with information regarding devices and consumption load curves, as well as different tables containing different events, validation records, etc.

4.2 Infrastructure

Due to the storage capacity and computational power required, as well as to be able to work optimally distributed, we have used the following hardware and software facilities:

- A cluster composed by 72 processing cores, 64 of them Intel (R) Xeon (R) Xeon (R) E7- 4820 CPUs @ 2.00GHz plus 8 Intel (R) Core (TM) i7-7700K CPUs @ 4.20GHz.
- 3 GeForce GTX 1080 GPUs with 2560 cores, Nvidia CUDA and 8 GB GDDR5X memory each.
- 128 GB RAM: 64 GB DD3 and 64 GB DDR4.
- A total storage capacity of 8 TB.
- Nodes interconnected through a Gigabit Ethernet network with a bandwidth of 1 Gbit/sec.
- AWS cloud computing services.
- Hadoop HDFS 2.8.0.
- Apache Spark framework 2.2.0

4.3 Experimental Setup

For this preliminary experimental study, the larger customers are been taking into account, thus, in order to accomplish the task of classifying this kind of customers in fraudulent or normal behavior, we have obtained a consistent and significant set of customer's daily consumption curves with both fraud and without it. The two options were considered to build the datasets that will train and test the classification algorithm are the following.

Setup #A: Dataset with subsequent measurement All the daily measurement that have occurred between a year before and three months after the start date of a sanction proceeding in the case of fraud customers have been selected.

For both the dataset to be balanced and to find a possible relationship between the consumption of the two types of customers, customers without fraud

that have the same number of measurements and at the same time, have the same type of business activity and the same postcode, have been selected.

With the aim of produce a model based on load curves of a customer for the period, a process to change quarter-hourly measurements to hourly measurement was necessary to create. In addition, it was counted the total number of estimated measurements of the hourly measurement.

Therefore, the final dataset is composed of the following fields:

- **customer code**
- **business activity code**
- **postcode**
- **number of estimated measurements**
- **dx** (being x the day, between 1 and 454. The field value is the addition of the measurements of the day)
- **label** (label with value 1 for fraud customers or 0 for those who are not fraud)

Setup #B: Dataset without subsequent measurement On the contrary to the previous dataset, in this case, only measurements occurred until a year before the start date of a sanction proceeding have been selected. Again, for both the dataset to be balanced and to find a possible relationship between the consumption of the two types of customers, customers without fraud that have the same number of measurement and at the same time, have the same type of business activity and the same postcode, have been selected. To build the dataset, the same process as for the creation of the previous dataset has been followed. Therefore, in the final dataset we count with the following fields:

- **customer code**
- **business activity code**
- **postcode**
- **number of estimated measurements**
- **dx** (being x the day, between 1 and 364). The field value is the addition of the measurements of the day)
- **label** (label with value 1 for fraud customers or 0 for those who are not fraud)

4.4 Classification results

The classification algorithm *Xgboost* has been chosen to be applied once the two datasets were created.

The *Xgboost* [4] algorithm trains, in an iterative way, decision trees to minimize a loss function. The specific method to tag again the records was defined by a loss function. The *Xgboost* algorithm decreases such loss function with the training data in every iteration.

For the purpose of establish the training, validation and testing procedure for the classification algorithm, the input datasets have been split into two parts:

training-validation part (70%) and test part (30%). Next, the training-validation part has been split again into two parts: training part (70%) and validation part (30%). The training part has been used to the algorithm training, the validation part has been used to cross-validation procedure [10] and the test part has been used as an external test for the resulting model.

For each experimental setup, we have built two classification models: one with all the attributes of the dataset and another excluding the number of estimated measurements. The aim of this decision was to prove the importance of the number of estimated measurements in the model generation, considering that we had previously detected that this value is critical by means of clustering analysis of the data. For each generated model and for each proof method (Cross Validation or Test) we the accuracy (ACC), the true positive ratio (TPR) and the true negative ratio (TNR) have been shown.

Setup #A results We can find these in Table 1. The complete model is the best one in terms of ACC and TPR, on the contrary, the model that exclude the number of estimated measurements has better values of TPR. In general terms, the complete model offers better results, reinforcing the importance of the number of estimated measurements as a key factor to classify fraudulent and normal customers.

Table 1. Setup A Results

Model	**Proof method**	**ACC**	**TPR**	**TNR**
Complete	Cross Validation	91.01%	87.67%	7.33%
Complete	Test	92.36%	86.16%	4.54%
Excluding #em	Cross Validation	72.26%	32.32%	8.96%
Excluding #em	Test	73.57%	39.51%	9.7%

Setup #B results They are presented in Table 2. In it, we can observe how the complete model is the best one in terms of ACC and TPR and how the model that exclude the number of estimated measurements has better values of TPR once again. The complete model offers once again better results in general terms, reinforcing the importance of the number of estimated measurements as a key factor to classify fraudulent and normal customers.

Table 2. Setup B Results

Model	**Proof method**	**ACC**	**TPR**	**TNR**
Complete	Cross Validation	89.16%	78.11%	5.45%
Complete	Test	87.55%	76.50%	6.58%
Excluding #em	Cross Validation	72.48%	52.16%	17.99%
Excluding #em	Test	73.45%	52.28%	16.29%

5 Conclusions

A real case for fraud detection in electricity consumption has been addressed in this paper. In particular, data from Spanish users have been processed, making use of big data technologies. Up to 1.5 TB of raw data were initially retrieved from different sources. After intensive preprocessing, data were cleaning and transformed into useful information, thanks to experts guidance. Later, machine learning algorithms have been applied in order to discover fraud consumption patterns in users. Two different scenarios have been considered, both of them reaching accuracies above 80%. The entire process has been carried out in a HDFS-Spark architecture, making the most of current big data technologies. Future works are directed towards the generation of models for different types of users and geographic areas.

Acknowledgments.

The authors would like to thank the Spanish Ministry of Economy and Competitiveness for the support under projects TIN2014-55894-C2-R and TIN2017-88209-C2-R.

References

1. Apache, S.F.: Parquet: A columnar storage format. `https://parquet.apache.org`
2. Armbrust, M., Xin, R.S., Lian, C., Huai, Y., Liu, D., Bradley, J.K., Meng, X., Kaftan, T., Franklin, M.J., Ghodsi, A., Zaharia, M.: Spark sql: Relational data processing in spark. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. pp. 1383–1394. SIGMOD '15, ACM, New York, NY, USA (2015)
3. Cabral, J.E., Pinto, J.O.P., Martins, E.M., Pinto, A.M.A.C.: Fraud detection in high voltage electricity consumers using data mining. In: Proceedings of the IEEE Transmission and Distribution Conference and Exposition. pp. 1–5 (2008)
4. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794. KDD '16, ACM, New York, NY, USA (2016)
5. Cody, C., Ford, V., Siraj, A.: Decision tree learning for fraud detection in consumer energy consumption. In: Proceedings of the IEEE International Conference on Machine Learning and Applications. pp. 1175–1179 (2015)
6. Coma-Puig, B., Carmona, J., Gavald, R., Alcoverro, S., Martin, V.: Fraud detection in energy consumption: A supervised approach. In: Proceedings of the IEEE International Conference on Data Science and Advanced Analytics. pp. 120–129 (2016)
7. Costa, B.C., Alberto, B.L.A., Portela, A.M., Maduro, W., Eler, E.O.: Fraud detection in electric power distribution networks using an ANN-based knowledge discovery process. International Journal of Artificial Intelligence and Applications 4(6), 17–23 (2013)
8. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The kdd process for extracting useful knowledge from volumes of data. Commun. ACM 39(11), 27–34 (Nov 1996)

9. Ford, V., Siraj, A., Eberle, W.: Smart grid energy fraud detection using artificial neural networks. In: Proceedings of the IEEE Symposium on Computational Intelligence Applications in Smart Grid. pp. 1–6 (2014)
10. Golub, G.H., Heath, M., Wahba, G.: Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21(2), 215–223 (1979)
11. Google: Snappy: A fast compressor/decompressor. `https://google.github.io/snappy/`
12. Lawi, A., Wungo, S.L., Manjang, S.: Identifying irregularity electricity usage of customer behaviors using logistic regression and linear discriminant analysis. In: Proceedings of the International Conference on Science in Information Technology. pp. 552–557 (2017)
13. Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D., Amde, M., Owen, S., et al.: Mllib: Machine learning in apache spark. The Journal of Machine Learning Research 17(1), 1235–1241 (2016)
14. Monedero, I., Biscarri, F., Len, C., Guerrero, J.I., Biscarri, J., Milln, R.: 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 and Energy Systems 34, 90–98 (2012)
15. Shvachko, K., Kuang, H., Radia, S., Chansler, R.: The hadoop distributed file system. In: 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST). pp. 1–10 (May 2010)
16. Spiric, J.V., Docic, M.B., Stankovic, S.S.: Fraud detection in registered electricity time series. International Journal of Electrical Power and Energy Systems 71, 42–50 (2016)