# Distilling Provider-Independent Data for General Detection of Non-Technical Losses

Jorge Augusto Meira, Patrick Glauner,
Radu State and Petko Valtchev
Interdisciplinary Centre for Security, Reliability and Trust
University of Luxembourg, Luxembourg
{first.last}@uni.lu

Lautaro Dolberg, Franck Bettinger and
Diogo Duarte
CHOICE Technologies Holding Sàrl, Luxembourg
{first.last}@choiceholding.com

***Abstract*—Non-technical losses (NTL) in electricity distribution are caused by different reasons, such as poor equipment maintenance, broken meters or electricity theft. NTL occurs especially but not exclusively in emerging countries. Developed countries, even though usually in smaller amounts, have to deal with NTL issues as well. In these countries the estimated annual losses are up to six billion USD. These facts have directed the focus of our work to the NTL detection. Our approach is composed of two steps: 1) We compute several features and combine them in sets characterized by four criteria: temporal, locality, similarity and infrastructure. 2) We then use the sets of features to train three machine learning classifiers: random forest, logistic regression and support vector vachine. Our hypothesis is that features derived only from provider-independent data are adequate for an accurate detection of non-technical losses. We used Area Under the Receiver-operating Curve (AUC) to assess the results.**

***Keywords*—*Artificial intelligence, big data, electricity theft, feature engineering, machine learning, non-technical losses.***

## I. Introduction

Electricity is a key factor to reduce the poverty and improve the quality of life all around the world. Around 84% of the global population has access to electricity and this number tends to grow according to *The World Bank*[1]. There are several sources of electricity, such as nuclear plants, hydroelectric, natural gas, wind turbines, etc. Once the electricity is generated, it is distributed using power grids. During the distribution phase losses happen quite commonly and are classified in two groups: technical or non-technical losses (NTL). Technical losses are naturally caused by dissipation, while non-technical losses include poor equipment maintenance, broken meters, un-metered supply and electricity theft [1]. In this paper we consider NTL as a black box, which means we make no distinction between the different types of NTL.

Non-technical losses occur especially but not exclusively in emerging countries. For example, the NTL estimation in India is around US$ 4.5 billion. In other emerging countries such as Brazil, Malaysia and Lebanon, NTL take up to 40% of the total electricity distributed. Developed countries, even though usually in smaller amounts, have to deal with NTL issues as well. UK and USA estimate these losses in a range between one and six billion USD. These facts have directed the focus of our work to the NTL detection.

We claim that a robust and portable approach may not rely on database particularities from different electricity providers. In this sense, we argue that a reliable set of features should be supported by common data to any electricity providers' database.

Overall, this work makes the following contributions:

- We compute several sets of features using data from a real Big Data base from Choice Technologies Holding Sarl company[2] (190M meter readings, 3.5M customers and 3M inspection results)
- We show the impact of the computed sets on supervised machine learning.
- We successful demonstrate that features supported only by raw consumption data presents satisfactory results when compared with "provider dependent" features.

This paper is organized as follows: Section II presents some background on machine learning and discusses related work. Section III presents our feature engineering step in detail. Section IV shows the outcomes and we conclude in Section V.

## II. Background and Related Work

In this section we first provide some background on machine learning and feature engineering. Next, we discuss the related work on NTL detection.

### A. Background

Machine learning is the ability of a software to learn autonomously [2]. We highlight two particular classes of machine learning algorithms used in this work: supervised and unsupervised learning. On the one hand, supervised learning algorithms are trained with data containing known labels (e.g., NTL or non-NTL) to produce a mathematical model as output. Later, this model is used to make predictions from unlabeled data. On the other hand, unsupervised learning algorithms use only unlabeled data in order to draw inferences from data (e.g, k-means) [3]. In both cases, it is essential to choose relevant data to create a meaningful set of features to support a robust learning model. This task is called feature engineering.

[1] http://www.worldbank.org/en/topic/energy/

[2] http://choiceholding.com/

978-1-5090-5550-0/17/$31.00 

TABLE I. SET OF FEATURES

| Set of features | Description |
|---|---|
| Notes (N) | Meter reader's notes |
| Consumption (C) | Fixed Interval + Fixed Lag |
| Consumption & Notes (CN) | Fixed Interval and Notes |
| Neighbourhood (Ng) | Intra Group (geographical neighbourhood) |
| Transformers (T) | Intra Group (Transformers) |
| Consumption Profile (CP) | Intra Group (k-means clustering) |
| C & Ng | Consumption and Neighbourhood |
| C & CP | Consumption and Consumption Profile |
| All | N+C+Ng+CP+T |

### B. Related Work

NTL detection can be treated as an anomaly or fraud detection problem. Comprehensive surveys of the field of NTL detection are provided in [4], [5] and [6]. Surveys of how an advanced metering infrastructure can be manipulated, are provided in [7] and [8].

One method to detect NTL is to derive features from the customer consumption time series, such as in [9]: average consumption, maximum consumption, standard deviation, number of inspections and average consumption of the residential neighborhood. These features are then grouped into c classes using fuzzy c-means clustering. Next, customers are classified into NTL or non-NTL using the fuzzy memberships. An average precision of 0.745 is achieved on the test set.

Daily average consumption features of the last 25 months are used in [10] for less than 400 out of a highly imbalanced data set of 260K customers. These features are then used in a support vector machine (SVM) with a Gaussian kernel for NTL prediction, for which a test recall of 0.53 is achieved.

The class imbalance problem has been addressed in [11]. In that paper, an ensemble of two SVMs, an optimum-path forest and a decision tree is applied to 300 test data. While the class imbalance problem is addressed, the degree of imbalance of the 1.5K training examples is not reported.

The consumption profiles of 5K Brazilian industrial customer profiles are analyzed in [12]. Each customer profile contains 10 features including the demand billed, maximum demand, installed power, etc. A SVM and k-nearest neighbors perform similarly well with test accuracies of 0.962. Both outperform a neural network, which achieves a test accuracy of 0.945.

In [13] authors particularly addressed the class imbalance of NTL detection and how to assess models in such environment by comparing Boolean and fuzzy expert systems. Addressing the same imbalanced class, the authors proposed neighbourhood features in [14] and demonstrated why these features are statistically meaningful.

## III. FEATURE ENGINEERING

Feature engineering is a key task to support learning algorithms, as mentioned in Section II-A. We compute several features using the following criteria:

1) Temporal: Seasonal, Monthly, Semiannual, Quarterly, Intra Year;
2) Locality: Geographical Neighbourhoods;
3) Similarity: k-means clustering using consumption profile[3];
4) Infrastructure: Transformers.

The unit to build the features is the monthly consumption of a given customer, described as follows:

$$C_i = [C_i^1, \ldots, C_i^n], \tag{1}$$

*where $C_i$ is a certain customer and $C_i^n$ is its consumption along $n$ months.*

Thus, each feature is calculated using the customer's consumption in a period of time according to a given criteria (*Temporal, Locality, Similarity, Infrastructure*).

*1) Fixed Interval:* The fixed interval calculates the difference between the current consumption and the average consumption in a period of time:

$$\text{K\_diff}_i^j = C_i^j - \left( \frac{1}{K} \times \sum_{k=j-K}^{j} C_i^k \right), \tag{2}$$

*where K assumes the set of values [3, 6, 12].*

*2) Fixed Lag:* The fixed lag calculates the Intra Year difference:

$$\text{I\_diff}_i^j = C_i^j - \left( C_i^{j-K} \right), \tag{3}$$

*where K = 12.*

*3) Window:* The window calculates the seasonal difference (Intra Year):

$$\text{W\_diff}_i^j = C_i^j - \left( C_i^{j-(K+1)} + C_i^{j-K} + C_i^{j-(K-1)} \right) \times \frac{1}{3}, \tag{4}$$

*where K = 12.*

*4) Intra Group:* The Intra Group is calculated over a grouping criteria (Locality, Similarity and Infrastructure):

$$N_i = [N_i^1, \ldots, N_i^n] \text{ with } N_i^j = \frac{\sum_{t \in N} C_t^j}{\#N}, \tag{5}$$

*where N assumes different group's **id** according to Locality, Similarity and Infrastructure criteria.*

### A. Sets of Features

We created nine sets of features (see Table I). These sets are composed by combinations of features computed using the criteria presented in the previous section.

### B. Features Correlation

Features correlation draws the similarities between features and supports the feature engineering task. In [15] the authors present the following hypothesis: "*A good feature subset is one that contains features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other*". We used the Pearson product-moment correlation coefficient, which gives us the linear dependence between two variables X and Y, pairs of features in our case. The Paerson's correlation is given by:

[3]The goal is to replace *Geographical Neighbourhoods* by creating groups of customers with similar consumption.

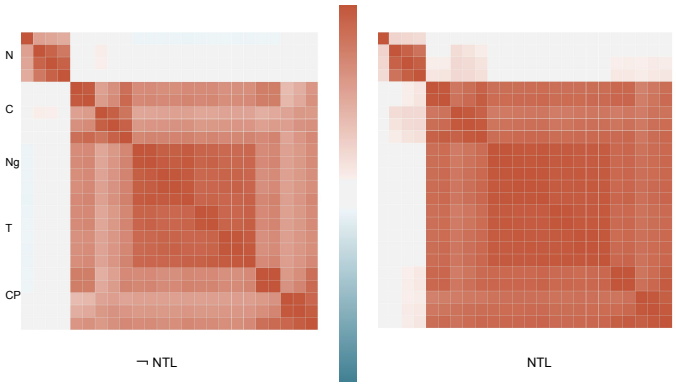



Fig. 1. Pearson's correlation of all features according to the label: in the left side non NTL and in the right side NTL, where 1 (dark orange) indicates a perfect positive linear correlation, 0 (white) indicates no linear correlation, and 1 (dark blue) indicates total negative linear correlation).

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{6}$$

*where* cov *is the covariance and* $\sigma$ *is the standard deviation of X and Y.*

The correlation between *NTL* customers and *non-NTL* costumers can be visualized in Figure 1. It shows an interesting finding related to *NTL* behaviour. The *NTL* customers presents higher features correlation when compared with *non-NTL* ones, which leads us to believe that *NTL* customers behave in a more similar way than *non-NTL* customers. Furthermore, we found that the set of features *"CP - Consumption Profile"* corroborates with the hypothesis mentioned in this section.

## IV. Evaluation

In this section, we evaluate the impact of different sets of features for NTL detection (see Table I). The experiments were run in a Intel(R) Xeon(R) CPU E5-2620 2.00 GHz machine with 128 GB RAM running Ubuntu 14.04.4. The code is written in Python using Apache Spark's (1.6.1) scalable machine learning library (MLlib)[4].

### A. Data

In historical data, up to 40% of the inspections end up in NTL. However, the model must be able to predict in different proportions of NTL and non-NTL. Thus, we generated data sets with NTL proportions from 10% to 90%, as follows: 10%, 30%, 40%, 50%, 60%, 70% and 90%.

### B. Metric

The performance measure used in the following experiments is the area under the receiver-operating curve (AUC) [16]. It plots the true positive rate or recall against the false positive rate:

$$\mathrm{AUC} = \frac{\mathrm{Recall} + \mathrm{Specificity}}{2}, \tag{7}$$

*where the recall measures the proportion of the true positives and the specificity measures the proportion of the true negatives.*

For a binary classification problem (e.g., NTL | non-NTL), a AUC score of 0.5 is equivalent to chance and a score of greater than 0.5 is better than chance.

### C. Setup

The data set is split into training, validation and test sets with a ratio of 80%, 10% and 10%, respectively. For each of the three models the trained classifier that performed the best on the validation set in any of the 10 folds is selected and tested on the test set to report the test AUC. This methodology is related to [13]. Overall, all three classifiers perform in the same regime, as their mean AUC scores over all NTL proportions are very close. This observation is often made in machine learning, as the actual algorithm is less important, but having more and representative data is generally considered to be more important [17]. This can also be justified by the no free lunch theorem, which states that no learning algorithm is generally better than others [18]. For this reason, in the next section we only present the results achieved by Random Forest classifier, which is slightly better than Logistic Regression and Support Vector Machine.

***Random Forest:*** A random forest is an ensemble estimator that comprises a number of decision trees [19]. Each tree is trained on a subsample of the data and feature set in order to control overfitting. In the prediction phase, a majority vote is made of the predictions of the individual trees.

### D. Results

Figure 2 shows the AUC performance of Random Forest classifier. First, Figure 2(a) draws a comparative overview between all sets of features presented on Table I using seven NTL proportions. The best results are zoomed in on Figure 2(b).

The set of features that cover all features overperforms any other set, but when compared with neighbourhood sets (i.e., *C & Ng, C & CP*) the difference is not relevant, around 1.5% better for NTL proportion of 70%. A more detailed comparison is presented on Table II.

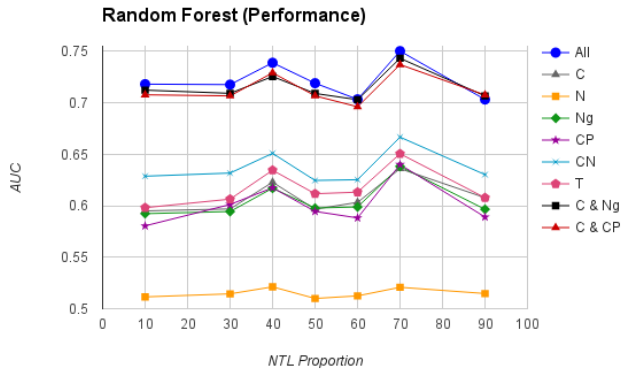
TABLE II. Random Forest performance on the best three set of features

| *NTL proportion* | **all** | **C & Ng** | **C & CP** |
|---|---|---|---|
| 10% | 0.7182 | 0.7123 | 0.7078 |
| 30% | 0.7177 | 0.7092 | 0.7067 |
| **40%** | **0.7389** | **0.7256** | **0.7292** |
| 50% | 0.7191 | 0.7091 | 0.7068 |
| 60% | 0.7034 | 0.7030 | 0.6962 |
| **70%** | **0.7503** | **0.7433** | **0.7372** |
| 90% | 0.7033 | 0.7067 | 0.7076 |

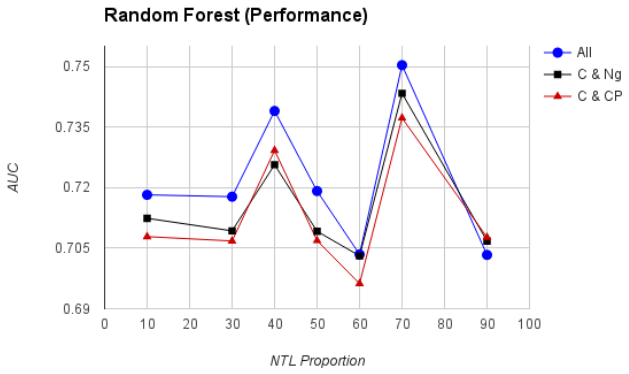
We pinpoint the performance with the NTL proportion of 40%, which is one of the most common NTL proportion found on real electricity distribution[5]. For this proportion, the

[4]http://spark.apache.org/mllib/

[5]Emerging countries such as Brazil, Malaysia and Lebanon, NTL take up to 40% of the total electricity distributed.




(a) All sets of features




(b) Best sets of features (zoom in)

Fig. 2. AUC performance of *Random Forest* on different NTL proportions trained with the sets of features presented in Table I.

performance difference for the best three sets of features is about 1%, and it is a case where the set of features "*C & CP - Consumption and Consumption Profile*" overperforms the set "*C & Ng - Consumption and Neighbourhood*".

### E. Discussion

Overall we highlight two sets of features: *"C & Ng"* and *"C & CP"*. These sets use the "Locality" and "Similarity" as criteria to compute Intra Group features. In the first case the intra group is based on Geographical Neighbourhoods and in the second case the intra group is based on consumption similarity. The performance of these sets are very similar to the performance of the complete set of features: *"All"*. Thus, we argue it is more likely to provide a "provider independent" classification model supported only by the set of features *"C & CP"*, since this set only uses features supported by raw consumption data.

## V. Conclusion

In this paper, we proposed a feature engineering approach to NTL detection. We evaluated three machine learning classifiers over several sets of features computed using four criteria: temporal, locality, similarity and infrastructure. The experimental results show that sets of features supported only by raw consumption data can achieve satisfactory performance when compared with sets composed of "providers' dependent features", such as notes or transformers. We also found out that for NTL detection the actual algorithm is less important than having representative set of features.

Based on our approach Company[6] carried out real inspections. The preliminary results show that common patterns for NTL, such as consumption downfall, are not a strict rule and costumers with consumption increasing may be also good targets.

In our future research, we intend to investigate covariate shift data issues in NTL (i.e., bias) and how to develop a robust classification method for this problem. We also plan to study in more detail the feature correlation in order to understand better customers' behaviour.

[6]Details omitted for double-blind reviewing.

## References


[1] T. B. Smith, "Electricity theft: a comparative analysis," in *Energy Policy*, 2004.

[2] A. Ng, "Machine learning," 2014.

[3] T. Pang-Ning, M. S., and Vipin, "Introduction to data mining, (first edition)," in *Addison-Wesley Longman Publishing Co.*, 2005.

[4] A. Chauhan and S. R. Rajvanshi, "Non-technical losses in power system: A review," in *Proceedings of International Conference on Power, Energy and Control (ICPEC)*, 2013, pp. 558–561.

[5] P. Glauner, A. Boechat, L. Dolberg, J. Meira, R. State, F. Bettinger, Y. Rangoni, and D. Duarte, "The challenge of non-technical loss detection using artificial intelligence: A survey," *arXiv preprint arXiv:1606.00626*, 2016.

[6] M. Kazerooni, H. Zhu, and T. J. Overbye, "Literature review on the applications of data mining in power systems," in *Power and Energy Conference at Illinois (PECI)*, 2014.

[7] R. Jiang, R. Lu, Y. Wang, and L. J., "Energy-theft detection issues for advanced metering infrastructure in smart grid," in *Tsinghua Science and Technology*, 2014.

[8] S. McLaughlin, D. Podkuiko, and P. McDaniel, "Energy theft in the advanced metering infrastructure," in *Lecture Notes in Computer Science*, 2009.

[9] E. W. S. Angelos, O. R. Saavedra, O. A. C. Cortes, and A. N. d. Souza, "Detection and identification of abnormalities in customer consumptions in power distribution systems," in *IEEE Transactions on Power Delivery*, 2011.

[10] 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," in *IEEE Transactions on Power Delivery*, 2010.

[11] M. D. Martino, J. Decia, F.and Molinelli, and A. Fernandez, "Improving electric fraud detection using class imbalance strategies," 2012.

[12] C. C. Oba Ramos, A. Nunes de Souza, D. Sinkiti Gastaldello, and J. Paulo Papa, "Identification and feature selection of non-technical losses for industrial consumers using the software weka," in *International Conference on Industry Applications*, 2012.

[13] P. Glauner, A. Boechat, L. Dolberg, R. State, F. Bettinger, Y. Rangoni, and D. Duarte, "Large-scale detection of non-technical losses in imbalanced data sets," in *Innovative Smart Grid Technologies Conference (ISGT), 2016 IEEE Power & Energy Society*. IEEE, 2016.

[14] P. Glauner, J. Meira, L. Dolberg, R. State, F. Bettinger, Y. Rangoni, and D. Duarte, "Neighborhood features help detecting non-technical losses in big data sets," in *3rd IEEE/ACM International Conference on Big Data Computing Applications and Technologies (BDCAT 2016)*, 2016.

[15] M. A. Hall, "Correlation-based feature selection for machine learning," 1999.

[16] W. B. van den Hout, "The area under an roc curve with limited information," in *Medical Decision Making*, 2003.

[17] M. Banko and E. Brill, "Scaling to very very large corpora for natural language disambiguation," in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ser. ACL '01. Stroudsburg, PA, USA: Association for Computational Linguistics, 2001, pp. 26–33. [Online]. Available: http://dx.doi.org/10.3115/1073012.1073017

[18] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," in *Neural Computation*, 1996.

[19] T. K. Ho, "Random decision forests," in *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, 1995.