Improving Hit Rates for and Operational Efficiency in Detecting Solar PV Panel Installations in the Philippines

Patrick H. Mosqueda
Business Analytics, Digital Technologies and Operations, Indra Philippines, Inc.
pmosqueda@indracompany.com

Abstract

The Philippines is one of the first countries in Southeast Asia to introduce the net-metering program which offers incentives to customers who produce their own supply and sell back the excess electricity to the grid. Distribution utilities (DUs), under the Renewable Act of 2008, are mandated to enter into a net-metering agreement with qualified end-users who will be installing renewable energy system. Despite this policy, majority of the end-users remain unregistered. A major DU in the country aims to identify these unregistered installations to adequately assess the power distributed through the network, monitor its effect to the grid stability, and assess possible revenue loss. The DU did a pilot study by collecting data via field inspection in target areas and manual examination of satellite imagery from maps available online. However, this proved to be tedious and inefficient, with only less than 1% hit rate. Acquisition of the latest advanced technologies, though effective, is very expensive. This paper aims to recognize the value of statistical machine learning (SML) in this situation. The DU's customer data, along with the data collected in the pilot study, was employed in generating predictive models using Microsoft Azure Machine Learning Studio, a cloud-based machine learning tool. With the SML techniques and advanced data analytics implemented, results showed a significant increase in hit rate and so a more efficient inspection process.

Keywords: solar; detection; net metering; statistical machine learning; analytics

1. Introduction

The Republic Act No. 9513, also known as the Renewable Energy Act of 2008, was signed into law to uphold the Philippine government's commitment to accelerate the utilization of renewable energy (RE) resources in the country. Under this law, *subject to technical considerations and without discrimination and upon request by distribution end-users, the DUs shall enter into net-metering agreements with qualified end-users who will be installing RE system*. Net metering, as defined in the law, *refers to a system, appropriate for distributed generation, in which a distribution grid user has a two-way connection to the grid and is only charged for his net* *electricity consumption and is credited for any overall contribution to the electricity grid*. DUs are authorized to provide the mechanisms for the physical connection and commercial arrangements necessary to ensure the success of the net-metering program.

Since its enactment, there has been a slow but steady rise in the installation of solar photovoltaic (PV) panels in both commercial and residential premises. However, despite the policies established to regulate the utilization of RE system, majority of these installations are not yet registered under the net-metering program. Aside from regulatory issues, unregistered installations may also pose safety threats to the physical stability of the electricity grid if not properly integrated and monitored.

To address the said risk while promoting the net-metering program, a study was conducted in 2017 by one major DU in the country to estimate the quantity of solar PV panel end-users and assess the effect of unregistered installations to the network. Data was initially collected through field inspections and manual rooftop examination based on satellite and aerial imagery from online public maps. For two months, three high-end subdivisions were monitored for residential services with visible solar PV panel installations (PVIs). Residential customers in high-end subdivisions were chosen as the target population as they have the means and higher likelihood to purchase and use solar PV panels. With only 25 unregistered PVIs confirmed out of 7,693 services examined, attaining only 0.32% hit rate, this system proved to be impractical to operationalize. Online maps are usually out-of-date and manual inspections are man-hour intensive.

Advanced technologies and data analytics that use satellite and aerial imagery for detection could be procured. Although highly effective, it is extremely expensive. Besides, there have been some experiments using detection algorithms on satellite imagery (Malof, Hou, Collins, Bradbury, & Newell, 2015) that can also be performed but data specific to the DU's franchise area was not yet available. It was also too tedious and unrealistic to manually create the dataset given the short timeline. The company then decided to tackle the problem with more cost-effective predictive analytics using their own customer data and available analytics tools to improve the inspection process and produce a more favorable hit rate than the physical approach.

Predictive models were built, tested, validated, and deployed through Microsoft Azure Machine Learning Studio (Azure ML), a cloud-based drag-and-drop machine learning platform with ready-to-use library of algorithms and modules. In this study, the two-class logistic regression algorithm via Azure ML (Two-Class Logistic Regression), which adopts an elastic net regularization method (Zou & Hastie, 2005), was used to detect unregistered PVIs. Elastic net, while addressing data heterogeneity and preventing overfitting through regularization, is ideal in achieving model parsimony and interpretability which are necessary requirements to align with the business understanding and decision-making. A synthetic minority oversampling technique (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) was also considered in the experiment in handling the highly imbalanced data brought about by the small percentage of net metering customers and previously identified unregistered PVIs.

2. Data

Data collected during the study were based on 30,421 residential customers living in high-end subdivisions under the DU's jurisdiction. Subdivisions are considered high-end when there is enough number of residents with high monthly electricity consumption. Based on the recommendations of business experts within the company and after several variable selection procedures, the following customer information using a snapshot of the customer database and other data sources were obtained: (1) average kWh consumption (12-month period); (2) contracted capacity (kW); (3) age of service (year); (4) age of contract (year); (5) number of billing discrepancies; (6) number of (other) end-users within the same street; (7) proximity to the nearest end-user (kilometer); and (8) net-metering indicator. These variables were also expected to be present in the final model.

The net-metering indicator indicates if a customer is registered under the net-metering program. To gather more information on and better understand the behavior and profile of solar PV panel users through an SML model, unregistered PVIs that were identified and monitored during the pilot study are tagged positive under the net-metering indicator. Hence, the net-metering indicator is instead modified to "end-user" indicator, referring to all customers with confirmed PVI, registered or unregistered. The final dataset contains 177 labelled end-users which represent only 0.58% of the target population. Summary measures (number of records and averages) are presented in Table 2.1.

Table 2.1 Descriptive Summary

End-User Indicator	1	0
Number of records	177	30,244
Average kWh consumption	1,612.0282	1,095.0542
Average contracted capacity (kW)	25.8820	10.6770
Average age of service (year)	18.4778	23.3082
Average age of contract (year)	9.3515	19.3544
Average number of billing discrepancies	0.0395	0.6055
Average number of (other) end-users within the same street	0.2599	0.2237
Average proximity to the nearest end-user (km)	0.2172	0.2908

The distance variable is calculated using the haversine formula (Inman, 1835) based on the meter location of a customer and the nearest end-user, regardless of the end-user's profile and location. Given the average distance, PVIs seem to emerge close to each other, with less than a kilometer apart. Although these installations are more clustered within an area, they may not necessarily be on the same street. This scenario is quite evident in Figure 2.1. Note that the variable for the number of (other) end-users within the same street does not include the observation in the count when it is currently tagged as an end-user itself. Basically, this is to properly assess the influence of end-users to the customer.

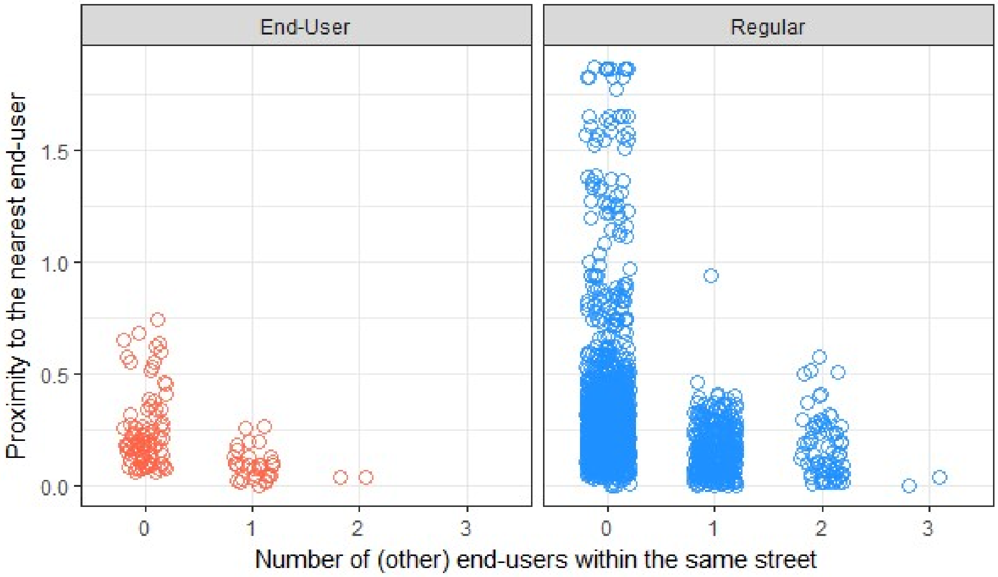

Figure 2.1 Density of PVIs of End-Users and Regular Customers

3. Methodology

The logistic regression is a generalized linear model (Nelder & Wedderburn, 1972) used in predicting the probability outcome of a binary response variable, and is commonly preferred due to its simplicity and interpretability. In this study, logistic regression is utilized via Azure ML to detect which among the customers are more likely to have PVIs based on their available information.

Let the response variable Y be the end-user indicator, where $Y = 1$ for end-users and $Y = 0$ for regular customers. We define the logistic regression model as:

$$\log\left[\frac{\mathrm{P}(\mathrm{Y}=1 \mid \mathbf{X})}{\mathrm{P}(\mathrm{Y}=0 \mid \mathbf{X})}\right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_7 X_7 \tag{3.1}$$

The model in equation (3.1) regresses for the log odds given a set of predictors $\mathbf{X} = (X_1, X_2, \dots, X_7)$ which signifies the variables mentioned in the previous section and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_7)$ which represents the regression coefficients. The $\boldsymbol{\beta}$ estimates are usually obtained by maximizing the likelihood function:

$$\mathrm{L}(\boldsymbol{\beta},\mathbf{X},\mathbf{Y}) = \prod_{i=1}^{n} \mathrm{P}(\mathrm{Y} = y_i \mid \mathbf{X} = \boldsymbol{x}_i) \tag{3.2}$$

$$= \prod_{i=1;\, y_i=1}^{n} \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_7 x_{i7})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_7 x_{i7})} \prod_{j=1; y_j=0}^{n} \frac{1}{1 + \exp(\beta_0 + \beta_1 x_{j1} + \cdots + \beta_7 x_{j7})} \tag{3.3}$$

Azure ML's two-class logistic regression model automatically standardizes the variables using a min-max normalization method. To prevent overfitting, the algorithm applies a naïve elastic net regularization framework (Zou & Hastie, 2005) by penalizing the earlier objective function:

$$\ell(\boldsymbol{\beta}, \lambda_1, \lambda_2, \mathbf{X}, \mathbf{Y}) = \ell(\boldsymbol{\beta}, \mathbf{X}, \mathbf{Y}) + \lambda_2 |\boldsymbol{\beta}|^2 + \lambda_1 |\boldsymbol{\beta}| \tag{3.4}$$

where

$$|\boldsymbol{\beta}|^2 = \sum\nolimits_{p=1}^{7} \beta_p^2 \tag{3.5}$$

$$|\boldsymbol{\beta}| = \sum\nolimits_{p=1}^{7} |\beta_p| \tag{3.6}$$

Equations (3.5) and (3.6) are L_2 and L_1 penalties, respectively. The $\ell(\boldsymbol{\beta}, \mathbf{X}, \mathbf{Y})$ term in equation (3.4) is the corresponding negative log-likelihood function of equation (3.2). The $\boldsymbol{\beta}$ estimates are computed by implementing an L-BFGS optimization procedure (Nocedal, 1980). Once the estimates are calculated, the model can be written as:

$$\log\left[\frac{\mathrm{P}(\mathrm{Y} = 1 \mid \mathbf{X}^*)}{\mathrm{P}(\mathrm{Y} = 0 \mid \mathbf{X}^*)}\right] = \hat{\beta}_0 + \hat{\beta}_1 X_1^* + \cdots + \hat{\beta}_7 X_7^* \tag{3.7}$$

$$\frac{\mathrm{P}(\mathrm{Y} = 1 \mid \mathbf{X}^*)}{\mathrm{P}(\mathrm{Y} = 0 \mid \mathbf{X}^*)} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 X_1^* + \cdots + \hat{\beta}_7 X_7^*\right) \tag{3.8}$$

where $\mathbf{X}^* = (X_1^*, \ldots, X_7^*)$ represents the normalized explanatory variables and $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_7)$ are the estimated model coefficients. The penalized logistic regression model can be interpreted using the equation (3.8) which solves for the odds. The effect of each coefficient is determined by means of *ceteris paribus*, or holding constant all other predictors in the model. Probability score is calculated similarly by using the estimated odds.

To address data imbalance, with the end-users representing less than 1% of the target population, oversampling and undersampling techniques are considered in building the working dataset. The minority class is oversampled through a synthetic minority oversampling technique (SMOTE) while the majority class is undersampled through a simple random sampling process.

Essentially, SMOTE attempts to generate artificial observations based on the characteristics of the original minority samples (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). The sampling techniques are performed only after randomly splitting the data to train and test dataset stratified by the end-user indicator. Hyperparameter tuning is also implemented to determine the model with the best fit. Figure 3.1 presents the difference between the original data and the analysis data after SMOTE.

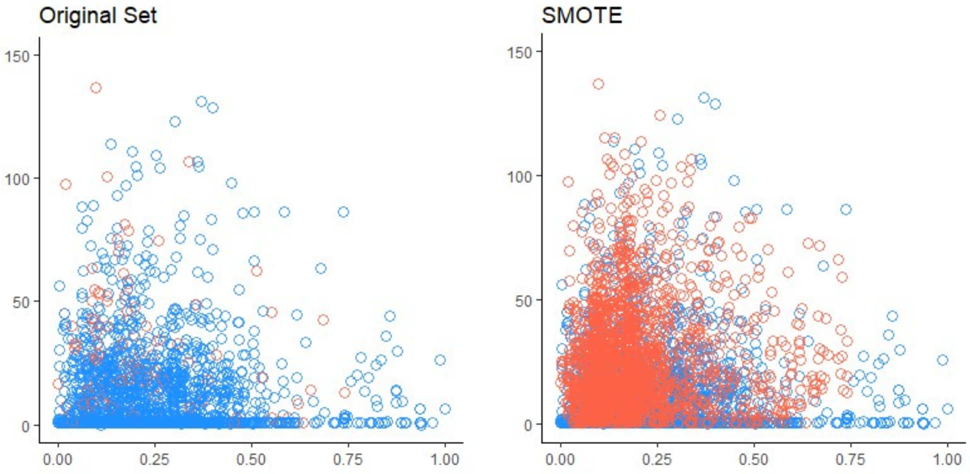

Figure 3.1. Implementation of SMOTE on actual dataset

4. Results and Discussion

The training dataset, generated through sampling techniques, has the following information in Table 4.1. Note that the engineered data does not deviate much from the original dataset in Table 2.1. However, SMOTE seems to have difficulty in fabricating the information on the number of (other) net metering customers within the same street, most likely due to the low variability in the feature and the rarity of confirmed end-users in general. Overall, the sampling techniques performed well in balancing the data.

Table 4.1 Data Summary with SMOTE and Random Sampling

End-User Indicator	1	0
Number of records	1,984	2,117
Average kWh consumption	1,658.8034	1,079.4119
Average contracted capacity (kW)	25.9112	10.8091
Average age of service (year)	17.6708	23.3962
Average age of contract (year)	9.7434	19.3790
Average number of billing discrepancies	0.0050	0.3042
Average number of (other) end-users within the same street	0.0938	0.2362
Average proximity to the nearest end-user (km)	0.1975	0.2856

The model is subject to 84.10% Area Under the Curve (AUC), 75.47% true positive rate, and 73.68% true negative rate on the test data using 50% probability threshold. Below is the final penalized logistic regression model.

$$
\begin{aligned}
\log\left[\frac{\mathrm{P(End{-}User = 1)}}{\mathrm{P(End{-}User = 0)}}\right] = {} & 1.27313 \\
& + 5.17663 \text{ (Average KWh Consumption)} \\
& + 5.65823 \text{ (Contracted Capacity)} \\
& - 0.30564 \text{ (Age of Service)} \\
& - 2.91369 \text{ (Age of Contract)} \\
& - 6.20447 \text{ (Number of Billing Discrepancies)} \\
& - 2.87240 \text{ [Number of (Other) End-Users Within the Same Street]} \\
& - 5.06105 \text{ (Proximity to the Nearest End-User)}
\end{aligned} \tag{4.1}
$$

The result of the detection model substantiated the customer profile described in the data summary in Section 2. From equation (4.1), it can be ascertained that the average kWh consumption and contracted capacity have positive influence on the probability of a customer using solar PV panels. Table 4.2 displays the effect of each variable to the odds of PVI assuming all other variables are held constant. A 100-kWh increment in the average energy usage means a 2.45% increase in the odds of PVI while an additional kW in the contracted capacity causes a 2.06% increase. There is a 0.62% and 7.5% decrease in the odds for each year older in the service age and contract age, respectively. A newer contract or a more recently created service indicates a very much higher probability that a customer is an end-user.

As observed earlier, energy usage fluctuation is quite uncommon in high-end subdivisions. So, it is somehow sensible that an increase in the number of billing discrepancies leads to a 46.23% decrease in the customer's odds of having PVI. Customers with billing discrepancy records are therefore expected to have lower probability scores due to this reason.

Table 4.2 Influence of Each Variable on the Odds of PVI

Variables (*unit*)	Factor
Average kWh consumption (*100 kWh*)	1.02447
Contracted capacity (*1 kW*)	1.02064
Age of service (*1 year*)	0.99379
Age of contract (*1 year*)	0.93027
Number of billing discrepancies (*1 count*)	0.53770
Number of (other) end-users within the same street (*1 count*)	0.38386
Proximity to the nearest end-user (*100 m*)	0.76291

It was initially established that PVIs normally appear close together in an area but it is quite uncommon that they are all on the same streets. The model says that a 100-meter nearer to the closest end-user raises the odds of PVI by 31.08%. This is high given the rarity of end-users, and this might also be the reason why an increase in the number of end-users on the same street decreases the odds of PVI by 61.61%, which is also quite extreme.

In summary, high odds of PVI is expected when a customer is relatively young, high-consuming, with consistent usage, and/or is closely surrounded by solar PV panel users.

5. Application

During a six-month post-study validation period, the probability scores of customers who had registered under the net-metering program following the experiment were examined. A high recall (true positive rate) was attained at 79.17% and 33.33% based on 50% and 80% threshold, respectively. Figure 5.1 displays the actual recall for each probability score threshold.

Finally, to make the field inspection feasible and to enhance the precision rate, the DU decided to set an 80% threshold to trim down the list of identified customers who are more likely to own unregistered PVIs. The final model (4.1) tagged 1,042 regular customers as possible end-users, excluding those that had already registered after the experiment's data gathering. These observations, based on the approximate meter locations, were then plotted in a map of the franchise area via a web-based dashboard used by the company so the business sectors could immediately tell which areas have high density of potential unregistered PVIs.

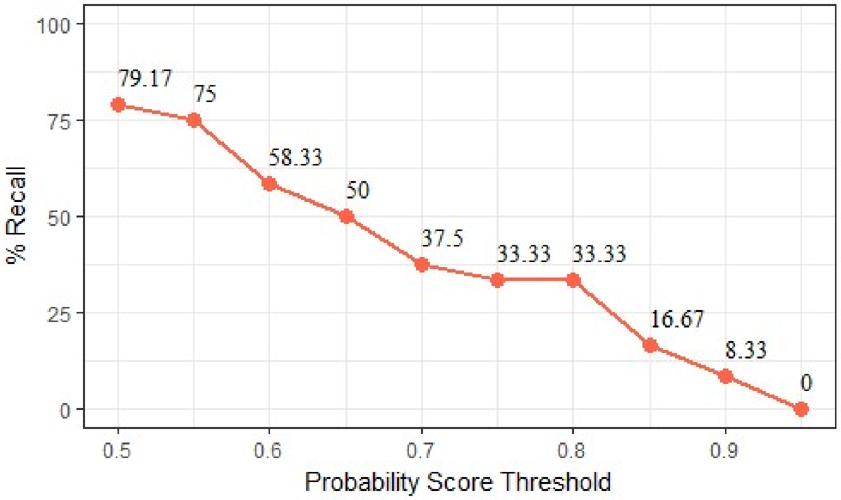

Figure 5.1. Post-study validation: recall by probability score threshold

The business sectors nominated PVI-dense subdivisions in major areas of the franchise and assembled inspection teams to validate the list of potential end-users. There were 225 residential customers tagged for field verification.

From Table 5.1, the result of the inspection based on advanced data analytics was remarkable. The study attained a significant hit rate which is 35.56 times better than the outcome of the manual search. Furthermore, there was a 96.55% improvement in the man-hours spent on field inspection.

Table 5.1 Field Inspection Results

Approach	Services Inspected	Unregistered PVIs	Hit Rate	Duration
Manual Search	7,693	25	0.32%	348 hours
Analytics	225	26	11.56%	12 hours

Using the probability scores produced by the SML model, the business sectors easily determined which areas to focus on and who among the customers to pay attention to. With a more efficient scheme to detect unregistered PVIs, the inspection efforts were optimized and a more substantial hit rate was achieved.

6. Conclusion

This study confirmed several advantages of utilizing advanced data analytics in detecting unregistered PVIs. The utility company established a more effective and workable inspection process with better results than the conventional, manual approach in terms of precision, work hours and labor cost. The results of the analysis not only corroborated the effect of each variable identified by the business experts but also determined the extent of their impact (e.g. proximity seems to be the major factor in detecting PVIs). With more end-users verified, the company could devise strategies to control the energy demand in areas with high density of PVIs, ensure network stability, and encourage electrical safety. It could produce better estimates on possible revenue loss from the foregone energy to alleviate the impact on sales and reassess its business model and capital investments. This study and the variables considered in the analysis may also be used as a guide for the company in building other new models on common business concerns, such as fraud detection, customer segmentation, and preventive maintenance.

Given the challenges in detecting PVIs, along with the current trend in RE and budding familiarity with solar energy usage, constant model refinement is recommended to guarantee a reliable predictive capability. With newly confirmed end-users added to the data, the model can get better in predicting unregistered PVIs. Additional variables commonly associated with any business (e.g. payment behavior) can be considered in the succeeding model iterations.

7. References

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321-357. doi:10.1613/jair.953

Inman, J. (1835). *Navigation and Nautical Astronomy: For the Use of British Seamen* (3 ed.). London, UK: W. Woodward, C. & J. Rivington.

Malof, J., Hou, R., Collins, L., Bradbury, K., & Newell, R. (2015). Automatic solar photovoltaic panel detection in satellite imagery. *2015 International Conference on Renewable Energy Research and Applications (ICRERA)*, (pp. 1428-1431). Palermo. doi:10.1109/ICRERA.2015.7418643

Nelder, J., & Wedderburn, R. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General), 135*(3), 370-384. doi:10.2307/2344614

Nocedal, J. (1980, July). Updating quasi-Newton matrices with limited storage. *Mathematics of Computation, 35*(151), 773–782. doi:10.2307/2006193

Two-Class Logistic Regression. (n.d.). Retrieved from Microsoft Azure: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-logistic-regression

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 67*(2), 301–320. Retrieved from https://www.jstor.org/stable/3647580