Data Mining based on Random Forest Model to Predict the California ISO Day-ahead Market Prices

Ashkan Sadeghi-Mobarakeh, Mahdi Kohansal, Evangelos E. Papalexakis, and Hamed Mohsenian-Rad
Department of Electrical Engineering and Computer Science, University of California, Riverside, CA, USA
e-mails: {asade004, mkoha002}@ucr.edu, epapalex@cs.ucr.edu, and hamed@ece.ucr.edu

***Abstract*—In this paper, an ensemble learning model, namely the *random forest* (RF) model, is used to predict both the exact values as well as the class labels of 24 hourly prices in the California Independent System Operator (CAISO)'s day-ahead electricity market. The focus is on predicting the prices for the Pacific Gas and Company (PG&E) default load aggregation point (DLAP). Several effective features, such as the historical hourly prices at different locations, calender data, and new ancillary service requirements are engineered and the model is trained in order to capture the best relations between the features and the target electricity price variables. Insightful case studies are implemented on the CAISO market data from January 1, 2014 to February 28, 2016. It is observed that the proposed data mining approach provides promising results in both predicting the exact value and in classifying the prices as low, medium and high.**

***Keywords*: Electricity market price regression, classification, data mining, random forest, California ISO day-ahead market.**

I. Introduction

Forecasting the market price is a key factor for the decision makers in determining the short term operating schedules and bidding strategies in the electricity markets. For example, a transmission company is interested in knowing the exact value of the future price to strategically bid into the market. As another example, a demand response market participant is interested in knowing whether the price is high or low to optimize its operation [1]. In this case, the price value forecasting problem reduces to a price classification problem.

One thread of related work focuses on forecasting the *exact value* of the electricity price using data mining techniques, e.g., in [2]–[7]. A method based on pattern sequence similarity was presented in [2]. In [3], the authors applied the regression trees and normalized radial basis function networks to the New England market price data. A neural network model was applied by [4], [5] to forecast electricity prices. Gonzalez et. al in [6] applied different regression methods using tree-based models to forecast the electricity price for the Spanish-Iberian market. The features such as load, hydro and thermal generation and wind energy production are considered. In [7], the authors applied a gradient boosting regression technique to forecast the exact value price. The average mean absolute error (MAE) was 7.13 \$/MWh which outperformed the average MAE, 8.64 \$/MWh, using ARMAX. A review of the state-of-the-art in forecasting the exact values of the electricity market prices is given in [8].

There is another, but less explored, thread of related work that focuses on electricity price *classification*, e.g., in [1], [9], [10]. The authors in [9] proposed a classification of future electricity market prices using support vector machines with data from Ontario and Alberta electricity markets. The authors in [10] addressed the short-term energy price classification based on the decision tree method. Two different methods for classifying the prices were applied in [1]. In the first one, the exact value of the electricity market price was obtained using multilayer perceptron (MLP) and then it was classified based on the pre-specified threshold. In the second one, the class labels of the prices were directly obtained using three techniques: Decision Trees (DT), Naive Bayes (NB), and K-Nearest Neighbor (KNN). The results showed that the second method outperforms the first one.

While data mining has previously been used in forecasting electricity prices, there is still great potential to enhance performance. In fact, in this field, the devil is in details. Specifically, the performance of a data mining approach highly much depends on: 1) the features selected, 2) the learning algorithms, and 3) the market under study. Furthermore, the results in one market can not be necessary generalized to other markets. For example, using the same features and approach, MAEs were obtained as 2.37 \$/MWh and 3.14 \$/MWh for New York ISO and the Australian Energy Market Operator, respectively [5]. To the best of our knowledge, the analysis in this paper is different from the literature in *all* these three distinctive aspects and it properly forecasts and classifies the electricity market prices at California Independent System Operator (CAISO). Accordingly, the contributions in this paper are:

- **New and Extensive Features**: We engineered several features to best capture the target variable characteristics, i.e., the day-ahead market price values. These features include the historical market prices at different nodes, the features related to year and month, the net demand, as well as the ancillary service requirements such as reserve, regulation and regulation mileage. The last two features are based on a relatively new market design platform by CAISO. Some features such as the prices at other locations and new ancillary service requirements, like regulation mileage, are entirely new in this paper and some other features previously used in other papers. Moreover, the combined set of the features we made in this paper is new and not been studied together before. For example, no ancillary service requirements were taken into account in [4]–[6]. The historical electricity prices and demand differences were the important features in [7]. Also, reference [1] only considered the historical electricity prices as the features.

This work is supported in part by National Science Foundation grant ECCS 1307756. The corresponding author is H. Mohsenian-Rad.

- **Effective Ensemble Learning Method**: The analysis in this paper is based on the ensemble learning method, namely the random forest (RF) method, which is known as one of the most powerful learning methods in data mining [11]. RF creates multiple decision trees in training phase and then aggregates the results and returns the solution in test phase. It is good in dealing with large datasets and handling over-fitting issue. It is also an overall robust approach. We experimentally adjusted the parameters for RF to increase the accuracy in our forecasting problem. Most importantly, RF is capable of taking advantages of our significantly diverse set of features; see the previous bullet point; compared to the existing literature.
- **Insightful Case Studies**: Our dataset comes from CAISO and our focus is on forecasting the day ahead market prices of the CAISO for the Pacific Gas and Electric (PG&E)'s default load aggregation point (DLAP).
  Our paper addressed predicting both the exact value and class label. We showed that our features and RF method could provide a promising results in both the exact value and class labels predictions. This is in contrast with [1], where the exact value prediction was not useful for class labeling. Our results also outperform in terms of MAE, mean absolute percentage error (MAPE) and mean percentage classification error (MPCE) in comparison with [1] and [5]. For Example, MPCE is reported as 9.21% when the approach in [1] is applied to our dataset compared to 6.53% for our approach. MAPE is also reported as 13.12% by applying the method in [5], while our method returns 2.13 $/MWh and 5.96% in terms of MAE and MAPE. Importantly, we also show *how* different features contributed to enhance forecast accuracy. Interestingly, it turns out that net demands, DLAPs at other locations and ancillary service requirements are effective features in our case study.

Finally, note that according to the literature survey reported in [4], most studies assess their models over four representative weeks across a year. Here, we examined our method on every day of 12 consecutive months. In such a case, MAPE may become larger and thus it is modified and smoothed to limit the effect of null and abnormal values as mentioned in [4]. However, in this study the evaluation framework involves the conventional definition of the MAPE.

II. California ISO Electricity Market

The CAISO market is multi-settlements consisting of two interrelated markets: day ahead market and real time market. Day ahead market is a forward market where generators and load commitments are determined for every hour of the next operating day, while the real time market is a spot market, where energy can be purchased at spot prices for each 15 minutes. Since the generation and load capacities traded at day ahead market are much more compared to the ones traded in real time market, the day ahead market has higher priority from the economic perspective such as price forecast studies.

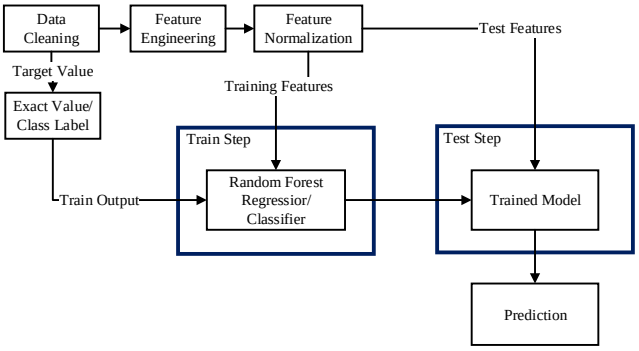

Fig. 1. The Main Components in Our Proposed Model

CAISO solves security constraint unit commitment (SCUC) problem to clear the market. The objective of SCUC is to maximize the social welfare of the system, and also to minimize ancillary service procurement cost subject to all grid constraints. In fact, the SCUC adjusts generation, load, import and export schedules based on the submitted energy supply and demand bids as well as ancillary service bids to meet ancillary service requirements, while managing grid congestion by enforcing transmission lines and generation units constraints [12]. The ancillary service requirements include: non-spinning/spinning reserve, regulation up/down, and mileage up/down.

Due to the complexity of locational marginal prices at different buses, CAISO is using the aggregation points to define the average prices of major regions in its territory. Those regions are based on the utilities which are responsible to buy energy on behalf of their consumers from CAISO market. One of the main utilities is PG&E which is in charge of the loads located in Northern California. Accordingly, PG&Es DLAP is the price average of purchasing energy in Northern California [12].

III. The Proposed Model

A. Methodology

In this paper, we gathered, and cleaned CAISO public available data [13] including the electricity market prices, net demand and ancillary service requirements. Then we engineered a reasonable set of features that could likely capture the target variable characteristics, i.e., the characteristics of the price values of the next operating day. Then we normalized our features. Then we split the data into two distinct parts named train set and test set. We studied the problem from both regression and classification perspectives.

In regression view, a random forest regressor was applied on the training set and then the performance of the model was evaluated on the test dataset using MAE and MAPE. In classification perspective, we used two different approaches. In the first one, we trained an RF regressor and predict the price values on the test set. Then, we compared values with the class labels using pre-specified threshold. In the second one, we converted all price features to pre-specified class labels and created an RF classifier to obtain the class labels directly. The performance of both approaches were evaluated by MPCE. Fig. 1 represents the main components of our proposed model.

TABLE I
FEATURE DESCRIPTION IN OUR PROPOSED MODEL

Features ID	Components	Feature Type
VeryShortTerm	Last 24 hours prices	Numeric
ShortTerm	Previous 25 to 48 hours prices	Numeric
LongTerm	Last week the same hour Last year the same hour	Numeric
Temporal	Weekdays Year Month	Binary Numeric Numeric
Geographical	The price at other DLAPs at 24, 25 and 26 hours ago	Numeric
NetworkCondition	Net Demand Ancillary Service Requirements	Numeric

B. Feature Engineering

We categorized our features into 6 groups, as shown in Table I. Next, we explain the features of each group.

Very Short Term and Short Term Features. The electricity market prices are highly autocorrelated. That is, the historical electricity market prices at previous hours provide effective features for predicting the future prices. As the focus of this paper is to predict and classify the prices for the day ahead market, 24 consecutive prices corresponding to 24 consecutive hours of the day should be predicted. As for predicting the price at hour H of the day of interest, one can ignore the effect of previous predicted prices at time slots $1, \ldots, H-1$. However, it is reasonable to take them into account [1]. In such a case, the predicted prices at previous hours are given as features. In this paper, the prices for the last 48 hours are considered. For example, if we aim to predict the electricity price at time slot 10:00 a.m., Dec. 8, 2015, the last 48 hours prices starting from 10:00 a.m., Dec. 6, 2015 to 9:00 a.m., Dec. 8, 2015 are taken as features. Needless to say that the prices starting from 1:00 a.m to 9:00 a.m., Dec. 8, 2015 are the predicted prices determined ahead by our proposed model.

Long Term Features. In price forecasting, it is important to take into account the long term trends in addition to the short term trends. We engineered relevant features based on historical data which lasts for longer period. The features such as last year on the same day at the same hour data, last week on the same day at the same hour prices are used in this paper.

Temporal Features. The day at which the price is determined is also of importance. For example, one can expect the lower prices at the weekends compared to a typical business day. Thus, we introduced 7 binary features, each corresponding to one day of the week, to capture this characteristic. We also consider the year and months as numeric features into account.

Geographical and Network Condition Features. The mechanism of market clearing process is another important factor to determine the market prices. The CAISO market is cleared by a co-optimization problem taking the ancillary services and network constraints into account. Therefore, we engineered three different features in our proposed model. These features are the historical price data for other important locations in the CAISO, net demand forecast and ancillary service requirements.

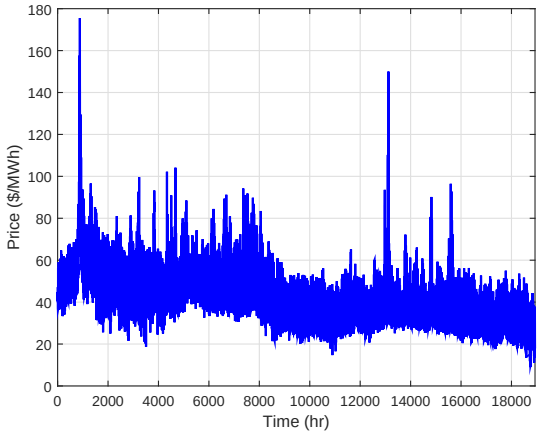

Fig. 2. The Price Data From Jan. 2014 to Feb. 2016.

C. The Learning Method

Ensemble methods learn several base estimators using a learning algorithm and combine their predictions. This approach can improve robustness over a single estimator. According to literatures, experimental evidences have shown that ensemble methods are often much more accurate than any single constituent model [11].

RF is one of the ensemble methods where each estimator is a decision tree. In this paper, we applied random forest with 150 estimators. To learn an estimator, a subsample of the training set is created using bootstrapping. The subsample size is equal to the training set size. For a given decision tree, the quality of each split is measured by Gini Impurity.

IV. CASE STUDY

A. Dataset Description

Our dataset consists of electricity market data obtained from California ISO Open Access Same-time Information System (OASIS). The data includes electricity market prices, net demand forecasts, ancillary service requirements during Jan. 1, 2014 to Feb. 28, 2016 [13]. The price trend in our dataset is presented in Fig. 2. The overall trend is that the day ahead electricity market price had been decreased during these two years. The price distribution is also presented in Fig. 3. The probability for each bin is equal to the number of the prices within the bin to the total number of prices. It is seen that the prices fall between 20 $/MWh and 60 $/MWh with more than 80% probability. However, there is still 20% chance that the price takes the higher or lower values. We created and labeled more than 80 features as shown in Table I. After cleaning the data, we found 18895 instances in total.

As for the classification problem, the class labels are named as low, medium and high. We first define two prices as 38 and 56 $/MWh. The price of 38 $/MWh is obtained as the mean of all the prices in the dataset. 56 $/MWh is at the boarder of the 10% highest prices in our dataset. We named the prices below 38 $/MWh as “low”, the price between 38 and 56 $/MWh as “medium” and the price above 56 $/MWh as “high” price.

As for predicting the exact value, we used two measures, namely MAE and MAPE [5]. As for the classification error,

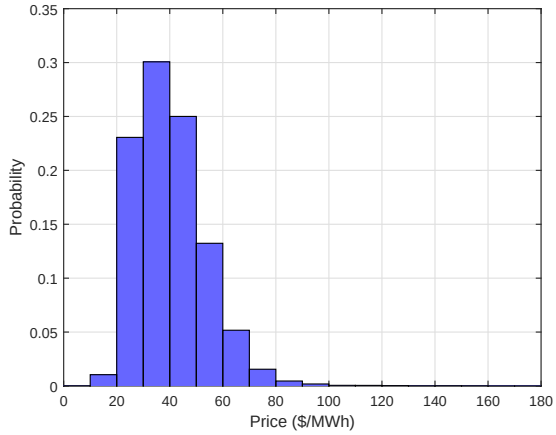

Fig. 3. The Distribution of Price Data From Jan. 2014 to Feb. 2016.

MPCE is a common evaluation method [1]. MPCE is defined as the total number of misclassification to the total number of instances in the test data. All the simulations were implemented on Python 2.7 using scikit-learn [14]. The computational time was 57 seconds in average for one day including train and test experiments.

B. Prediction Results: Exact Values and Class Labels

In this section, the exact values and class labels of the predicted prices for 12 months are obtained as shown in Table II. We foretasted the prices for each hour of a day starting from Mar. 2015 to Feb. 2016. For each specific day, the training data includes all the days starting from Jan. 2014 to the day before it. The average MAE and MAPE for each month are shown in Table II. Note that, the average MAE and MAPE errors for each month is calculated as the average of all MAE and MAPE errors over all days of the month. In average, the proposed model reported the 2.13 $/MWh and 5.96% for MAE and MAPE, respectively. According to Table II, the largest MAPE error occurs at Feb. 2016. The reason may be related to not considering the fuel price and temperature features.

We also determined the class labels of the prices. The MPCE error in average for all test data is 6.98% showing a bit drop compared to the exact value prediction. We also considered the method where the class labels are directly obtained. The average MPCE is 6.53% which is a promising result. Unlike the reference [1] that the authors observed a huge difference between these two approaches, our proposed method is showing a small difference. Note that, considering a deterministic threshold could have a negative impact on predicting the class labels. For example, the prices 37.5 $/MWh and 38.5 $/MWh are considered as a low and a medium price, even though it may not make a noticeable difference to a market participant. Addressing this issue could be a future research direction.

C. Comparison with Literature: Exact Values and Class Labels

We used KNN and NB methods for the classification. We combined KNN and NB with mutual information feature selection methods. These two methods worked well in [1]. Table III shows the results for the classification. As it can be seen, our proposed method outperforms these two methods. We also

TABLE II
THE AVERAGE ERROR RESULTS FOR OUR PROPOSED METHOD DURING MAR. 2015 TO FEB. 2016

Month	Exact Value Method			Direct Classification Method
	Exact Value Prediction		Class Label Prediction	
	MAE ($/MWh)	MPAE (%)	MPCE (%)	MPCE (%)
Mar.	1.78	5.86	4.76	6.25
Apr.	2.24	6.68	8.33	8.06
May	2.12	6.36	9.14	7.39
June	2.75	6.17	9.17	9.17
Jul.	2.39	5.23	9.01	8.06
Aug.	2.16	5.66	11.02	8.47
Sep.	2.17	5.36	9.03	8.47
Oct.	1.77	4.84	5.65	6.59
Nov.	1.8	6.02	5.32	4.6
Dec.	1.71	5.92	5.24	3.9
Jan.	1.64	5.69	4.30	4.44
Feb.	1.75	7.77	2.83	2.98
Mean	2.03	5.96	6.98	6.53
Std.	0.337	0.758	2.573	2.081

TABLE III
THE AVERAGE ERROR RESULTS FOR THE METHODS IN THE LITERATURE DURING MAR. 2015 TO FEB. 2016.

Month	Method	
	Naive Bayes	KNN
	MPCE (%)	MPCE (%)
Mar.	7.74	9.67
Apr.	12.08	12.92
May	10.35	13.84
June	13.47	15.56
Jul.	10.89	17.74
Aug.	9.68	13.98
Sep.	12.22	14.72
Oct.	9.81	10.89
Nov.	6.9	7.61
Dec.	6.45	8.47
Jan.	6.32	6.85
Feb.	4.61	4.02
Mean	9.21	11.36
Std.	2.777	4.098

compared our regression method with the one similar to the neural network model in [5]. The average MAPE was reported as 13.12%. This shows that the data mining approaches that are used to forecast the prices in other markets may not work well for CAISO; which further justifies the need for new studies based on the specific features of each market, such as our work in this paper.

D. More Detailed Discussions: The Exact Value Prediction

We considered Jul. 2015 and analyzed its actual and predicted prices as shown in Fig. 4. The average price for each hour and its deviation are shown in this figure. The prices at 1:00 a.m. to 8:00 a.m. did not show a high variation, thus make it easy for our model to estimate the exact value. However, the price deviation around 15:00 to 21:00 is high. For example consider the price at hour 18:00 where the actual price deviation is the highest. The price varies from 30 $/MWh to above 70 $/MWh at that hour. However, our proposed method could not adopt itself with these deviations and it only deviates from 42 $/MWh to 62 $/MWh. It means that there may be other

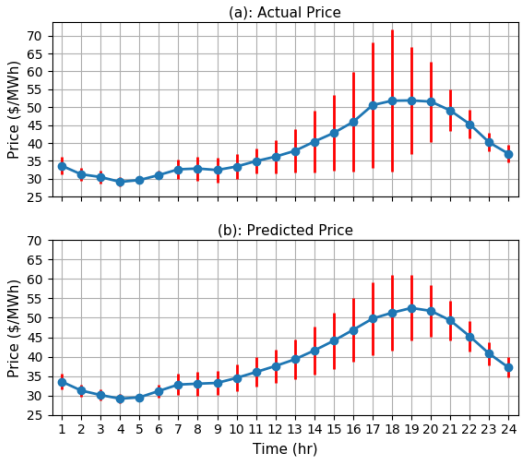

Fig. 4. The actual and predicted price variation in Jul. 2015.

TABLE IV
THE CONTRIBUTION OF THE PROPOSED FEATURES ON THE ERROR DECREASE FOR MAY 2015

Features	Error				
	Avg MAE ($/MWh)	MAPE (%)			
		Avg	Max.	Med.	Min.
VeryShrotTerm	2.7	8.07	20.79	6.4	2.87
+ ShorTTerm	2.69	8.13	18.97	6.81	2.88
+ Temporal	2.61	7.86	18.04	6.66	2.45
+ Geographic	2.53	7.71	17.78	6.56	2.1
All Features	2.12	6.36	11.33	5.85	2.53

features that we did not consider, but they may improve the performance of our model. It would be interesting to obtain high-variation hours and add more corrections to the predictor model to obtain the higher accuracy. Another important issue is to analyze how our proposed features contribute to improve the performance accuracy in our model. As the day ahead prices are highly autocorrelated to the previous days, we first consider those features as the only features and obtain MAE and MPAE for the day ahead prices in May 2015. According to Table IV, considering only the previous day prices have a great impact on the price accuracy. This is not a surprising result as it is shown in the previous literature. We then add our proposed features in Table I and see how the errors would change. It is interesting to see that adding our proposed features in our RF model decreases the average MAE and MAPE errors. Another interesting observation is about the maximum MAPE error. The maximum MAPE error is the highest error between all the days in May 2015. As we consider the features related to the network conditions, i.e., the prices at different location for previous day, net demand forecasts and ancillary service requirements, the error decreases significantly by 9.46 % compared to only considering the last 24 hours prices. Accordingly, we can conclude that those features can help the model to evaluate more accurately the conditions that cause the unexpected prices.

V. CONCLUSION

This work applied an ensemble learning model named RF to predict the exact value and to classify the prices in the CAISO day ahead market. Several features such as historical prices in the location of under study, i.e., PG&E DLAP and other locations, net demand, calender and ancillary service requirements such as the new product named mileage requirements have been engineered. The model was implemented on the CAISO market from Jan. 2014 to Feb. 2016. The result was promising. The average MAE and MAPE were 2.13 $/MWh 5.96% during one year test data. The average MPCE result for classifying was also reported as 6.53%. The model was compared with the literature and outperformed them. It was observed that introducing our features helped the model to reduce the maximum MAPE by 9.46% in average compared to considering only the historical prices. The future research would be on introducing the new efficient features to capture the price characteristics in the hour with high price fluctuations more appropriately. It is also interesting to Consider the impact of other features, such as temperature and fuel prices, on exact value prediction as well as on the classification. Last but not least, the comparison with more papers is a necessity to evaluate better the effectiveness of the proposed model.

REFERENCES

[1] D. Huang, H. Zareipour, W. D. Rosehart, and N. Amjady, "Data mining for electricity price classification and the application to demand-side management," *IEEE Trans. on Smart Grid*, vol. 3, no. 2, pp. 808–817, 2012.
[2] F. M. Alvarez, A. Troncoso, J. C. Riquelme, and J. S. A. Ruiz, "Energy time series forecasting based on pattern sequence similarity," *IEEE Trans. on Knowledge and Data Engineering*, vol. 23, no. 8, pp. 1230–1243, 2011.
[3] H. Mori and A. Awata, "Data mining of electricity price forecasting with regression tree and normalized radial basis function network," in *Proc. of IEEE International Conference on Systems, Man and Cybernetics*, 2007, pp. 3743–3748.
[4] Panapakidis, Ioannis P and Dagoumas, Athanasios S, "Day-ahead electricity price forecasting via the application of artificial neural network based models," *Applied Energy*, vol. 172, pp. 132–151, 2016.
[5] B. Neupane, K. S. Perera, Z. Aung, and W. L. Woon, "Artificial neural network-based electricity price forecasting for smart grid deployment," in *Proc. of Computer Systems and Industrial Informatics (ICCSII)*, 2012.
[6] González, Camino and Mira-McWilliams, José and Juárez, Isabel, "Important variable assessment and electricity price forecasting based on regression tree models: classification and regression trees, Bagging and Random Forests," *IET Generation, Transmission & Distribution*, vol. 9, no. 11, pp. 1120–1128.
[7] Barta, Gergo and Nagy, Gyula Borbely Gabor and Kazi, Sandor and Henk, Tamas, "GEFCOM 2014probabilistic electricity price forecasting," in *Intelligent Decision Technologies*, 2015, pp. 67–76.
[8] R. Weron, "Electricity price forecasting: A review of the state-of-the-art with a look into the future," *International Journal of Forecasting*, vol. 30, no. 4, pp. 1030–1081, 2014.
[9] Zareipour, Hamidreza and Janjani, Arya and Leung, Henry and Motamedi, Amir and Schellenberg, Antony, "Classification of future electricity market prices," *IEEE Trans. on Power Systems*, vol. 26, no. 1, pp. 165–173, 2011.
[10] J. Reston Filho, C. Affonso, and R. de Oliveira, "Energy price classification in north brazilian market using decision tree," in *Proc. of International Conference on the European Energy Market (EEM)*, 2015.
[11] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.
[12] https://www.caiso.com/rules/Pages/BusinessPracticeManuals/Default.aspx.
[13] http://oasis.caiso.com/mrioasis/logon.do.
[14] Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.