International Conference on Applied Energy 2019
Aug 12-15, 2019, Västerås, Sweden
Paper ID: 777

DATA-DRIVEN ANALYSIS OF WEATHER IMPACT ON THE OUTAGES IN URBAN POWER DISTRIBUTION SYSTEM

Yang Zhang[1], Andrea Mazza[1*], Ettore Bompard[1], Emiliano Roggero[2], Giuliana Galofaro[2]

1 Department of Energy, Politecnico di Torino, Italy

2 IRETI SpA, Gruppo IREN, Italy

ABSTRACT

In this paper, a data-driven approach is explored to evaluate the impact of weather conditions on the reliability of urban distribution system. The severity of power outages is divided into two levels according to the number of days with outages in one week. The actual outage records from the local utility are used for the analysis in this study. First, the difference of weather conditions under the two outage levels are intuitively described with the Kernel Density Estimation (KDE). Then, an extreme gradient boosting algorithm is applied to build a classification model for evaluating the outage levels of the local distribution system under given weather conditions. The importance of weather features on the outage level is discussed with the built model. Finally, the performance of the proposed data-driven model is assessed with the Receiver Operating Characteristic (ROC) curve.

Keywords: kernel density estimation, distribution system, outage, ROC

NONMENCLATURE

Abbreviations	
KDE	Kernel Density Estimation
ROC	Receiver Operating Characteristic
PDF	Probability Density Function
TPR	True Positive Rate
FPR	False Positive Rate
Symbols	
MH	Weekly minimum relative humidity
MT_{15}	Average value of maximum temperature in 15 days
MH_{15}	Average value of maximum relative humidity in 15 days
$Prec_{15}$	Average value of precipitation in 15 days

1. INTRODUCTION

As the terminal of power grid, distribution network is characterized with various electrical equipment, multi-type feeders and wide area, which bring various causes for an outage, including aging of equipment, damages from adverse weather conditions, human's error, and so on. A reliable power supply is always the goal of distribution utilities' pursuit as it plays a significant role in the customers' satisfaction [1]. Therefore, an evaluation of outages in the system is necessary to the utilities as an important reference for predictive maintenance and secure operation.

In order to address the uncertainty of outages, a lot of researches have been published to identify the potential threats to the reliability of power distribution system. For example, a wind speed severity scale model is proposed in [2] to discuss the wind-related failures of overhead lines. The reliability evaluation of the system is improved by considering the stochastic characteristic of a wind speed time series. Moreover, the risk of wind storms on the distribution systems in the Northeast U.S. is analyzed in [3] based on the historical records of storm events. Apart from the wind characteristics, some other environmental variables are taken into consideration for a two-step prediction procedure introduced in [4], which is used to get a spatiotemporal forecast of outages in the electrical power system.

Selection and peer-review under responsibility of the scientific committee of the 11th Int. Conf. on Applied Energy (ICAE2019).

In this paper, the impact of weather conditions on the outages in distribution system will be discussed. The outage severity of one week is defined according to the number of days with outages, i.e. outage days. If there are more than two days appearing in the outage records within one week, the outage severity of this week is labeled as Level II, indicating a fault-prone period; and the other case is labeled as Level I. In such case, this task could be regarded as a binary classification problem with various weather conditions as input and two outage levels as output.

2. DATA DESCRIPTION

2.1 Outage records

In this study, the outages of the local distribution network are collected from 2014 to 2017. There are 52 complete weeks in each year taken for analysis, whose outage days are shown in Fig 1.

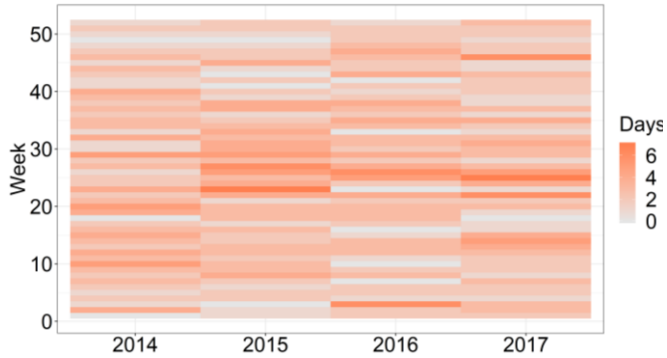


Fig 1 Number of days with outages in every week

As can be seen from the figure above, the weeks in the middle of the year contain more outage days, especially in 2015 and 2017 when the city suffered more severe heat waves than normally years. It indicates that the reliability of distribution network is under higher risk in summer and could be affected by certain weather conditions. For a better understanding of the weather's impact on the outages, a criterion is applied in this paper to identify the outage severity as previously described, which classifies the weeks with more than 2 outage days as Level II, and all the others as Level I.

2.2 Weather records

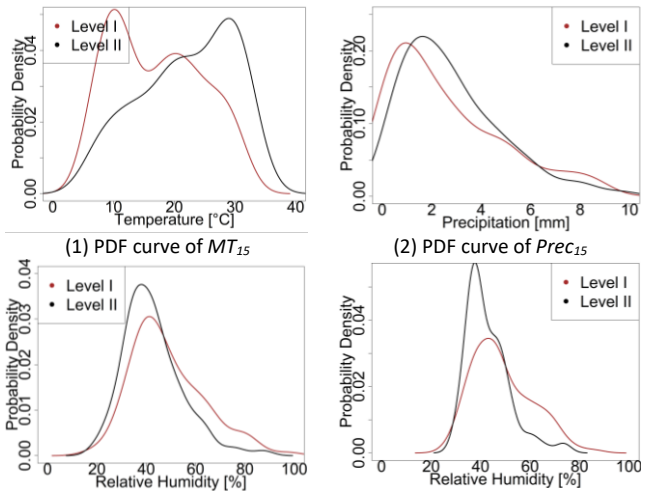
The weather information collected for our study includes the maximum temperature, minimum relative humidity, solar radiation, precipitation, average wind speed as well as duration without wind (i.e., calm duration) of each week. In order to take the continuous impact of weather conditions into consideration, the average values of the first four features within 15 days are further calculated to avoid the bias due to some abnormal observations. Moreover, the number of days without any precipitation in 31 days before an outage is counted as a new feature for analysis.

With the outage severity defined to each week, the weather conditions corresponding to different outage levels are supposed to be distinct to each other. In this paper, the PDF curve of each weather feature under different outage levels is to be estimated with the KDE method, which is a widely used non-parametric method without priori assumptions about the distribution. Given a random set of data points x_1, …, x_n from an unknown continuous distribution, the probability density could be estimated as equation (1)

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n}\phi\left(\frac{x-x_i}{h}\right) \quad (1)$$

where $\phi(\cdot)$ is a kernel function and *h* is the bandwidth. The kernel function is supposed to provide a higher weight to the data points near x when calculating the probability density at the point x.

With the kernel function chosen as Gaussian distribution function and the bandwidth as the standard deviation of the smoothing kernel, the PDF curves of some of the weather features are shown in Fig 2.



(1) PDF curve of MT_{15} (2) PDF curve of $Prec_{15}$

(3) PDF curve of *MH* (4) PDF curve of MH_{15}

Fig 2 PDF curves of four weather features

In Fig 2, there is a clear difference between the two distributions of weather records under different outage severity. In particular in the distribution of MT_{15}, the peak probability density of weeks labeled as Level II appears at a higher temperature, showing that it is one of the important threats of the reliability in distribution system. The similar situation is also shown in the PDF of $Prec_{15}$, where the weeks of severe outages concentrate more with a little higher precipitation. As in the probability density curves of *MH* and MH_{15}, the average value of the

daily minimum relative humidity in 15 days shows a higher difference than the week's minimum value of this feature. Therefore, it is reasonable to take the weather conditions in a longer time period for considering the weather's continuous impact.

3. CLASSIFICATION

As discussed above, the evaluation of weather's impact on the reliability of urban distribution system could be carried out as a binary classification. By dividing the outage severity into two levels, the relations between the given weather conditions and probability of outage severity is studied with the XGBoost algorithm in this paper.

3.1 Extreme gradient boosting (XGBoost)

Boosting is one of the ensemble techniques which aims to create a strong learner based on a sequence of weak learners whose performance is slightly better than random guessing [5]. The misclassified samples from earlier weak learners will be assigned with more weight in each iteration. At the end, all of the successive learners will vote for the classification problem, creating a strong learner. In XGBoost algorithm, the concept of Boosting is implemented with bifurcated decision tree as weak learners.

Given a dataset denoted as $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where n is the number of samples, $\mathbf{x}_i$ is a m-dimensional corresponding to the features of the i-th sample and y_i is the class label, then the generalized additive model [6] for the final strong leaner in XGBoost can be written as

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^{K} f_k(\mathbf{x}_i) \quad (2)$$

where $\hat{y}_i$ is the prediction of the final model, $f_k(\mathbf{x})$ is the model of k-th decision tree, and K is the total number of weak learners. In the algorithm, the k-th decision trees cannot affect the parameters in the previous ones. Therefore, the output of the first (k-1) trees is a constant value $\hat{y}_i^{(k-1)}$ in the calculation of k-th iteration. The objective function for the parameters of the k-th decision tree is as follows:

$$\mathcal{L}^{(k)} = \sum_{i=1}^{n} \ell\left(y_i, \left(\hat{y}_i^{(k-1)} + f_k(\mathbf{x}_i)\right)\right) + \Omega(f_k) \quad (3)$$

where ℓ is a differentiable convex loss function that measures the difference between output of the final model $\hat{y}_i$ and the real class label y_i. This means that the performance of prediction model is supposed to improve with new decision trees added. In order to control the complexity of the tree's structure and avoid the overfitting problem, a regularization term is added in the objective function, whose formula is in equation (4).

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\mathbf{w}\|^2 \quad (4)$$

where artificial coefficients γ and λ are set to adjust the impact of the number of leaves T and weights of these leaves $\mathbf{w}$ in the new decision tree.

In XGBoost algorithm, the second-order approximation of equation (3) is used as for the optimization of weights in the new added decision trees. With the constant term removed, the simplified objective function could be written as

$$\tilde{\mathcal{L}}^{(k)} = \sum_{i=1}^{n}\left(g_i f_k(\mathbf{x}_i) + \frac{1}{2}h_i f_k^2(\mathbf{x}_i)\right) + \Omega(f_k) \quad (5)$$

where g_i and h_i are the first and second order gradient statistics on the loss function. The above equation could be regarded as a quadratic function of the weights in the k-th decision tree. Finally, the optimized weights of leaves in the new decision tree could be derived as follows:

$$w_j^* = -\frac{\sum_{i\in\mathbf{I}_j} g_i}{\sum_{i\in\mathbf{I}_j} h_i + \lambda} \quad (6)$$

where $\mathbf{I}_j$ is the set of samples belonging to the j-th leaf in the new decision tree.

3.2 Importance of weather features

In this study, 208 complete weeks are extracted from the outage records of local distribution system for analysis. The number of weeks in two outage severity levels are shown in Fig 2.

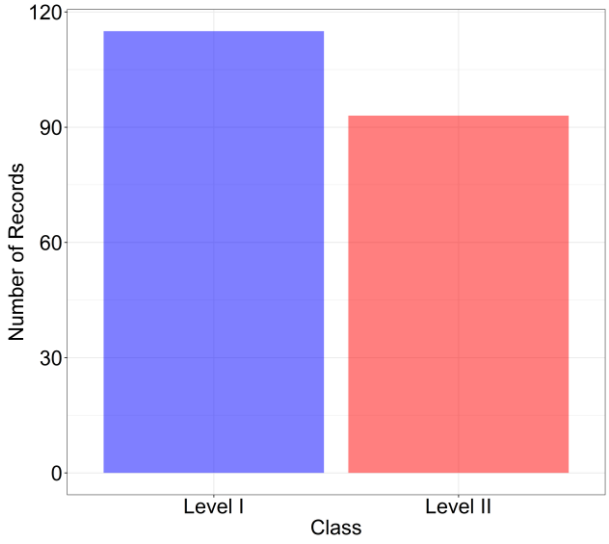


Fig 2 Outage levels of weeks from 2014 to 2017

As can be seen from Fig 2, less than half of the records are labeled as outage severity Level II, which indicates a pretty reliable performance of the distribution network. By randomly sampling from these weeks, 70% of them are used as training dataset for

building the binary classification model with XGBoost, and the rest will be used for evaluating the performance of the model.

When constructing the classification model with boosting technique, the feature's importance could be calculated and ranked at the same time. it indicates how useful this feature is in building the model. For decision tree, the impurity of samples under the leaves are normally represented with the Gini index. The feature importance is evaluated by how much the impurity could be decreased after splitting this feature in the iterations. The final importance is the averaged value of all the decision trees. With the highest feature importance normalized as 100, the top 6 weather features affecting the outage level are shown in Fig 3.

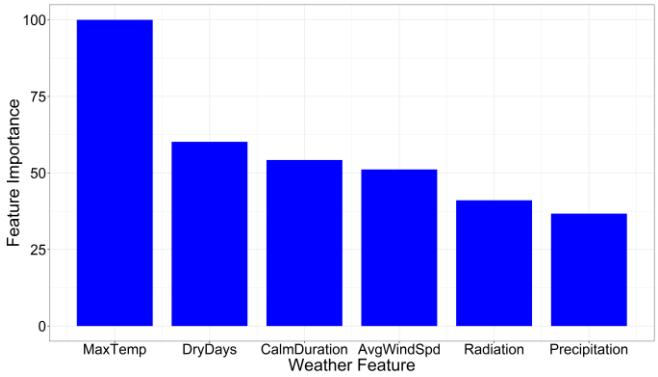


Fig 3 Weather Feature Importance

3.3 Performance of the classification model

As a typical binary classification model, the initial output of our model is the probability of each outage level with given weather conditions. Therefore, a probability threshold is needed for identifying the different outage levels. Instead of using 0.5 as default threshold, this study compared the FPR and TPR of the classification results regarding to different probability threshold and built the ROC curve [7]. When the threshold equals to 0.442, the classification model could get the highest TPR and lowest FPR. The area under ROC curve reaches 0.756, which shows an acceptable performance of the classification model.

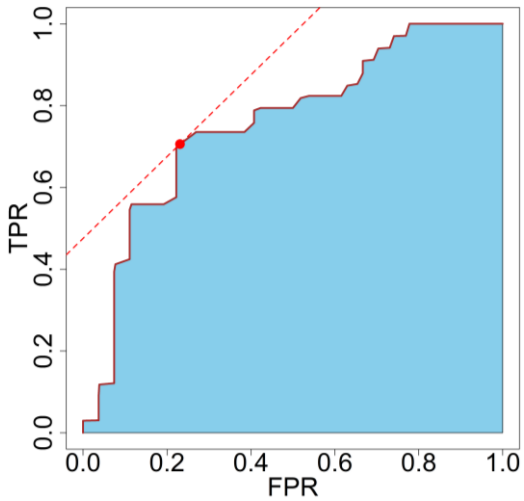


Fig 4 ROC of the classification model

4. CONCLUSION

In this paper, a data-driven model is built for revealing the impacts from weather conditions on the reliability of the distribution system. With the weekly performance defined as outage levels, a binary classification model is built with the ensemble learning technique. The weather features' importance is evaluated to show their different influence on the outages. The final output of the classification model is improved with the ROC curve to get the balance between TPR and FPR. The future work will focus more on a wider collection of effective features and apply the prediction results for predictive maintenance.

REFERENCE

[1] Doostan M, Chowdhury BH. Power distribution system fault cause analysis by using association rule mining. Electric Power System Research, 2017, 152: 140-147.

[2] Costa IC, Venturini LF, Rosa MA. Wind speed severity scale model applied to overhead line reliability simulation. Electric Power Systems Research, 2019, 171: 240-250.

[3] Li GF, Zhang P,Luh PB, et al. Risk analysis for distribution systems in the northeast U.S. under wind storm. IEEE Transactions on Power Systems, 2014, 29: 889-898

[4] McRoberts DB, Quiring SM, Guikema SD. Improving hurricane power outage prediction models through the inclusion of local environmental factors. Risk Analysis, 2018, 38: 2722-2737

[5] Nobre J, Neves RF. Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets. Expert Systems With Applications, 2019, 125: 181-194

[6] Chen TQ, Guestrin C. XGBoost: A scalable tree boosting system. The 22nd International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA — August 13 - 17, 2016

[7] Darzi MRK, Niaki STA, Khedmati M. Binary classification of imbalanced datasets: the case of coil. Expert Systems with Applications, 2019, 128: 169-186