Prediction Model of Electricity Energy Demand for FCU in Colombia Based on Stacking and Text Mining Methods

Javier H. Velasco Castillo(✉) and Andrés M. Castillo

Escuela de Ingeniería de Sistemas y Computación,
Universidad del Valle, AA 25360, Cali, Colombia
javier.hernan.velasco@correounivalle.edu.co

Abstract. Electric energy demand prediction models are used by energy marketers to plan the demand hour by hour of the following week. Using these predictions, the company of experts in markets (XM) plans the allocation of the generation of electric power to the different generators connected to the network. The precision of these models is extremely important because with these the purchase of energy is planned and this allows the control of operating costs to be controlled. A bad prediction by the FCUs will be put to detailed monitoring before the National Operation Council (NOC) due to the performance of the forecasts, in this article we will present some of the models that have been implemented to carry out the demand forecasts of energy, and a model is proposed using stacking methods (based on statistical methods) and text mining, which improves the deviations of energy demand forecasts and that could be implanted by different FCU in Colombia.

Keywords: Prediction model · Statistics model · Electricity energy demand Forecast · Text mining · Stacking

1 Introduction

The short-term planning of electricity demand in Colombia is a task under the administration of the Company of Experts in Markets (XM), which is a subsidiary of the Colombian state transmission company ISA. The company is in charge of receiving the daily offers that the generators present in the energy exchange and it plans the purchase for each generator hour so that it can meet the demand for energy the following day[1]. The current regulation and the National Operation Council (NOC) through the agreement 349 of 2006 approves the creation and modification of the demand Forecast Control Units (FCU) to guarantee the delivery of the electricity demand prediction to the economic dispatch of XM. In the short-term planning of XM, it must report the demand prediction for the following week no later than Friday at 1:00 p.m. of the week prior. This must be done to determine the demand hour by hour

[1] http://www.xm.com.co/corporativo/Paginas/Nuestra-empresa/que-hacemos.aspx.)

J. E. Serrano C. and J. C. Martínez-Santos (Eds.): CCC 2018, CCIS 885, pp. 291–300, 2018.
https://doi.org/10.1007/978-3-319-98998-3_23

of the week, to determine the amount of energy to negotiate with the generators connected to the electrical system, to program the generation units, to fix the spot price of electricity supply in the market and to avoid the cost overrun in the production of energy [7].

The energy generated cannot be stored at low cost like other products. It must be attended to instantaneously and the Colombian electricity system must be able to meet the demand [3]. For this reason, the prediction models of each FCU must represent an accurate demand to avoid large deviations in the predictions that may imply large operating costs [7]. The FCU have focused on reducing the percentage of deviation of the demand forecast, so they have proposed different models and have combined different methods. However, these all conclude that there are complex external factors that affect their predictions. Each FCU has tried to include this knowledge in their models, sometimes through human appraisals, but it has not been possible to systematically include these factors in the prediction models.

Because of this, each FCU has implemented its own demand prediction model based on its knowledge of the business and the expertise of its analysts. Most use statistical methods (Least Squares, Medium, Average [2] and Moving Averages) combined with knowledge of environment, such as, consider holiday periods, periods of drought or strong winters, etc. However, considering and modeling all external factors is a complex task since there are no reliable systems in Colombia that record them. In this document we describe a demand prediction model based on text mining methods that can infer external conditions from climate predictions made by several internet pages and that when combined with statistical methods and classification, allow to increase the accuracy of all prediction models tested with historical data.

2 Background

Within the demand prediction models most used are statistical models (their basis is historical information), empirical models (which depend on intuition and human judgment) and artificial intelligence [3]. Statistical models are a great predictive tool, but it is noteworthy that energy demand is a dynamic stochastic process composed of several individual components that can be influenced by several external factors related to energy consumption in a given hour such as climate, temperature, economy, demography, important sport events, exodus of people, etc. [9].

It has been determined that there are patterns that repeat over time [9]. Therefore, a method of extracting these patterns is required to validate the historical data and collaborate to generate more accurate predictions. Hence, the use of extraction methods such as text mining for the representation of words in vectors is proposed [4]. A clear example is documented in the studies carried out by Mesa Bustelo [15], which show how the variation in the temperature in Madrid affects the energy demand either in a positively or negatively manner. The variation in this case is estimated to be at around 12%, and the explanation is that the lower the temperature, the lower the energy consumption and the higher the temperature, the higher the energy consumption due to the use of air conditioners. So far, several models, methods and techniques have been

proposed to predict energy demands and this is how XM provides a base forecast of energy demand approximation for each FCU.

2.1 ARIMA and SARIMA Model

The Autoregressive Integrated Moving Average model (ARIMA) is a time series analysis characterized by its application to make predictions where data with trends are considered [17]. The ARIMA model applied in the prediction of energy demand has drawbacks due to structural changes in the trajectory of the series [18]; in addition, the model must be adjusted at each hour of the day [1]. Other studies found that the Seasonal Moving Autoregressive Integrated Moving Average model (SARIMA) is a prediction model which closer prediction the reality of the behavior of daily energy. SARIMA better explains the structural changes in a time series by incorporating significant variables such as delays, fictitious or level variables [10].

The classic ARIMA model requires numerous historical data for its estimation. In addition, factors such as season, climate, temperature, among other aspects, affect the demand, but these cannot be easily incorporated since factors are related to energy demand, and each relationship is assigned a certain weight of energy. According to the new predictions that will be taken in the week S + 1, we can take as an example the cloudiness in hours where the sun rises or is hidden since at a lower light intensity there is a greater probability of greater energy consumption.

2.2 Dynamic Bayesian Model

The dynamic Bayesian model tries to predict the demand of week t + 1, from the previous n days. In a traditional Bayesian model, it is necessary to know in advance the probability distributions of the data. This prior knowledge is usually determined by specific knowledge of the problem, or by using a significant sample of the data. However, in cases in which the priori distribution of the data is not known and there is not sufficient data to estimate it, a set of "artificial data" can be constructed using Markov chains. The parameter of the artificial data are optimized using simulation location of Monte Carlo, to make an estimate of the apriori distribution. This distribution must be recalculated each time from the available historical data and hence the term "Dynamic" of the method [11]. The Bayesian models are also used to evaluate the distance between the predicted value and the real value. This model generates an adjustment component that evolves according to the behavior of the demand, and as information is obtained, the a priori distribution is the analyzed again [8]. This model makes predictions with historical data of the last 15 days. It does not reflect the real trend of energy demand in the long term, and does not consider the influence of some events related to the time and the celebrations of the country. For example. in Fig. 1 we can see that during the holidays of 2017 there are decreases in consumption, since when starting school holidays, the demand for energy is lower at 6 a.m.

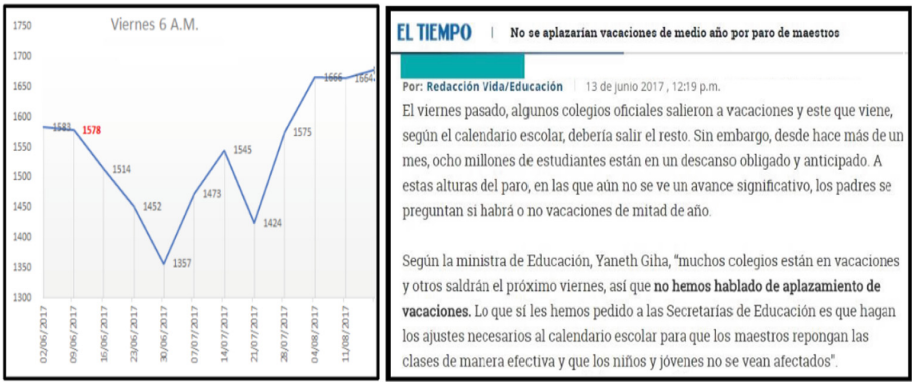

Fig. 1. Relationship Day Friday 6 a.m. FCU Codensa June 2017 and Note from the newspaper EL TIEMPO about vacations in June 2017

2.3 Model of Artificial Neural Networks

Computational models such as artificial neural networks (ANN) have been used to construct models that improve the prediction of demand of energy, with the advantage that they consider the relationship between variables as a non-linear function [6, 12]. The use of artificial intelligence for normal weeks is reliable, but it is deficient for atypical weeks such as Easter and holiday periods. For this reason, it is necessary to make the separation of weeks in groups [9]. For example: the demand during normal weeks, weeks with holiday Mondays, weeks during the holiday period, week corresponding to Holy Week. And weeks with holidays different from Monday. Because the training data in the RNAs for the last three groups are scarce and the results are deficient [9], they also do not take into account external factors that affect the prediction and only work with historical data for their training.

2.4 Regression Model

Regression models analyze the effect of a series of explanatory variables x on a response variable y [15]. Regression models can be simple or multiple. In simple regressions we have a single predictor variable but when two or more predictor variables are required, a multiple regression is used [16]. The linear regression method is one of the methods applied by the Codensa CPU for the realization of the demand prediction per hour and has 15 historical data as shown in the Fig. 2 [2].

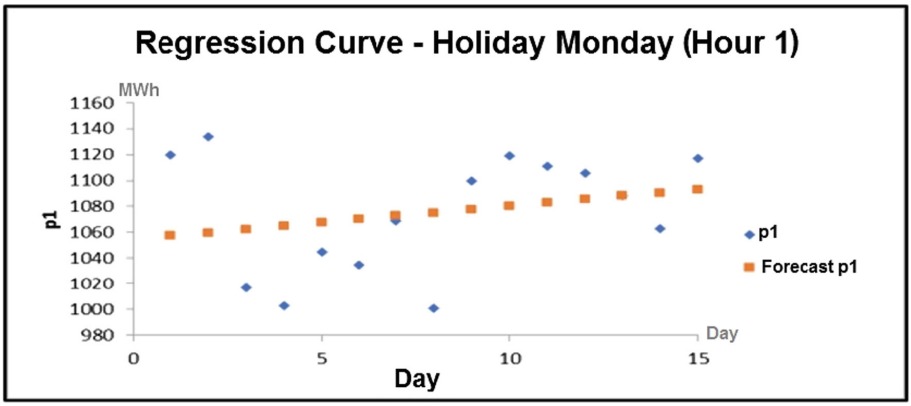

Fig. 2. Simple linear regression curve

3 Proposed Model

A model for the prediction of energy demand in Colombia is proposed based on obtaining information from historical data corresponding to border reports of the FCU and published in the XM portal. These reports are affected by programmed jobs and unmet demand (UD). This data is the property of the FCU for the calculation of the real demand. The demand for energy is related to the location, the scheduled jobs (are framed by requests from the industry, projects and network maintenance) and unscheduled jobs (events that are monitored by the control center in the meters in voltage level 2, 3 and 4). The location of the energy demand is in commercial, industrial and residential sectors, the latter burden can be concentrated only in the headwaters as well as they can be dispersed between the populated areas and the rural areas. For this reason, concentration in residential sectors is more affected by external factors that influence the prediction of demand either positively or negatively; in the extraction of knowledge, some stages have been defined, where one of the stages is data mining whose objective is the identification of valid patterns and models, based on prediction techniques such as text mining that focuses on data textual [13] Social networks such as Twitter provide unstructured information and is a great source of subjective information in real time, because millions of users share opinions on Twitter and all this information is downloaded and processed with a high volume of historical data [14].

The propose model (see Fig. 4) has its beginning from the methodology and statistical methods used by the FCU UCodensa [2]. from where 3 of the statistical models were extracted:

- Method 1: Minimum Squares with Total Daily Energy: Last 5 days of data, the total energy is predicted with respect to the median of the last 5 data and weighted each hour with respect to the last data of the day.
- Method 2: Minimum squares per hour: Last 5 data, is calculated with linear regression for each hour.
- Method 3: Minimum squares per hour: Last 12 data, is calculated with linear regression for each hour.
- Method 4: Median: Last 15 data, it is calculated with the median for each hour.

To determine the data to be used, we started from the XM studies, where they determined that there are types of days (42), that is to say that the behavior of a Sunday before a holiday Monday is different from a Sunday before a normal Monday. A principal component analysis (PCA) was carried out on the information of the Codensa FCU, obtaining as a result that Monday, Saturday and Sunday are easily separable, but on Tuesdays and Fridays it is not possible to separate them as observed in the Fig. 3 so it was concluded that three types of algorithms can be used:

- Algorithm 1: Similar Days: Days of the previous weeks that present the same type of day.
- Algorithm 2: Days with Similar Weeks: Weeks with the same types of days, as an example: M-T-W-R-F-S-U, M-T-W-R-F-SBHM and UBHM.
- Algorithm 3: Tuesday, Wednesday, Thursday and Friday as Similar Days: Data are selected from only those days and weight is given to the day to be predicted.

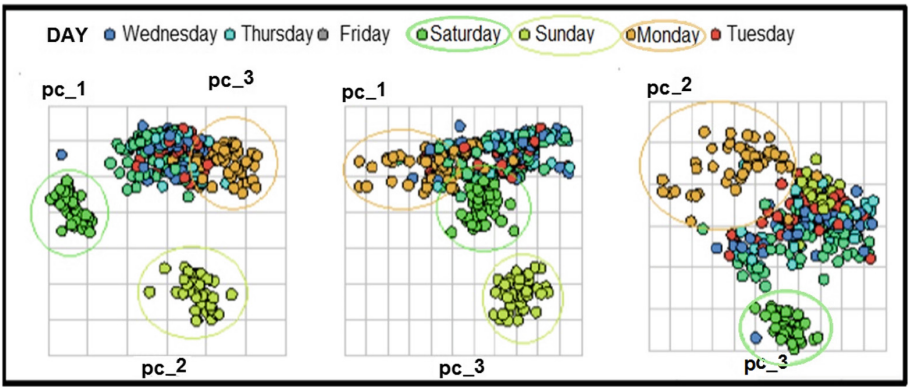

Fig. 3. PCA data FCU Codensa.

As a first phase we have the result of the combination of the 3 algorithms with the 4 methods, for a total of 12 prediction combinations and for each algorithm the median will be calculated, generating 3 new predictions for a total of 15 predictions; after comparing the values obtained in the pre-liminal tests against the real values, the values proposed by XM and the one reported by the FCU Ucodensa, similar results are obtained.

For the second phase, the classifier in the meta-level will combine 12 predictions (algorithms for methods) and then a plurality vote is take. Each prediction emits one vote and the prediction with the most votes is selected. This works well when the predictions have an acceptable precision, otherwise meta-learning will be used, which consist of learning how to integrate the results from multiple algorithms of learning, using an arbitrator that is a classifier (Artificial Neural Networks) that is trained to solve differences between the 12 predictions [5]. As a result of the Stacking method, the energy demand forecast of week S + 1 will be obtained. The analyst responsible for

carrying out the demand forecast configures the parameters previously and modifies the forecast according to the Energy Not Supplied (ENS) for Programmed Jobs, finally a message will be displayed with the external factors that could affect the forecast (data previously stored in Basis of Data).

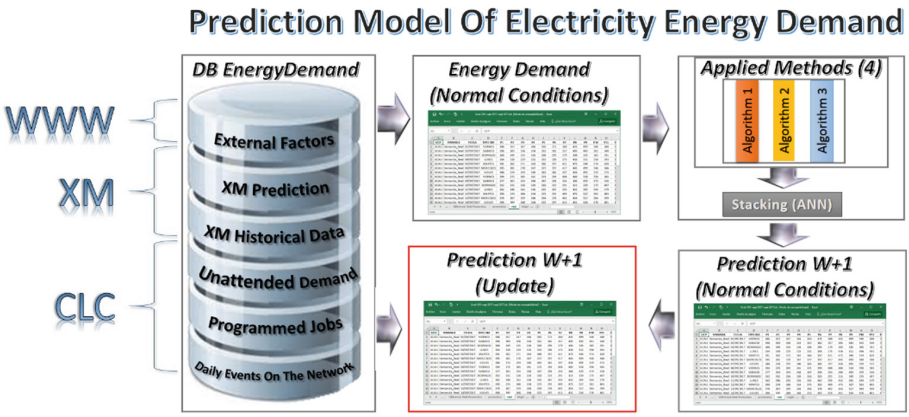

Fig. 4. Proposed energy demand prediction model

4 Results

To test the model we selected 9576 records from the XM dataset corresponding to Codensa collected from 1st February of 2017 to 30th April of 2018. The records corresponding to December and January were excluded because there were not enough data to train a model for these type of days. The data set was split into train (70%) and test (30%). Each data file is composed of 14 columns: day, hour, pred_1, pred_2, pred_3, pred_4, pred_5, pred_6, pred_7, pred_8, pred_9, pred_10, pred_11, pred_12, codensa_prediction, real_consumpsion.

Several feed forward ANNs that were trained by the back propagation algorithm have been assessed using the training dataset and a 10 k-fold cross validation. We select the ANN that better perform a regression model on the training data. The cross validation range the parameters from 1 to 14 neurons in the input layer, and from 1 to 12 in the hidden layer, using the mean root square error as target function. The source code is available in github at: https://github.com/javiervelasco/RNA_Backpropagation.

Once we select the best ANN for this problem, we apply it to the test data set, for making the prediction of the 2856 records. We compared the resulting predictions against the Codensa predictions for the same records, using the average daily cumulative error as performance descriptor. We compare the overall accuracy of both predictors, and also the accuracy depending on the type of day and hour of the day.

The training was perform in Desktop PC with an Intel Core i3, and 4 GB of RAM, running on Windows 8.

5 Discussion of Results

The 10 fold cross-validation yields a ANN with 6 neurons in the input layer, 10 neurons at the hidden layer, and 1 neuron at the output layer. The training took 16109 iterations to reach an average accuracy of 0.9946. These preliminary results have shown that the proposed staking model can outperform the Codensa model for days between Monday to Thursday, but still presenting high discrepancies for the weekend days (Fig. 5).

Fig. 5. Cumulative error by type of day for the model and for the Codensa model

In the Fig. 6, we show the hour by hour prediction of the ANN and the Codensa model versus the real consumption reported by XM. On those 4 types of days is clear how the ANN model predict in a very accurate way the XM data.

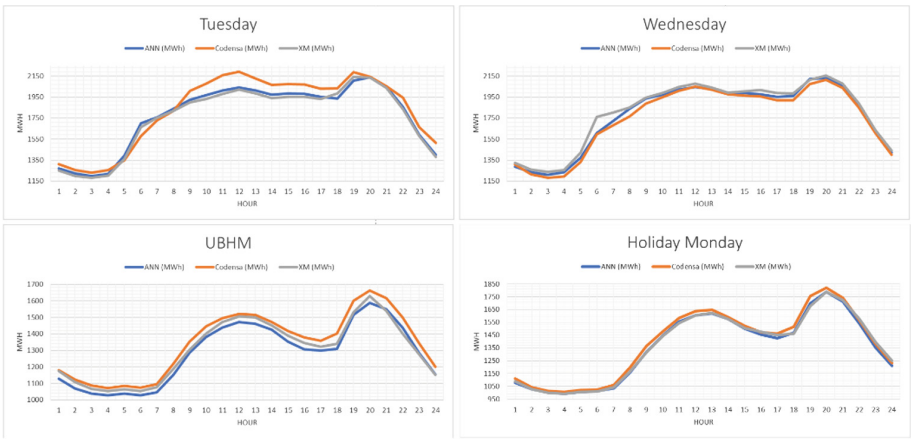

Fig. 6. Prediction Tuesday, Wednesday, UBHM and Holiday Monday

6 Conclusions

The implementation of the four methods to simulate the predictions of Codensa led us to investigate what type of data were used as input to make the prediction. As a first source of information, the real data reported by Codensa to XM was used. As a result, the prediction values of Codensa are above the National Office Center (NOC) proposal (reference data) and the preliminary test showed values similar to the NOC proposal. The results obtained are partial, due to the fact that the research has not been completed, but promising results have been obtained since, with our current model, we can make forecasts for any FCU that are in the worst case as good as those of XM and that on average show a decrease in MAPE. It is also clear that our current approach of having a single ANN for all the days is the not best choice, because as shown by the PCA there is at least 2 groups of days that behaves very similar. In the next steps of the work, we will train different ANNs for the different groups of days.

References

1. Barrientos, A.F., Olaya, J., González, V.M.: Un modelo spline para el pronóstico de la demanda de energía eléctrica. Revista Colombiana de Estadística **30**(2), 187–202 (2007)
2. CODENSA, Grupo Enel: Metodología pronóstico oficial de demanda Codensa. https://www.cno.org.co/sites/default/files/documentos/noticias/Pronostico_Codensa_22072016_v1.pdf. Visitado 29 Sept 2017
3. Franco, C.J., Velásquez, J.D., Cardona, D.: Micromundo para simular un mercado eléctrico de corto plazo. Cuadernos de Economía **31**(58), 229–256 (2012)
4. García Gutiérrez, Á.: Machine learning en bases de datos de lenguaje natural. Trabajo de grado en Ingeniería Informática y Matemáticas, Escuela Politécnica Superior y Universidad Autónoma de Madrid (2016)
5. Moreno Garcia, M.N., Segrera Francia, S.: Multiclasificadores: métodos y arquitecturas. Departamento de Informática y Automática Universidad de Salamanca (2006)
6. Ortiz Parra, D.A.: Aplicación de redes neuronales artificiales en el pronóstico de la demanda eléctrica a corto plazo en el sni. http://dspace.ups.edu.ec/bitstream/123456789/6672/6/UPS-KT00835.pdf
7. Rueda, V.M., Velasquez Henao, J.D., Franco Cardona, C.J.: Avances recientes en la predicción de la demanda de electricidad usando modelos no lineales. DYNA **78**(167), 36–43 (2011)
8. Boada, A.: Modelo Dinámico Bayesiano y Modelos Estadísticos Predictivos. Bayesian Dynamic. Dimensión Empresarial **15**(1), 25–41 (2016)
9. Sarmiento Maldonado, H.O., Villa Acevedo, W.M.: Inteligencia artificial en pronóstico de demanda de energía eléctrica: una aplicación en optimización de recursos energéticos. Revista Colombiana de Tecnología Avanzada **2**(12), 94–100 (2008)
10. Tabares Muñoz, J.F., Velasquez Galvis, C.A., Valencia Cardenas, M.: Comparación de técnicas estadísticas de pronóstico para la demanda de energía eléctrica. revista Ingeniería Industrial-Año **13**(1), 19–31 (2014)
11. Valencia Cardenas, M., Correa Morales, J.: Un modelo dinámico bayesiano para pronóstico de energía diaria. Revista Ingeniería Industrial **12**(2), 7–17 (2013)

12. Vasquez Macedo, R.R., Mamani Huacani, D.N.: Pronóstico de la demanda de energía eléctrica para Bolivia Aplicación de inteligencia artificial. Instituto de Estudios Sociales y Económicos IESE–UMSS, Instituto Nacional de Estadística INE, Agosto (2013)
13. Mariñelarena, D., Errecalde, M.L., Castro Solano, A.: Extracción de conocimiento con técnicas de minería de textos aplicadas a la psicología. Revista Argentina de Ciencias del Comportamiento **9**(2), 65–76 (2017)
14. Maté, A., Peral, J., Fernandez, A., Gil, D., Trujillo, J.: A hybrid integrated architecture for energy consumption prediction. Future Gener. Comput. Syst. **63**, 131–147 (2016)
15. Fernández De Mesa Bustelo, M.: Análisis y mejora de la predicción de la demanda eléctrica en periodos de alto ECM. Info:eu-repo/semantics/bachelorThesis, julio de 2016
16. Carrasquilla Batista, A., Chacon Rodriguez, A., Nuñez Montero, K., Gomez Espinoza, O., Cerdas, J.V., Guerrero Barrantes, M.: Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal. Revista Tecnología en Marcha **29**(8), 33–45 (2016)
17. Amaris, G., Ávila, H., Guerrero, T.: Applying ARIMA model for annual volume time series of the Magdalena River. Tecnura **21**(52), 88–101 (2017)
18. Broz, D.R., Valentina, N.V.: Predicción de precios de productos de Pinus spp. con modelos ARIMA. Madera y bosques **20**(1), 37–46 (2014)