Paper 3406-2015

Modeling to improve the customer unit target selection for inspections of Commercial Losses in Brazilian Electric Sector - The case CEMIG

Sérgio Henrique Rodrigues Ribeiro, CEMIG; Iguatinan Gischewski Monteiro, CEMIG

ABSTRACT

Electricity is an extremely important product for the society. In Brazil, the electric sector is regulated by ANEEL (Agência Nacional de Energia Elétrica) and one of the regulated aspects is the power loss in the distribution system. In 2013, 13,99% of all injected energy was lost in the Brazilian system. The commercial loss is one of the power losses classifications, which can be countered by inspections of the electrical installation in the search for irregularities in power meters. The CEMIG (Companhia Energética de Minas Gerais) currently serves approximately 7.8 million customers, which makes unfeasible to inspect all customer units, in financial and logistic terms. Thus is essential the selection of the potential inspection targets. In this paper, logistic regression models, decision tree and Ensemble were used to improve the target selection process in CEMIG. The results indicate an improvement in the positive predictive value from 35% to 50%.

INTRODUCTION

Electricity is an extremely important product for the society, being used for simply to generate light or helping to keep a patient alive while allowing machines to operate in an intensive care unit in a hospital. In Brazil, the energy sector is composed by independent agents that operate in Power Generation, Transmission, Distribution and Commercialization. The ANEEL (Agência Nacional de Energia Elétrica) has the function to regulate, monitor e mediate all these stages of the power system as well the agents.

One of the parameters regulated by ANEEL is the power loss in the distribution network through the Normative Resolution 414/2010 (ANEEL, 2012). In case the energy utilities do not meet the losses limits stipulated by ANEEL, they must afford the difference. There are two categories of losses in the distribution network: Technical Losses and Non-technical Losses. Technical loss is related to the energy loss inherent to the conductors and equipment involved in the power distribution. Non-technical loss results from all the others kind of losses, obtained by the difference between the total loss and the technical loss (ANEEL, 2014). Also known as Commercial losses, the non-technical loss happens, for example, due to measurement errors of the energy consumed, power theft, energy meter frauds, errors on billing process, or energy meter failure. In Brazil, 13,99% of all energy injected in the system is lost, being 8,39% related to technical losses and 5,60% to commercial losses (ABRADEE, 2014). An essential procedure to reduce the commercial losses is the inspection of electrical installations in the customer units. The inspection is made by an electrician in the customer point of power delivery, seeking to identify any irregularities, like frauds or electromechanical failures of the meters. The inspection of all customers installations is financially and logistically unfeasible, which makes necessary the selection of some inspection targets.

CEMIG (Companhia Energética de Minas Gerais) is an important electric energy utilities in Brazil. CEMIG’s concession area extends throughout nearly 96.7% of the State of Minas Gerais, whose territory is situated in the Southeastern region of Brazil and covers an area of 567,478 thousand square kilometers, which corresponds to the territorial extension of a country the size of France. CEMIG has a predominantly hydroelectric energy matrix and produces electricity to supply more than 17 million people living in the state’s 774 cities and consume 45,089 GWh (INFOVEST CEMIG, 2014).

Basically, nowadays, the process of electrical installations target selection to inspections is as follow: an algorithm is applied to the data base composed by all 7.8 million installations. This algorithm, based in business rules, complaints made by electricians and variations on energy consumption over time, generates signs with inspections motives. Then, the responsible staff make the final selection analyzing the units based on the register information, consumption history and the results of the algorithm. The installations that they judge that have more possibility to revenue recover and more probability to find irregularities are selected to be inspected. Annually, about 60.000 units are inspected, of which 35% present commercial losses.

In this paper, logistic regression models, decision tree and Ensemble were used to improve the customer unit selection process to inspections. The SAS® Enterprise Guide 4.3 was used in the processing of the data base and creation of explanatory variables based on daily billed energy consumption. SAS® Enterprise Miner Client 6.2 was used to create and seek for the best model.

METHODOLOGY

The used data base contains the history of performed inspections between December 2009 and January 2014, comprising a total of 222.681 inspections. Also, 176 variables derived from the daily billed energy consumption (resulting by the division between the month billed energy and the number of billed days) and 9 variables related to electrical, locality and social-economical characteristics are present. The response variable of the model is given by the inspection performed in the energy meter, indicating whether or not there Commercial losses. This variable is binary and assumes the value "Founded" (Target = 1) where there is commercial loss or "Unfounded" (Target = 0) otherwise. The nominal explanatory variables are Number of Phases, Consumption Class, Network Regions and Sub regions.

The variables derived on the daily billed energy capture the consumption variation, seeking to explain the existence of the commercial loss. They are based on the change rate, means, standard deviation, coefficients of variation and counters, created on 24 consumptions months, which are represented by the Daily Billed Energy. Some examples of the created variables are as follow:

- M4TANO2: mean of the 4th quarter of the year 2.
- TXV4T: change rate between the mean of the 4th quarter of the year 2 and the 4th quarter of the year 1.
- TXV2SA2_4TA1: change rate between the mean of the 2nd semester of year 2 and the mean of the 4th quarter of year 1.
- CV_A2_EFD: Coefficient of variation of year 2.
- MEDIA_A1_EFD: Mean of year 1.
- TXD_2423: Change Rate of the 24th and 23rd months.
- TXD_24_A2T3: Change rate between 24th month and the mean of the 3rd quarter of year 2.
- N_TX_MENOR_N50: counter indicating how many monthly rates were less than -50%.
- N_RUIDO_50: counter indicating how many monthly rates were less than -50% or greater than 50%.

Wherein Year 2 refers to the most recent as the 24th Month, being considered a historical window of 24 months from the date the inspection was performed.

The base was divided in three: Training, Validation and Scoring. The scoring base separated for the final model evaluation contains data related to the inspections performed in the last three months (November2013. December2013, January2014) comprising a total of 13.131 inspections. The 209.550 inspections left use to the adjustments of the models are balanced, which means 50% were considered "Founded" and 50% were "Unfounded". Of these inspections, 70% were allocated in the training base and 30% in the validation, using the Data Partition Node.

The models adjusted in SAS® Enterprise Miner were two logistic models, which the second differs from the first since it allows interactions between the factors. Also, two decision trees were adjusted, and in the second the node division rule to the interval variables was changed to "Variance" (reduction in the square error from node means). To the variable selection, the logistic models used the stepwise algorithm and the decision trees used the Importance Value. After adjusting the four models, the Ensemble model was created based in the first ones. Finally, the five models were compared using the model comparison node, which has as criterion for selection of the best model the ROC curve. In this case, the model with the largest area below the ROC curve is chosen. An example of the diagram used in SAS® Enterprise Miner is shown in Figure 1:

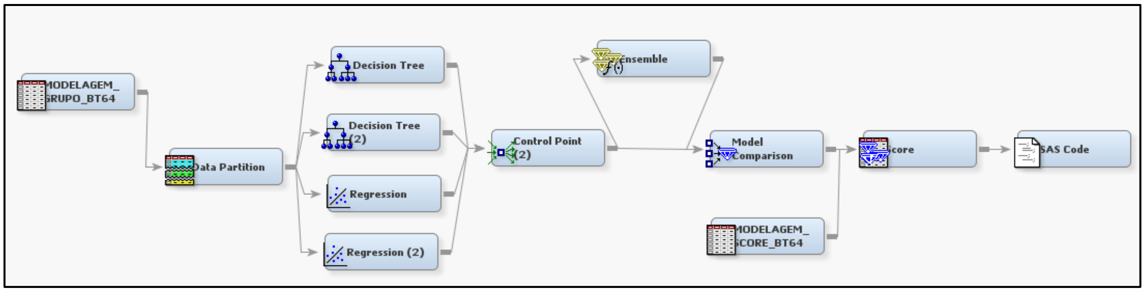

Figure 1: SAS® Enterprise Miner Modeling Diagram

In summary, following the flow illustrated in Figure 1:

1. The data base is divided in the bases Training and Validation bases (Data Partition Node), after the Scoring base being separated previously.
2. The four models are adjusted – 2 logistic models and 2 decision tree models.
3. The Ensemble model is adjusted (Ensemble Node).
4. The five models are compared and the best one is chose (Model Comparison Node).
5. The chose model is applied to the Scoring base (Score Node).
6. The data base with the models results is saved (SAS Code Node).

Then, an evaluation is made using the following statistics: Correct Classification, Sensibility, Specificity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV). To understand these statistics, consider the Table 1.

		Real		Total
		Positive (1)	Negative (0)	
Predicted	Positive (1)	**True Positive** a	False Positive b	a + b
	Negative (0)	False Negative c	**True Negative** d	c + d
Total		a + c	b + d	a + b + c + d

Table 1: Statistic Construction to evaluate the model

Table 1 compares the predicted results of the model and the observed in the reality during inspections, classifying as positive the “founded” units (with commercial losses) and as negative the “unfounded” units (without commercial losses). Models that maximize the True Positive and the True Negative are sought. The statistics are defined as:

- Correct Classification: (a+d)/N.
- Positive Predictive Value: a/(a+b).
- Negative Predictive Value: d/(c+d).
- Sensibility: a/(a+c).
- Specificity: d/(b+d).

The correct classification is a measure of global accuracy of the model. Given all evaluated installations, it counts, in percentage, the amount of units that the model correctly classified. It counts the true positive and the true negative.

The PPV is the chance of an unit to be “founded” given that the model classified it as “founded”, while the NPV, analogously, represents the chance of a unit to be “unfounded” given this classification by the model.

The sensibility represents the chance of the model to classify an installation as “founded” since it really is “founded”, i.e. is the chance of the model to detect commercial losses given that it exists in the unit. Similarly, the specificity is the chance of the model to classify an installation as “unfounded” given that it really does not present irregularities.

RESULTS

The model chosen, based on the ROC index, was the first Decision Tree, that uses the p-value of F-Test associated to the node variance as criterion of node partition to the interval variables and uses the p-value from Pearson Chi-squared statistic for target variable versus the branch node. The results of the logistic and Ensemble models were close, although a little inferior to the results from the Decision Tree model.

		Real		Total
		Positive (1)	Negative (0)	
Predicted	Positive (1)	**3.783**	3.805	7.588
	Negative (0)	764	**4.779**	5.543
Total		4.547	8.584	13.131

Table 2: Final Model Result for Scoring Base

Table 2 shows the results of the final model for scoring base. The observed evaluation statistics were:

- Correct Classification: Among the 13.131 evaluated units, the model correctly classified 8.562 (3.783 + 4.779) installations, getting a Correct Classification of 65%.
- PPV and NPV – Among 7.588 installations indicated as inspections targets by the mode, 3.783 (50%) really presented commercial losses. While, among the 5.543 units that the model did not indicated as a target, 4.779 (86%) didn’t present irregularities.
- Sensibility and Specificity – Considering the 4.547 units with Commercial losses, the model was capable to identify the existence of irregularities in 3.783 (83%). And among the 8.584 units that do not present non-technical losses, the model indicated the absence of commercial losses in 4.779 (56%).

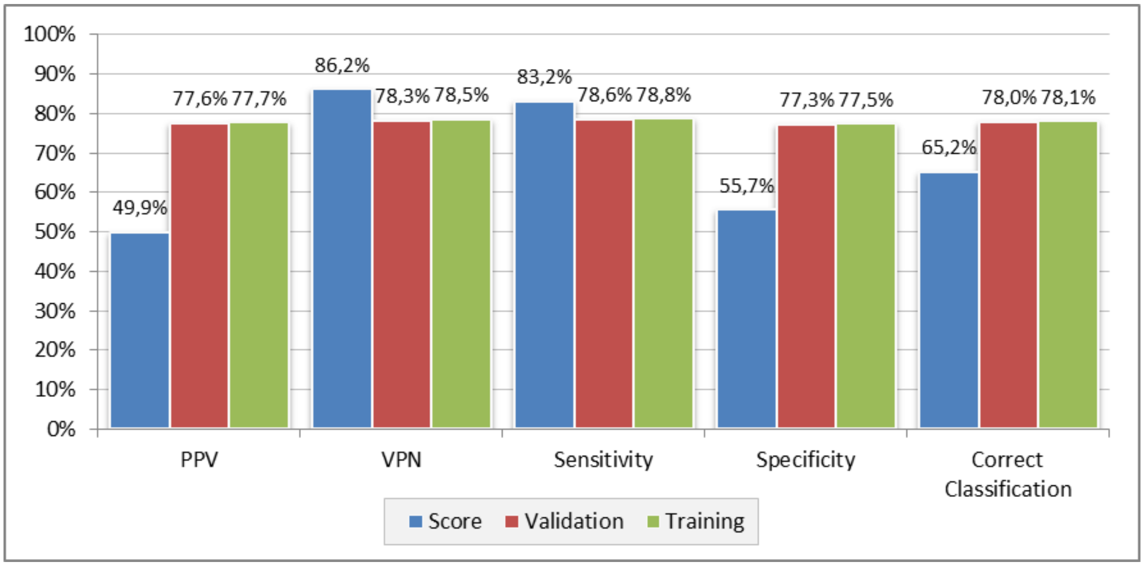

Figure 2: Evaluation Statistics of the final model

Figure 2 presents the evaluation statistics of the final model for all three data bases. The results for the Scoring base were discussed previously. Regarding the Training and Validation bases, all statistics are around 79%. Thus, good results were obtained with the model.

EXPECTED FINANCIAL RETURN

A comparison between the currently results of the practiced process in CEMIG and the results obtained by the model was made. Currently, 35% of the inspections performed by CEMIG find units with Commercial Losses. In the tests with the model, this percentage increases to 50%. Monthly, about 5.000 inspections are performed. Therefore, assuming this historical percentage of 35%, it is expected to find 1.750 units with irregularities, while with the model results this number would increase to 2.500. In other words, the model would represent an increase of 750 units with commercial losses detected by month.

Historically, we know that around 40% of the energy meters with Commercial Losses are liable to revenue recovery after the evaluation in laboratory. Hence, considering the 750 units, it is expected that about 300 installations would be likely to be charged to revenue recovery. In average, R$ 518 are recovery by unit. Thus, for this units, the expected recovery would represent R$ 155.440,00 monthly, or R$ 1.865.280,00 per year.

APPLICATION IN 2014

In order to complement the analysis, the final model was applied to April, May and June, for the units of the cities within the metropolitan region of Belo Horizonte, that internally in CEMIG, is divided between the SM/MP and SM/MT sub regions. The observed PPV was 48%, meeting the observed PPV of the scoring base of 50%.

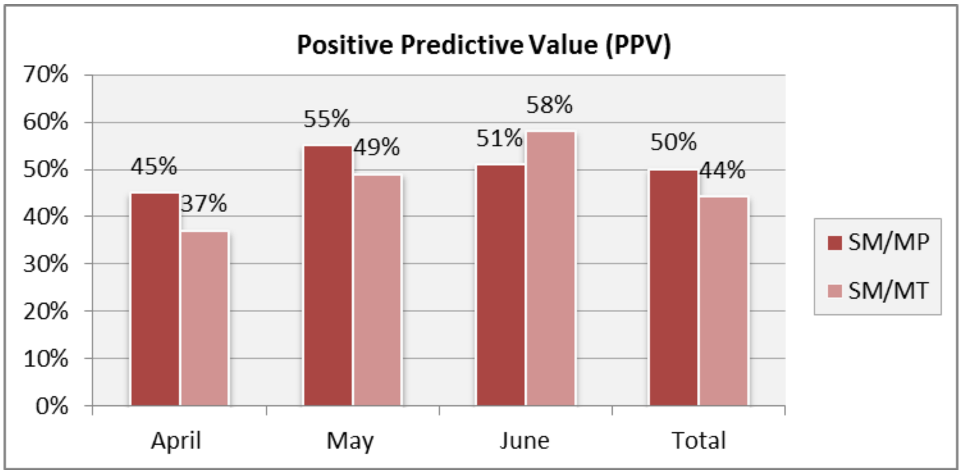

Figure 3: PPV of final model, applied to April, May and June of 2014

Figure 3 shows the PPV in each of these months, observed for both sub regions. In SM/MP the PPV was 50% and in SM/MT of 44%.

CONCLUSION

The results indicated that the model is capable to improve the target selection process aiming inspections, reducing the costs of "unfounded" inspections and increasing the revenue recovery by increasing the inspections accuracy.

Some other tests performed in the modeling process include the increase of the historical window to 64 consumption months, neural network models, transformations of the explanatory variables, use of Principal Component Analysis based on explanatory variables, creation of different models for each region of operation, models adjustments based on a random sample enabling unbiased inferences about the entire area.

REFERENCES

Agência Nacional de Energia Elétrica – ANEEL. 2014. "Perdas de Energia" Accessed October 21, 2014. http://www.aneel.gov.br/area.cfm?idArea=801/.

Agência Nacional de Energia Elétrica – ANEEL. 2012. *Resolução Normaiva nº 414/2010 – Condições Gerais de Fornecimento de Energia Elétrica.* Brasília: ANEEL.

Associação Brasileira de Distribuidores de Energia Elétrica - ABRADEE. 2014. "Furto e Fraude de Energia" Accessed October 20, 2014. http://www.abradee.com.br/setor-de-distribuicao/perdas/furto-e-fraude-de-energia.

INFOVEST CEMIG. 2014. "Who We Are" Accessed October 21, 2014. http://cemig.infoinvest.com.br/?idioma=enu.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Sérgio Henrique Rodrigues Ribeiro
Companhia Energética de Minas Gerais - Cemig
+ 55 31 3506-2115Phone
sergio.hribeiro@cemig.com.br