Predicting and Explaining Price-Spikes in Real-Time Electricity Markets

Christian Brown[#1], Gregory Von Wald[#2]

[#]*Energy Resources Engineering Department, Stanford University*
367 Panama St, Stanford, CA 95304

[1]csbrown@stanford.edu

[2]gvonwald@stanford.edu

***Abstract*—The electricity market is designed to optimize the generation and delivery of power. When the grid is under stress, price spikes may occur, yielding up to a 100-fold increase in the electricity price. In order to predict the likelihood of a real-time price spike occurrence, a suite of supervised classification algorithms were trained based on the weather, day-ahead market information, and temporal characteristics. Specifically, a logistic regression, a random forest classifier, and a gradient boosting classifier were trained and tested on New England ISO electricity prices. The Gradient Boosted classifier achieved the highest accuracy of 95.1%, with a positive recall of 97.2%. As electricity markets continue to grow more volatile, tools for the prediction of electricity price spike occurrence will add value to the grid operator's ability to mitigate price-spikes as well as energy trader's ability to hedge their bets and avoid hours of high risk.**

I. Introduction

The wholesale electricity market is designed to ensure optimal generation and delivery of power. When markets are operating properly, the total system cost is minimized. However, when the grid is under stress, price-spikes may occur in the real-time market. These anomalies can yield up to a 100-fold increase in the price of electricity. This poses a market inefficiency during which the grid is operating at a suboptimal level. Price-spikes are typically very difficult to predict as by definition they only occur when the grid-operator could not predict the anomaly. We hope to utilize machine learning techniques to improve the accuracy of predicting these hours of volatility and explaining their underlying causes such that the grid operator may make informed decisions to mitigate or avoid these situations.

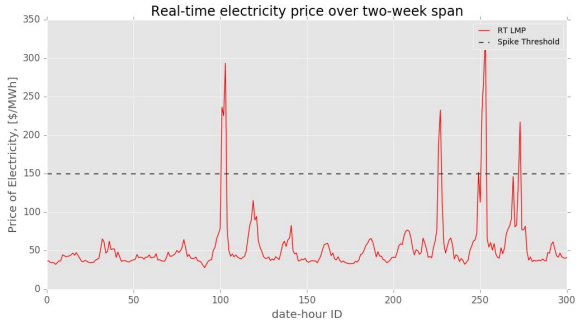

Figure 1. Hourly real-time electricity price from 2011-07-27 to 2011-08-09 with numerous price spikes

The input to our algorithm is a set of weather, day-ahead market, and temporal features of a given hour. The output of our model is the predicted probability that this hour will contain a price spike. In this manner, the objective is to train a binary classification algorithm with a set of marked examples in order that these price-spikes can be predicted ahead of time and grid operators as well as energy traders can act with caution during these hours of volatility.

Section II will provide a high-level overview of some of the previous literature on predicting price volatility with machine learning. Section III will discuss the datasets obtained and the features selected for this analysis. Section IV will describe the learning algorithms that were utilized. Section V outlines the experiments conducted, the final results obtained, as well as a discussion of these results in the context of the market. Finally, Section VI will offer conclusions and future directions for this work.

II. Related Work

Forecasting the price of electricity is a well-covered topic in the literature as this is essential for wholesale power providers as they determine their bidding strategy into the market and wholesale power purchasers as they determine their procurement strategy. A large body of research exists for the forecasting of both electricity demand and electricity prices [1-7].

However, price-spikes fall outside of the typical profile and these methods are typically ineffective at predicting these volatile spikes in the electricity market. As such, several machine learning based approaches have sought to improve forecasting of price spikes[8-13].

Lu et al. developed a data mining based electricity price forecast framework leveraging a neural network based forecast model for the normal prices, coupled with a data mining approach to predict price spikes. Using only market

information, and ignoring weather data, this group was able to produce a Bayesian classifier that could predict spikes based solely on the demand-supply balance in the market[8].

Amjady et al. propose a probabilistic neural network for the forecast of price spike occurrence[9]. Mount et al. utilize a regime-switching model to represent the volatile behavior of wholesale electricity prices[10]. Huang et al. display the applications that these methods could have for demand-side management, where operational decisions are made based on whether the price for electricity will fall above a certain threshold[11]. For a more detailed discussion of the state-of-the-art in electricity price forecasting, Weron has written a detailed review article[14].
However, one major shortcoming of many of these papers is that the demand-supply balance data and reserve margin data are not available ahead of time. Many of these methods require perfect information and can only be helpful in hindsight. Additionally, many of these researchers did not leverage weather data in their analysis.

It is with these shortcomings in mind that we propose a binary classifier to predict the likelihood of price spike occurrence based on day-ahead market information, temporal characteristics, and weather data.

III. Dataset and Features

The primary data used for the model was grid and power market data from ISO New England (ISO-NE) and New England Local Climatological Data (LCD) from the National Oceanic and Atmospheric Administration (NOAA). Preprocessing was required to combine data from these two sources, remove measurements that were incompatible with our models, and to convert certain features into variables that could be used by our model. Additionally, the training set needed to be balanced prior to training to predict optimally.

The data from ISO-NE were zonal time-series data discretized by date and hour. For each date and hour, a datapoint was reported for one of the eight ISO-NE zones (figure 3). Some features from this dataset included day-ahead (forecasted) energy demand and electricity price, along with the real-time electricity price. Based on the real-time price of a particular date, hour and zone, the observation was assigned a binary tag indicating whether or not it would constitute a spike. For the majority of the analysis in this paper, a spike-threshold of $150/MWh was used. This spike tag was used as the target for the binary classifiers used in this project.

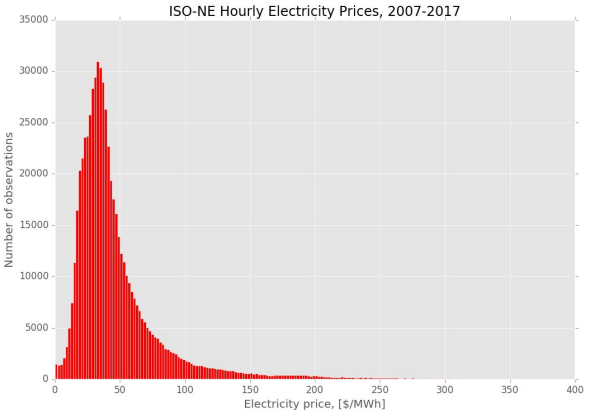

Figure 2. Distribution of electricity prices in NE-ISO service territory for the years 2007-2017.

NOAA provides timestamped weather and climate measurements from the hundreds of weather stations that it manages across the country. Eight weather stations in the New England region were assigned to each of the eight New England zones, to best capture the relationship between weather and grid behavior. Weather station measurements were averaged over each date and hour so that it would map directly to a particular date-hour measurement given in the ISO-NE data. Once the weather data was joined to the ISO-NE data using the date-hour-zone mapping, observations with incomplete measurements were dropped.

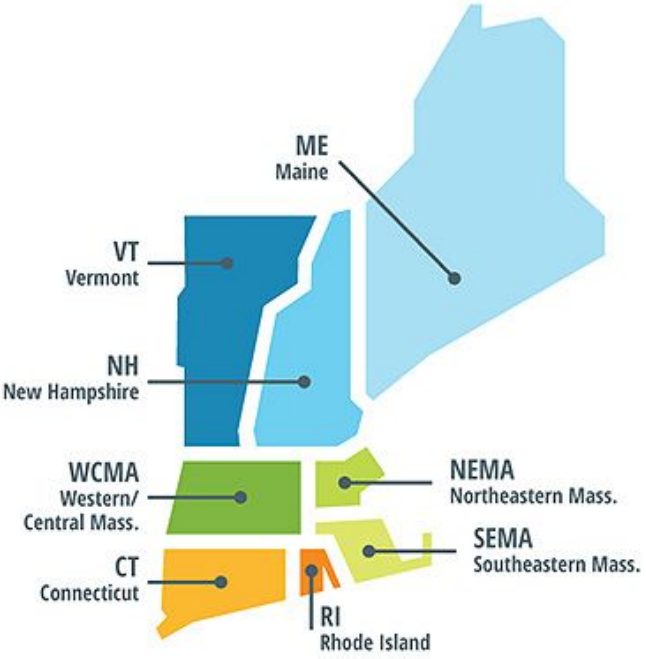

Figure 3. Zones of ISO-NE used in analysis.

In addition to grid and weather data, calendar features were introduced to capture the relationship between price and time. Some of these features included day of the week, weekend or

weekday indicators, US holiday indicators and whether the hour constitutes on-peak or off-peak hours, as defined by ISO-NE. Additionally, features such as hour, month, and day of week were converted from single integer variables to a pair of cyclical features, to capture the cyclic nature of time.

Ten years of hourly data was extracted for each ISO-NE zone and corresponding weather station. After preprocessing, the overall dataset included approximately 611,000 observations. As real-time price is not distributed uniformly, the dataset is imbalanced with respect to the two classes of observation the model is trying to predict (spike and non-spike). Using the $150/MWh threshold for classifying an observation as a spike, approximately 2.8% of observations were classified as a spike. The overall dataset was initially spilt into a train-validation set and test set using an 80-20 ratio. Following this, the train-validation set was split into a train set and validation set using an 80-20 ratio. For training, the dataset was treated into a dataset balanced 50-50 by classification via first extracting all samples labeled as a spike, and sampling a corresponding number of non-spikes from the remainder of the training set. Using a spike threshold of $150/MWh yielded a balanced training set size of about 21,000 observations. Classifiers were trained on the balanced training set, then were validated and eventually tested using the remaining imbalanced data. Due to computational constraints, the validation set and test set were limited to a size of 30,000 observations, and were obtained by sampling from their corresponding sets.

IV. Methods

Three binary classifiers were used in this analysis. They were logistic regression, random forest classifiers and gradient boosted classification trees. The data was managed with Python's pandas library, and the models were implemented using Python's scikit-learn library.

Logistic regression is a regression technique used for categorical variables. In particular, the sigmoid function is used as the hypothesis in order to output a value between 0 and 1. This output, which can be interpreted as a probability, can be used to assign a classification to a particular sample. The objective of the model is to maximize the log likelihood function.

$$max \sum_{i=1}^{m} \hat{y}^i log y^i + (1 - y^i) \log(1 - \hat{y}^i) - \lambda \|\theta\|^2{}_2$$

Additionally, a regularization term is included to reduce the variance of the model. This is achieved by limiting the magnitude of the model parameters. The strength of regularization can be tuned by adjusting the regularization coefficient.

The random forest classifier is an ensemble learning method where the mode of the output of a number of decision trees is used as the output. One representation of the objective of a random forest classifier is as follows:

$$min \sum_{i=1}^{m} l(y^i, \hat{y}^i) + \sum_{k=1}^{K} \Omega(f_k)$$

$$with\ \hat{y}^i = \frac{1}{K} \sum_{k=1}^{K} f_k(x^i)$$

$$and\ \Omega(f_k)\ is\ complexity\ of\ tree\ k$$

The complexity of the tree can be defined in a variety of different ways, and is generally used to specify the parameters of each tree. This analysis used scikit-learn's implementation of the random forest classifier, where some of the tree parameters that can be altered include the number of features taken into consideration when growing an individual tree and the number of samples required for a node to be considered an external node. As such, the complexity of this analysis's model was determined by scikit-learn's definition of an individual decision tree.

In addition to the tree parameters, random forest parameters such as the number of decision trees to train were specified. Although a greater number of trees will produce more robust results, the relative performance gains achieved by increasing the number of trees in the forest is eventually outweighed by the increase in the computation time of the model.

The gradient boosted classification tree is another ensembling method that utilizes decision trees. However, unlike the random forest classifier which grows all trees in parallel, decision trees for the gradient boosted classification tree are grown in sequence. One representation of the minimization objective function of a gradient boosted tree model is shown as follows:

$$\sum_{i=1}^{m} \left[g_i f_t(x_i) + \frac{1}{2} h_i {f_t}^2(x_i) \right] + \Omega(f_t)$$

$$with\ g_i = \partial_{\hat{y}^{t+1}} l(y_i, {\hat{y}_i}^{t-1}), h_i = \partial^2_{\hat{y}^{t+1}} l(y_i, {\hat{y}_i}^{t-1})$$

The number of trees to be grown in sequence is specified beforehand. Trees are optimized in sequence, taking into account the previous tree's output. The final tree's output is used as the final prediction. As such, many of the tree parameters specified for a random forest model are also used in the gradient boosted classification tree model.

Additionally, model parameters such as learning rate, are defined to specify the influence that each tree has on the tree that succeeds it. Due to this form of regularization, one can interpret the gradient boosted classification tree as a decision tree based model that has lower variance than that of a random forest classifier.

The three models mentioned above are fitted on the balanced training set described in the previous section, and assessed on

an imbalanced validation set. During this stage, each model's hyperparameters (coefficient of regularization, number of trees, learning rate, etc.) were selected via grid-search and manual tuning to obtain the best overall performance.

The performance of the model was evaluated by taking the average of a model's overall accuracy, true-recall rate, and precision.

$$Model\ Score = mean(accuracy, true\ recall\ rate, precision)$$

This scoring criteria was used to capture a model's ability to successfully predict a price-spike should it occur, while also valuing overall model accuracy and reliability. For instance, evaluating model solely based on overall accuracy may bias the hyperparameters towards models that tend to over-predict non-spikes, since there are far more observations labeled as non-spikes compared to spikes.

Once optimal hyperparameters for each model was selected, each model was retrained on a balanced subset of both the train and validation set. Following this, the models were analyzed based on their performance on the same subset of the test set.

V. Experiments/Results/Discussion

In order to achieve the best classification performance, several metrics were utilized to understand the accuracy of each classifier. Overall accuracy was used as a component of scoring each model performance. However, in the context of predicting price spikes, the positive recall is the most important metric as it provides the fraction of spikes that you are able to predict accurately. Since this is intended as a risk mitigation tool,the main objective is ensuring that you are not caught off guard by a price spike. Finally, precision was included as a component in our model scoring metric as we also do not want the model to be too conservative and we sought a measure of confidence that when our model predicts a spike, there will actually be a spike.

As mentioned above, the composite "score" for model performance that we used was the arithmetic mean of the three aforementioned metrics. This score was then used to tune the hyperparameters of each model.

A. Hyperparameter Tuning

As discussed in the Methods section, the hyperparameters for each model were selected via grid-search and manual tuning in order to achieve the best performance, as measured by the composite score developed given a spike threshold of $150/MWh.

For logistic regression, the only hyperparameter that needed to be selected was the regularization coefficient. An inverse regularization coefficient of 6e-5 proves to be optimal.

For the random forest classifier, the parameters tuned included number of trees in the forest, minimum number of samples to be considered an end node and number of features used in each decision tree. The optimal hyperparameters, such as the number of samples to consider when defining an end node, indicated that the model was optimal when it had a relatively high variance.

For the gradient boosting classifier, the aforementioned tree parameters were tuned, in addition to the learning rate and maximum tree depth. Similar to the random forest classifier, the optimal hyperparameters suggested that the model put an emphasis on variance.

	Inverse Regularization Strength				
Logistic Regression	6e-5				
	Number of Trees	Minimum Samples leaf	Minimum samples split	Max number features	
Random Forest Classifier	70	1	2	Sqrt(n)	
	Number of Trees	Minimum Samples leaf	Maximum Tree Depth	Max number features	Learning Rate
Gradient Boosted Classifier	125	1	10	Sqrt(n)	.25

Table 1. Tuned hyperparameter selections for the models when subject to a $150/MWh spike threshold

B. Model Performance

The logistic regression model performed the worst out of the three, while the random forest classifier and the gradient boosted classifier yielded better results. While all the overall accuracy metrics lie within a few percent of each other, the tree-based methods produced significantly better positive recalls.

Model	Training (m=26,640)	Test (m=30,000)
	Accuracy	Accuracy *(recall/score)*
Logistic Regression	88.90%	93.2% *(88.2%/.711)*
Random Forest Classifier	99.99%	94.46% *(97.0%/.764)*
Gradient Boosted Classifier	100%	95.1% *(97.2%/.777)*

Table 2. Model performance from training to test set

V.B.1 Threshold Modification

One assumption that we had made throughout our analysis was that the price threshold that indicated a price spike was $150/MWh. For this experiment, we modified the threshold and measured the positive recall of each of the models. As

displayed below in Figure 4, the model performance declines as the spike threshold increases. This is largely attributable to the fact that there are fewer spikes from which to train on, and therefore, even though our training set is balanced, the classifier is less accurate at properly classifying the spikes of greater magnitude.

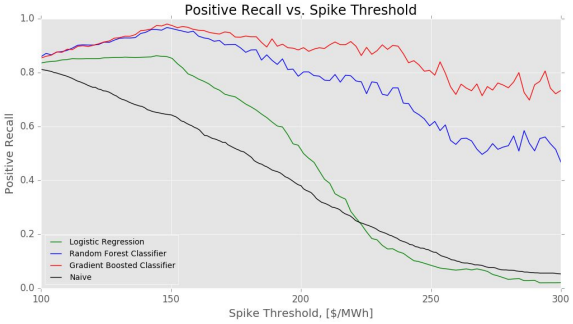

Figure 4. Different models' positive recall rate when subject to different price-spike thresholds

V.B.2 Comparison to Naive Model

As displayed above in Figure 4, we developed a Naive Model that simply looks at the Day Ahead market clearing price of electricity and if it is above the threshold, predicts a spike for this hour in the real-time market. The naive model's positive recall falls quickly to about 60% when the threshold for a price-spike is at $150/MWh. Our model is able to perform significantly better than naively relying solely on the day ahead price, correctly predicting about 50% more of the real-time market spikes.

C. Sample Model Output

A two-week span of the test set was classified and visualized below to display the gradient boosting classification model's ability to predict all of the spikes over this stretch. However, there are also some misclassifications seen as red dots below the threshold. This visually depicts our models high positive recall, but also lower precision as there are a fair amount of Type II errors present.

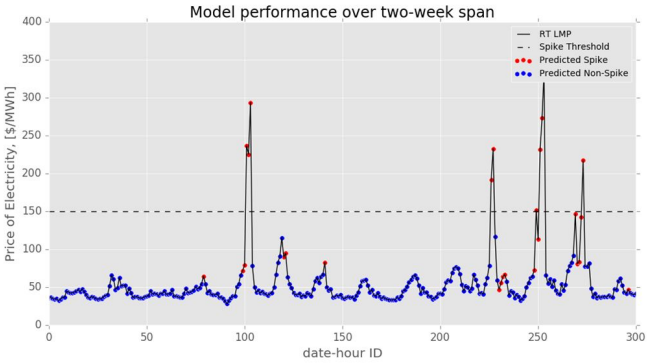

Figure 5. Time-series of hourly electricity prices, price-spike threshold, and model predictions for dates of 7/27/2011 to 8/9/2011

VI. Conclusions

The gradient boosted classifying tree model had the best overall performance compared to the random forest classifier and logistic regression model. The model's high positive recall rate of 97.2% on the test set showed that it could be used as a valuable tool to predict incoming price spikes. The model's main source of error was from Type II (false positive) errors. This was deemed acceptable, as the ability to predict price-spike events was prioritized above overall accuracy. Additionally, the spike-threshold sensitivity analysis showed that the gradient boosted classifying tree had the most robust predictions, retaining a positive recall rate of above 70% as the threshold approached $300/MWh.

Future work for the model includes more extensive model reduction to identify the features most valuable to the model, limiting the featureset to day-ahead predictions so that it can be used directly in practice as a 24 hour ahead hedging tool, and experimenting with polynomial expansions of features. Additionally, the definition of a price-spike event can be revised from a fixed threshold to a function of the previous hour's spikes. This will help capture sudden increases in price relative to surrounding hours, as opposed to periods of consistently high prices that would be misinterpreted as spikes using the current framework.

This model can be used in tandem with other predictive models, such as a regression model that predicts the actual value of real-time electricity price, in order to improve overall accuracy.

Contributions & Acknowledgements

Christian contributed: Preprocessed dataset, developed infrastructure to train, test, and evaluate different models using Python and Pandas/Scikit-learn libraries.

Greg contributed: Literature review, raw data acquisition, assisted with debugging code, hyperparameter selection and evaluation of models.

We would also like to acknowledge Alphataraxia for the inspiration to pursue this topic and for providing a high-level overview of the scope.

References

[1] Raviv, E., Bouwman, K. E., & van Dijk, D. (2015). Forecasting day-ahead electricity prices: Utilizing hourly prices. *Energy Economics*, *50*, 227-239.

[2] Amjady, N. (2006). Day-ahead price forecasting of electricity markets by a new fuzzy neural network. *IEEE Transactions on power systems*, *21*(2), 887-896.

[3] Amjady, N., & Hemmati, M. (2006). Energy price forecasting-problems and proposals for such predictions. *IEEE Power and Energy Magazine*, *4*(2), 20-29.

[4] Amjady, N., & Keynia, F. (2011). A new prediction strategy for price spike forecasting of day-ahead electricity markets. *Applied Soft Computing*, *11*(6), 4246-4256.

[5] Singhal, D., & Swarup, K. S. (2011). Electricity price forecasting using artificial neural networks. *International Journal of Electrical Power & Energy Systems*, *33*(3), 550-555.

[6] Contreras, J., Espinola, R., Nogales, F. J., & Conejo, A. J. (2003). ARIMA models to predict next-day electricity prices. *IEEE transactions on power systems*, *18*(3), 1014-1020.

[7] Yamin, H. Y., Shahidehpour, S. M., & Li, Z. (2004). Adaptive short-term electricity price forecasting using artificial neural networks in the restructured power markets. *International journal of electrical power & energy systems*, *26*(8), 571-581.

[8] Lu, X., Dong, Z. Y., & Li, X. (2005). Electricity market price spike forecast with data mining techniques. *Electric power systems research*, *73*(1), 19-29.

[9] Amjady, N., & Keynia, F. (2010). Electricity market price spike analysis by a hybrid data model and feature selection technique. *Electric Power Systems Research*, *80*(3), 318-327.

[10] Mount, T. D., Ning, Y., & Cai, X. (2006). Predicting price spikes in electricity markets using a regime-switching model with time-varying parameters. *Energy Economics*, *28*(1), 62-80.

[11] Huang, D., Zareipour, H., Rosehart, W. D., & Amjady, N. (2012). Data mining for electricity price classification and the application to demand-side management. *IEEE Transactions on Smart Grid*, *3*(2), 808-817.

[12] Christensen, T. M., Hurn, A. S., & Lindsay, K. A. (2012). Forecasting spikes in electricity prices. *International Journal of Forecasting*, *28*(2), 400-411.

[13] Zhao, J. H., Dong, Z. Y., Li, X., & Wong, K. P. (2007). A framework for electricity price spike analysis with advanced data mining methods. *IEEE Transactions on Power Systems*, *22*(1), 376-385.

[14] Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International journal of forecasting*, *30*(4), 1030-1081.