Employing Gradient Boosting and Anomaly Detection for Prediction of Frauds in Energy Consumption

Beatriz Albiero[1], Estevo Uyrá[1], Ramon Vilarino[1], Juliano Andrade Silva[2], Tales Fonte Boa Souza[2], Ricardo dos Santos[1], Sami Yamouni[1], Renato Vicente[1]

[1]Latam Datalab Serasa Experian
Alameda Vicente Pinzon, 51 - Vila Olímpia
04547-130 – So Paulo – SP – Brazil

{beatriz.albiero, estevao.uyra, ricardo.santos3,

sami.yamouni, renato.vicente}@br.experian.com

[2]CPFL Energia
Rua Jorge de Figueiredo Correa, 1632 - Jardim Professora Tarcília
13087-397 – Campinas – SP – Brazil

{julianoandrade, tfonteboa}@cpfl.com.br

Abstract. *Energy fraud is a critical economical burden for electric power organizations in Brazil. In this paper we present the application of cutting-edge Machine Learning algorithms, namely XGBoost and Isolation Forest, for prediction of irregularities in electrical energy consumption. By using a Logistic Regression model as a benchmark, we show that the use of XGBoost results in a significant improvement in the F1-score for fraud predictions in two different scenarios: with and without inspection history features. Moreover, we also propose the use of the Isolation Forest algorithm for detection of anomalies in electrical energy consumption. We show that this approach may be useful in the case of lack of inspection history features, surpassing dummy classifiers.*

1. Introduction

The last few decades have witnessed a drastic increase in global energy consumption, driven by the accelerated growth of industry and technology. In special, world electricity consumption almost duplicated in the last quarter-century with new demands such as accessibility to home appliances and transportation [ENERDATA 2019]. In the future, despite the development of more efficient devices and production processes, it is expected a steady 1% annual growth in electricity demand [Antunes Lima 2019].

Competition for market share in electrical utility industry has raised the energy loss during distribution as a major concern for generation companies in order to increase efficiency ([Management Solutions 2017]). The reasons for energy loss in distribution lines can be separated in two types: technical and non-technical [Antunes Lima 2019, Doukas et al. 2011, Management Solutions 2017]. The first case is inherent to the physical properties of electricity transport through grid, when a fraction of energy is converted and dissipated through heat or lost due to inductive and capacitive effects [Antunes Lima 2019, Doukas et al. 2011, Management Solutions 2017]. The second type of energy loss comprises in fraudulent practices by consumers that deliberately modify

energy measuring devices in order to reduce household bills or perform new illegal electricity connections on main power cables in the neighborhood [Smith 2004, Ford et al. 2014, Cody et al. 2015a, Coma-Puig et al. 2016]. Despite being a crime in many countries, energy fraud is a widespread practice encouraged by the difficulty in verification, which relies mainly on *in situ* inspections.

In Brazil, the high incidence of energy fraud is a critical economical burden for electric power organizations. According to ANEEL (the Brazilian Electricity Regulatory Agency), it was estimated a loss of 14% of total electrical energy available for distribution on 2016, with fraud practice accounting for approximately half of this deficit and a total burden of $1.2 billion at the same year [Antunes Lima 2019, Maia 2017]. This amounts to 3.6 times the last year's budget for the National Council for Scientific and Technological Development (CNPq[1]).

Recently, advancements in machine learning algorithms and computing power provided novel solutions to boost efficiency in detection of energy frauds. Based on patterns and anomalies identification in consumption, predictive models can highlight potential candidates for *in situ* inspection and reduce the cost of energy fraud detection [Ford et al. 2014, Cody et al. 2015a]. Herein, we describe the application of state-of-the art machine learning techniques for fraud detection in electrical energy consumption. These studies resulted in robust predictive models for fraud occurrence based on gradient boosting applied to registry profiles and energy consumption records. Moreover, we also propose a generalized and unsupervised model for fraud detection based on consumption anomalies inferred by isolation forests.

2. Related Work

Several applications of supervised and unsupervised machine learning algorithms for prediction of fraud and irregularity in electric utility can be found in literature. [Messinis and Hatziargyriou 2018]. Examples of implementations of usual supervised methods include the application of support vector machines to identify customer's abnormal consumption behavior based on previous energy usage data [Nagi et al. 2010, Alfarra et al. 2018]. Other case studies of well-established machine learning methods focused on fraud detection in electricity consumption are the use of decision trees [Monedero et al. 2012, Cody et al. 2015b], logistic regression, linear discriminant analysis ([Lawi et al. 2017]) and time series [Nogales et al. 2002]. Additionally, recent studies have provided new insights with the use of more complex machine learning models, such as Neural Networks [Nizar et al. 2008, Monedero et al. 2006, Costa et al. 2013] and rough set theory [Spiri et al. 2014]. Within the scope of Unsupervised Learning, Cabral et al. in [E. Cabral et al. 2008] present self-organizing maps that learns historical consumer energy consumption behavior. This study is focused in high voltage electricity consumers. Furthermore, [Angelos et al. 2011] proposes a two step methodology in order to find consumers with similar consumption profiles and hence potential fraudsters. This methodology consists of: (i) a C-means-based fuzzy clustering and (ii) a fuzzy classification system to rank users according to their irregular patterns.

[1]http://www.portaltransparencia.gov.br/orgaos/20501?ano=2018

3. Methodology

This section describes the methodology used in the present research. First, we discuss the two datasets provided by CPFL Energia [2]. The datasets basically consist of the same features and also share the same target distribution, but diverge with relation to historical information. The first dataset containing outdated information was used to fit the model. Later, a second dataset containing more recent information was used in an out-of-time validation. Secondly, we introduce the model used for this fraud classification task, the XGBoost model. We also provide a list of hyperparameters in which we performed a grid search.

3.1. Datasets

In order to develop predictive models for fraud in energy consumption, the following primary datasets were considered: (i) reports of local inspections and (ii) history of energy consumption for each registry. The datasets were provided by CPFL Energia, a utility company distributing electricity. They contain data of a medium-size brazilian city with around 700 thousands customers between February of 2014 and September of 2018. Considering [Messinis and Hatziargyriou 2018] definitions, the dataset contains low-resolution energy data, with a time resolution of one day, at consumer level. After feature engineering and categories aggregation a total of 64 features was employed in statistical modeling studies. For confidentiality reasons, the variables information are condensed in classes as described in Table 1. The resulting features are very similar to the ones described in [Messinis and Hatziargyriou 2018]. The fraud event variable describes exclusively the fraudulent or non-fraudulent events observed by the company investigators. Any irregular behaviours with no proved malicious intent has been discarded.

Figure 1 shows the fraud distribution in the two datasets used in the present article. The first one will be the basis to fit the classification model contains nearly 35 thousand records, while the second one will be used to validate the model out-of-time and contains nearly 7 thousand records. It is straightforward to see the conservation of fraud proportion between both datasets. They both present an unbalanced dataset with around 72% of regular events and 28% of irregular or fraud events.

3.2. XGboost classifier for fraud consumption

We approach the problem of fraud detection as a supervised binary classification problem and set the fraud_event feature (described in Section 3.1) as our target. The remaining features are used to fit a XGBoost model.

The XGBoost algorithm [Chen and Guestrin 2016] is a decision-tree-based ensemble model that has been first introduced in 2014. It uses gradient boosting, an iterative and additive approach where new models are trained to predict the residuals of prior models. Since we defined fraud detection as a binary classification problem, we use logistic regression (LR) for binary classification as our learning objective.

We perform a grid search over 5 hyperparameters: (i) Number of Estimators, (ii) Sub-sampling of Columns, (iii) Maximum Depth and regularization (iv) Gamma and (v)

[2]CPFL Energia, Rua Jorge de Figueiredo Correa, n 1632, Jardim Professora Tarclia CEP 13.087-397, Campinas/SP, Brazil.

Table 1. Fraud dataset description.

Feature	*Description*
Fraud_event	Target variable labelling fraud (1) and non-fraud events (0), as reported by local inspection
meter_ID	Meter identifier related to the consumer
t_0	Inspection date used as reference for feature values
Inspection history	Variables describing the history of inspections results for a specific meter equipment.
Meter characteristics	Set of variables describing different meter characteristics (e.g.: age of equipment, model, manufacturer brand).
Geographic location	geographic variable calculated from the meter location.
Consumption at inspection date	Total electrical energy consumption in kwh, as measured at t_0
Consumption history	Historical energy consumptions before t_0
Consumption statistics	Set of statistic variables calculated from consumption history (e.g.: coefficient of variation).

Min Child Weight (Minimum sum of instance weight (hessian) needed in a child leaf). We set our learning rate to 0.1. Remaining hyperparameters are set to default values. Table 2 shows the hyperparameters and the respective values on which we have performed the grid search.

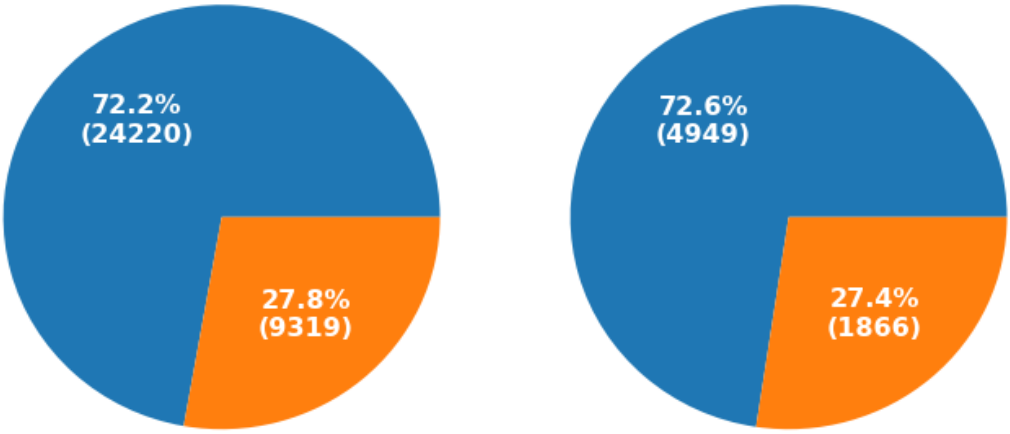
Table 2. Hyperparameter grid search for XGBoost

Hyperparameter	**Values**
n estimators	10-560
col subsampling	0.3, 0.5
max_depth	1-12
gamma	1, 4, 10, 20
min_child_weigth	1, 4, 10

The grid search is performed using 5-Fold [Raschka 2018] cross-validation. For evaluating the model performances, we use the F1-score. It is a metric defined as the harmonic mean between Precision and Recall, and is considered a parsimonious metric when dealing with unbalanced problems, which is our case.

3.3. Unsupervised analysis

We apply an anomaly detection analysis, making use of the less amount of data per observation. In special, we made no use of "fraud" labels when training the model, making it unsupervised. The anomaly detection model produces an "anomaly score" for each example, analogous to how a binary classifier would produce a score for the "True" label. We treated the unsupervised anomaly score as a fraud score, such that "common" examples (with low anomaly score) were considered as legitimate, and "odd" examples (with high



(a) Original **(b)** Out-of-time

Figure 1. Fraud distribution in original (a) and out-of-time (b) datasets. Regular and irregular situations are displayed in blue and orange, respectively. The figures in bracket represent the true numbers of investigated customers for each category.

anomaly score) were considered fraudulent. We fitted an Isolation Forest model ([Liu et al. 2008]), using all the data in which we had no label (no inspections were made), and evaluated the predictions on the inspected population. We made no parameter tuning, relying on the default parameters of the scikit-learn implementation [Liu et al. 2012]. After fitting, we ignored the default labels and have chosen the threshold that maximized the F1-score, based on the precision-recall curve.

4. Results and Discussion

In this section we compare the results obtained with the XGBoost model and a LR model. Both models were tested over two different datasets, one containing inspection history data, and a second without such features. We also show how we conducted feature selection to refine the XBoost model. Furthermore, there is a section dedicated to the topic of anomaly detection over energy consumption. In this, we show that anomaly scores may be used as indicative of fraud.

4.1. Models Comparison for Fraud Detection

Inspection history variables demonstrated to be among the most important features for fraud prediction. Hence, we created a new dataset by excluding inspection history data, namely Newcomers. We chose to compare XGBoost with a LR model since it would be a good benchmark for a binary classification task. Therefore, we test XGBoost and LR performances over both datasets.

Three metrics were used to assess of these algorithms: F1-score, precision and recall. F1-score is the more adapted when addressing unbalanced classification problems, as it is the case here. In Table 3, results show that XGBoost outperforms LR in both settings,

in particular for the newcomers customers. And, for all metrics, the "All customers" case displays higher performance than the "Newcomers"'s one. It is expected since customer historical data are taken into account.

Table 3. Fraud metrics for comparing performance of Logistic Regression (LR) and XGBoost models.

	All customers		Newcomers	
	LR	XGBoost	LR	XGBoost
F1	0.67	0.8	0.13	0.61
Precision	0.79	0.9	0.37	0.81
Recall	0.58	0.71	0.08	0.48

4.2. Out-of-time model validation

The out-of-time validation process checks the model robustness on a later dataset than the one on which the model has been fitted. It is useful when the application of a model to a population is changing over time such as the energy consumption. Results considering all customers as well as newcomers are displayed in table 4, both using XGBoost algorithm. Considering "All customers" column, we note a slight and consistent decrease in all the metrics. This trend is also observed for the newcomers except for the recall where the increase is not relevant regarding the standard deviation.

Table 4. Out-of-time validation.

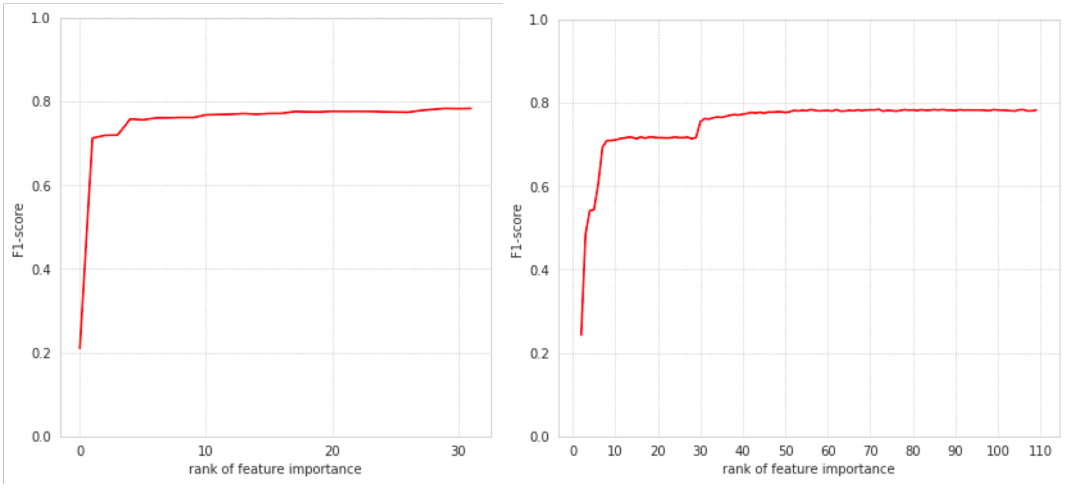
	All customers	Newcomers
F1	0.72	0.59
Precision	0.82	0.66
Recall	0.64	0.53

4.3. Model refinement

The XGBoost model automatically provides a list of the features ranked by their importance on the predictive model problem. Following this list as an importance rank, we gradually increased the number of features in order to observe the smallest set of features with higher importance that could provide a high predictive accuracy. We performed this study considering both the full dataset and the group of consumers with no previous inspection history (newcomers).

Figure 2 shows the change in F1-score for both cases with the gradual addition of features based in importance rank. As can be observed for the more general case considering all type of consumers (2(a)), F1-score drastically increases when the first 4 features are considered. These first 4 features consist with information of the coefficient of variation of energy consumption along last year, sum of previous fraud events for the location, and features that describes meter equipment age and geographical location.

When removing features related to historical inspection (Figure 2(b)), we observe a change in the pattern of F1-score increase, with a more gradual growth in performance and two main substantial increases: (i) after including the 8th and (ii) the 30th feature ranked by importance. In order to keep a predictive model with highest F1-score, we



(a) all consumers **(b)** newcomers

Figure 2. Refinement of predictive models in order to keep only the most relevant features for predictive efficiency for (a) all costumers and (b) costumers with no previous inspection history (newcomers).

considered for the case of newcomers the top 42 features to be considered in the final model. Table 5 summarizes the results obtained for external cross-validation and external hold-out datasets.

Table 5. Refinement of models.

	F1 external CV	F1 hold-out
All consumers	0.80	0.76
Newcomers	0.78	0.60

Moreover, a pattern change in the final prediction models that is worthy to mention is the role of the most important features for fraud distinction. While historical fraud events are crucial features for the most general model considering all consumers, in the absence of historical features (newcomers), the obtained models still present considerable external predictive power with the geolocation and consumption coefficient of variation being the most important contributors for fraud discrimination.

4.4. Anomaly detection may help when there is no labeled data

Extending our analysis to contexts when even less data is available, we removed the correct labels from our training set, making the problem unsupervised. Assuming that data irregularities could be indicative of fraud, we use Isolation Forest ([Liu et al. 2008]), an anomaly detection that does not have label supervision, as described in 3.3. To test this, we collected a subset of the data in which the proportion of frauds and was approximately 11%, comparing the results with the expected within our subset of data. Because we have no prior information about the target (it is unavailable in this setting), the precision does not raise above the true proportion, which is 11%. Hence, the best f1 score achievable in this setting is when all samples are categorized as frauds (at least reaching 100% recall).

The results, shown in Table 6, indicate that anomaly scores can be used as a proxy

for fraud. Specially when the amount of unlabeled data is big, the anomaly results are a considerable improvement over a random baseline.

Table 6. Frauds detected as anomalies.

	Random	Random with same recall	Anomalies
F1	0.20	0.18	0.30
Precision	0.11	0.11	0.21
Recall	1.00	0.50	0.50

5. Conclusions and Future work

In this paper, we presented both a supervised and an unsupervised approaches to detect fraud using data from CPFL Energia, a utility company distributing electricity.

Regarding the supervised study, models using XGBoost algorithms outperformed the benchmark logistic regression models displaying a F1-score of 0.8. We explained this difference by the fact that XGBoost models perform better on unbalanced datasets, as it is the case here. The resulting model has then been successfully validated on an out-of-time dataset and newcomers, which are populations without any investigation historic. A refinement study was also conducted, using the XGBoost feature importance list as a reference. For the model considering the general case of all type of consumers, the filtered predictive models presented very low decrease in F1-score metric even when considering only the four most important features. Among these features, there are data that accounting consumption changes that suggests to be able to identify the changes in the customers behavior, historical fraud observations and a geographic variable related to the meter location.

On the other hand, for the unsupervised study, we ran an anomaly detection algorithm using Isolation Forest and it has shown promising results.

The results presented in this article only concern one mid-size city, and should be apply to other geographic regions. By doing so, we would potentially create a generic version of the fraud detection model.

Acknowledgments

This work was supported by ANEELs research & development program (Project ID PD-0063-3039/2018) in partnership with CPFL ENERGIA group companies.

References

Alfarra, H., Attia, A., and S. M. El Safty, C. (2018). Nontechnical loss detection for metered customers in alexandria electricity distribution company using support vector machine. *Renewable Energy and Power Quality Journal*, 1:468–474.

Angelos, E., Saavedra, O., Carmona Cortes, O., and Souza, A. (2011). Detection and identification of abnormalities in customer consumptions in power distribution systems. *Power Delivery, IEEE Transactions on*, 26:2436–2442.

Antunes Lima, D. (2019). Perdas de energia - aneel (brazilian electricity regulatory agency). `https://www2.camara.leg.br/atividade-legislativa/`

comissoes/comissoes-permanentes/cme/audiencias-publicas/2018/audiencia-publica-16-05-2018/ANEEL\%20-\%20\%20Perdas\%20Eletricas\%20-\%20Davi\%20Lima.pdf.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *arXiv e-prints*, page arXiv:1603.02754.

Cody, C., Ford, V., and Siraj, A. (2015a). Decision tree learning for fraud detection in consumer energy consumption. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 1175–1179. IEEE.

Cody, C., Ford, V., and Siraj, A. (2015b). Decision tree learning for fraud detection in consumer energy consumption. *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 1175–1179.

Coma-Puig, B., Carmona, J., Gavalda, R., Alcoverro, S., and Martin, V. (2016). Fraud detection in energy consumption: A supervised approach. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 120–129. IEEE.

Costa, B., L. A Alberto, B., M. Portela, A., W, M., and O.Eler, E. (2013). Fraud detection in electric power distribution networks using an ann-based knowledge-discovery process. *International Journal of Artificial Intelligence & Applications*, 4:17–23.

Doukas, H., Karakosta, C., Flamos, A., and Psarras, J. (2011). Electric power transmission: An overview of associated burdens. *International Journal of Energy Research*, 35(11):979–988.

E. Cabral, J., Pinto, J., M. Martins, E., and M. A. C. Pinto, A. (2008). Fraud detection in high voltage electricity consumers using data mining. pages 1 – 5.

ENERDATA (2019). Global energy statistical yearbook 2019. https://yearbook.enerdata.net/electricity/electricity-domestic-consumption-data.html.

Ford, V., Siraj, A., and Eberle, W. (2014). Smart grid energy fraud detection using artificial neural networks. In *2014 IEEE Symposium on Computational Intelligence Applications in Smart Grid (CIASG)*, pages 1–6. IEEE.

Lawi, A., Wungo, S. L., and Manjang, S. (2017). Identifying irregularity electricity usage of customer behaviors using logistic regression and linear discriminant analysis. *2017 3rd International Conference on Science in Information Technology (ICSITech)*, pages 552–557.

Liu, F., Ting, K., and Zhou, Z.-H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data*, 6(1):1 – 39.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE.

Maia, C. (2017). Perdas de energia custam mais de r$8 bi aos consumidores em 2016. https://www.valor.com.br/empresas/5219107/perdas-de-energia-custam-mais-de-r-8-bi-\aos-consumidores-em-2016.

Management Solutions, M. (2017). Fraud management in the energy industry. https://www.managementsolutions.

com/sites/default/files/publicaciones/eng/fraud-management-in-the-energy-industry.pdf. Accessed: 2019-07-11.

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

Monedero, I., Biscarri, F., Leon, C., Guerrero, J. I., Biscarri, J., and Millan, 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*, 34:90–98.

Monedero, I., Biscarri, F., Len, C., Biscarri, J., and Milln, R. (2006). Midas: Detection of non-technical losses in electrical consumption using neural networks and statistical techniques. pages 725–734.

Nagi, J., Yap, K. S., Tiong, S. K., Ahmed, S. K., and Mohamad, M. (2010). Nontechnical loss detection for metered customers in power utility using support vector machines. *IEEE Transactions on Power Delivery*, 25:11621171.

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

Nogales, F., Contreras, J., J. Conejo, A., and Espinola, R. (2002). Forecasting next-day electricity prices by time series models. *Power Engineering Review, IEEE*, 22:58–58.

Raschka, S. (2018). Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. *arXiv e-prints*, page arXiv:1811.12808.

Smith, T. B. (2004). Electricity theft: a comparative analysis. *Energy policy*, 32(18):2067–2076.

Spiri, J. V., Stankovi, S. S., Doi, M. B., and Popovi, T. D. (2014). Using the rough set theory to detect fraud committed by electricity customers. *International Journal of Electrical Power & Energy Systems*, 62:727 – 734.