Analysis of Outage Frequency and Duration in Distribution Systems using Machine Learning

Tumininu Lawanson*, Vinayak Sharma*, Valentina Cecchi*†, Tao Hong‡
*Department of Electrical & Computer Engineering
†Energy Production & Infrastructure Center
‡Department of Systems Engineering and Engineering Management
University of North Carolina at Charlotte
Charlotte, USA
{tlawanso, vsharm12, vcecchi, tao.hong}@uncc.edu

2020 52nd North American Power Symposium (NAPS) | 978-1-7281-8192-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/NAPS50074.2021.9449708

***Abstract*—Outage duration plays an important role in assessing the impacts of distribution system outages. Moreover, whenever an outage occurs, customers are most concerned about when power will be restored, i.e. the duration of the outage. Hence, this paper presents an analysis of the frequency and duration of outages using outage data from a distribution system network. In addition, this study performs a feature importance analysis by using random forests and gradient boosting regressors to determine which features in the outage dataset are most important in predicting the duration of an outage. The results show that climatic description, failed equipment and wind speed are the most important predictors of outage duration in the dataset used in this analysis.**

***Index Terms*—gradient boosting, random forests, outage duration, outage frequency, outage management systems**

I. Introduction

Outage frequency and duration impact system reliability and customer satisfaction. With regards to outages, customers are most concerned about the duration of outages [1]. A major priority of utilities is to reduce the amount of time that customers are left without power. The impact of outages on customers can range from inconvenience and stress (for residential customers) to loss of revenue and man-hours (for commercial and industrial customers). The study in [2], which was conducted in 2015, estimates that the average cost per event for a momentary outage in the United States ranges from around $3.9 for residential customers to as high as $12,952 for medium and large commercial and industrial customers. Results from the same study show that the average cost per event rises as the outage duration increases. Hence, there is a need to study factors that significantly impact the duration of an outage.

Much of the current literature pays attention to analyzing outages based on their frequency and causes; typically these studies focus on using machine learning techniques to predict the cause of outages or to analyze outages based on a particular outage cause (typically, trees and animals) [3]–[7].

Conversely, fewer studies have analyzed factors that impact outage duration. Reliability indices can be improved by reducing not only outage frequency, but also outage duration. One of the most common distribution system reliability indices, Customer Average Interruption Index (CAIDI), represents the average time to restore service after an outage. Authors in [8] investigate the impact of several variables on time of outage restoration (TOR) in distribution systems using statistical methods and measures such as the chi-square approximation to Kruskal-Wallis test and the coefficient of determination (R^2). The variables considered in the analysis were categorized under time (hour of day, day of week and month), consequence (number of phases affected and protection device activated) and external factors (weather condition and outage cause). Similarly, [9] presents an analysis to assess the impacts of different features on outage duration in a distribution network. Some features considered in the study include outage cause, action taken by repair crew, weather conditions, clearing device, number of customers and calendar variables such as year, month, and hour of day. On the other hand, [1] uses recursive neural networks (RNN) to predict the duration of distribution system outages in real-time. Data used in this study include weather information, outage reports and repair logs. Outage causes are identified by applying natural language processing to utility outage reports.

This paper analyzes the impact of several features on outage frequency and duration in a distribution network using random forest and gradient boosting regression. The analysis uses the frameworks presented in [8] and [9]; the features considered in this analysis include: outage cause, interrupted phase, voltage level of the affected circuit, climatic description, and calendar variables. The impact of these features are ranked using random forest and gradient boosting regression.

The rest of the paper is organized as follows: section II discusses the data used in this study including the features and data processing conducted; section III presents an exploratory analysis on the frequency and duration of the outages based on several features in the outage dataset; the feature importance analysis is discussed in section IV, and section V concludes the paper and provides some recommendation for future work.

II. Data

This study uses outage data obtained from an electric power utility in southeastern United States. The dataset, which

This work is supported in part by the National Science Foundation under Grant No. 1839812.

TABLE I
SUMMARY OF FEATURES IN OUTAGE DATASET

Features	*Classes*
Climatic Description	Calm, Precipitation-Rain, Thunderstorm Wind & Precipitation
Day of Week	Mon, Tue, Wed, Thu, Fri, Sat, Sun
Interrupted Phase	A, AB, ABC, AC, B, BC, C
Month	Jan, Feb, Mar, Apr, May, Jun Jul, Aug, Sep, Oct, Nov, Dec
Outage Cause	Third Party, Animal, Equipment Failure Event Response, Lightning, Other, Tree, Unknown
Season	Fall, Spring, Summer, Winter
Voltage Level	4 kV, 12 kV, 46 kV, 161 kV
Year	2016, 2017, 2018

comprises over 20,000 entries, includes outage information from 2016 to 2018 for an electric power distribution network. Prior to analyzing the data, data cleansing is performed by removing duplicates and missing entries from the dataset.

Features in the dataset include: climatic description during the outage, voltage level of the circuit affected by the outage, outage cause, outage duration, interrupted phase and failed equipment. In addition to the features in the original dataset, the date of the outage is decomposed into new features: year, month, day of the week and season. Table I presents a summary of the features from the outage dataset along with their respective classes. The failed equipment feature (not listed in Table I) comprises over 20 classes, some of which include: transformer, switchgear, regulator, meter, and conductor.

This study also uses weather information for the distribution network location, sourced from OpenWeatherMap, an online weather data service [10]. The weather variables considered are: temperature (Fahrenheit), wind speed (miles/hour) and humidity (%).

III. Data Analysis

This section presents results from exploratory analysis of the outage data. The features listed in the previous section are analyzed based on outage frequency and average outage duration (in minutes).

A. Number of Outages

Fig. 1 shows a breakdown of the outage events by cause. Outages due to trees are the most frequent and account for 38.5% of the outages, while outages caused by a third party are the least frequent, and account for 3.4% of the outages. Outages attributed to third party include outages caused by vehicle accidents and contractor dig-ins. It is interesting to note that the cause of nearly 7% of the outages is categorized as Unknown. Outages categorized under *Event response* refer to outages caused by opening a protection device for repair purposes, whereas outages categorized as *Other* include outages that do not fall into any of the other cause categories shown.

Fig. 2 presents plots of outage frequency with respect to each feature in the outage dataset.

Fig. 2a shows the number of outages categorized by the climatic condition at the time of the outage. The climatic description feature has four classes: calm, wind and precipitation, precipitation-rain, and thunderstorm. About 75% of the outages occur in calm weather, while 18% of the outages occur during thunderstorms. The precipitation-rain class has the lowest number of outages.

Fig. 2b presents the number of outages categorized by each day of the week. This is done to identify any seasonality that might be present due to changing load profiles for different days of the week. Monday has the highest number of outages, followed by Saturday, while Friday and Sunday have the least

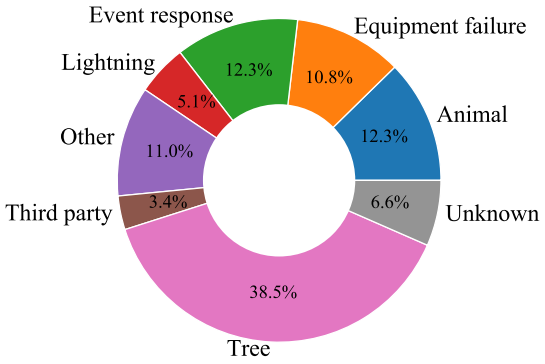
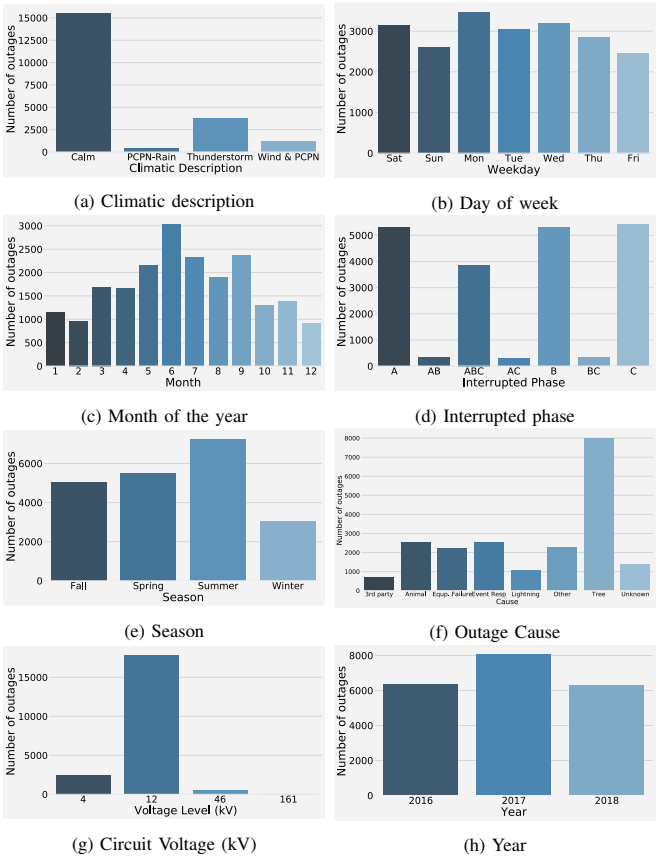


Fig. 1: Distribution of outage frequency by cause



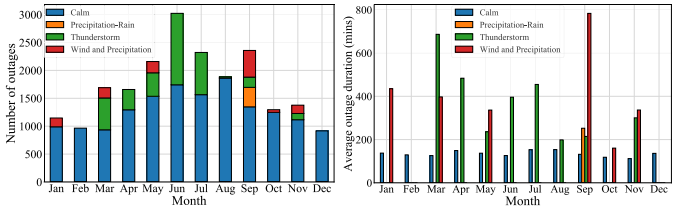
(a) Climatic description (b) Day of week

(c) Month of the year (d) Interrupted phase

(e) Season (f) Outage Cause

(g) Circuit Voltage (kV) (h) Year

Fig. 2: Outage Frequency with respect to different features in outage data set



(a) Outage Count (b) Average outage duration (mins)

Fig. 3: Monthly plots of outage count and average outage duration categorized by climatic description

outages. Thursday, Tuesday and Wednesday have very similar number of outages.

Similarly, Fig. 2c presents the number of outages categorized by the months of the year. The month of June stands out with the maximum number of outages, nearly 15% of the total. It can be observed that in the initial months of year, i.e. January, February and March, the outages are low. The outages begin to rise in the months of April, May, peaking in June. After June, the number of outages begin to decrease until September, which has higher number of outages and then the number of outages decrease till December.

Fig. 3a shows the distribution of outages by climatic description and by month of the year. The highest number of outages during thunderstorm occurs in June. It is worth noting that September is the only month that has all four climatic description classes present. Further investigation revealed that the highest frequency of outages during wind and precipitation occurred in September 2017, and this coincides with the period Hurricane Irma struck the US. On the other hand, June and July account for the most outages during thunderstorms, while August accounts for the most outages during calm weather.

Fig. 2d presents the number of outages by the interrupted phase. 75% of the outages affect only a single phase (A, B, or C) with phase C having the most outages. This is not surprising as single-phase faults are most common faults in distribution systems [11]. 18% of the outages affect all three phases (ABC) simultaneously. On the other hand, less than 5% of the outages affect only two phases (AB, AC or BC) at the same time.

Fig. 2e shows the number of outages by season. The months of the year are categorized into four seasons as follows: Spring (March to May), Summer (June to August), Fall (September to November), and Winter (December to February). As was observed from the monthly plot, the maximum outages are in the summer months and the least in winter. The lower number of outages during the winter could be attributed to the location of the the distribution network location, which typically experiences mild winters. However, this location experiences a significant number of tornadoes, hurricanes and thunderstorms, in the other three seasons. Hence, the number of outages in each of these seasons are at least 1.5 times as much as the number of outages in the winter.

Outage frequency by cause categories is presented in Fig. 2f. As previously stated, outages caused by trees are the most frequent, while outages caused by a third party are the least frequent. Fig. 2g shows the number of outages by voltage level (kV) of the affected circuit. The most number of outages occur in the 12 kV circuits. Most of the circuits in the distribution network in this study operate at the 12 kV level.

Fig. 2h shows the number of outages per year. 2017 has the highest number of outages.

B. Average Outage Duration

This section presents a visual analysis for the average outage duration in minutes by each category. The histogram in Fig. 4 shows the distribution of outage duration (in minutes) of

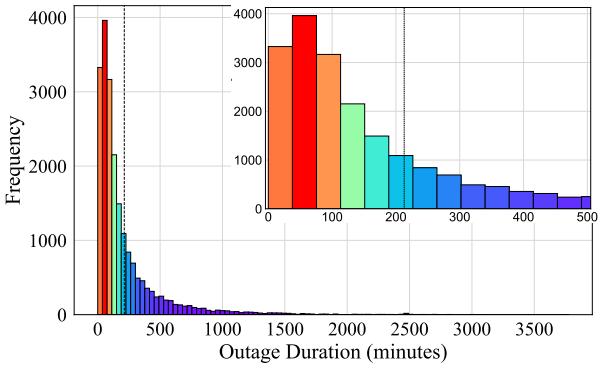
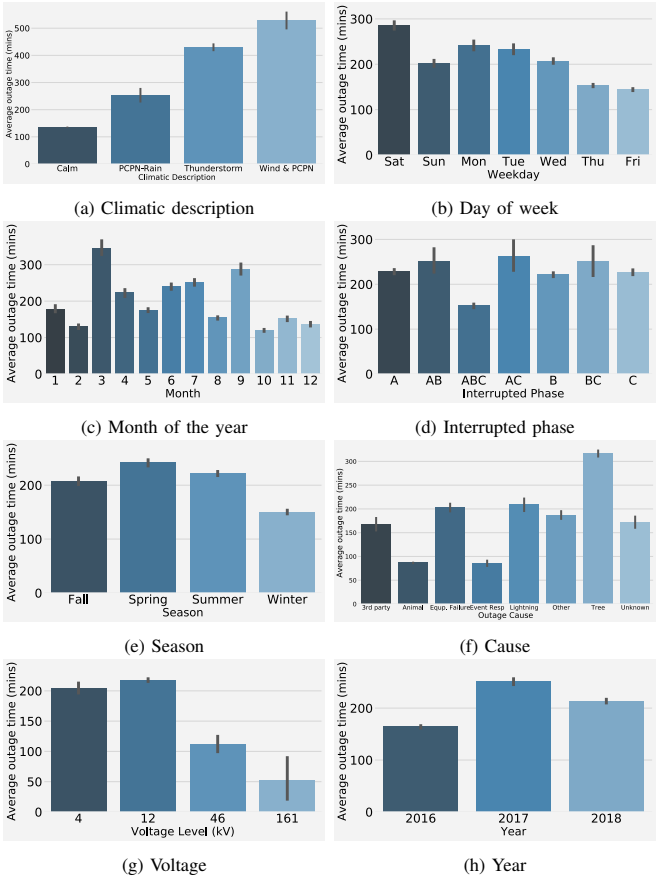


Fig. 4: Histogram of outage duration (minutes)



(a) Climatic description (b) Day of week

(c) Month of the year (d) Interrupted phase

(e) Season (f) Cause

(g) Voltage (h) Year

Fig. 5: Average outage duration (minutes) with respect to features in outage data set

outages in this analysis. As seen in the histogram, most of the outages last between 38 to 76 minutes. The shortest outages lasted for only 1 second, while the longest outage lasted for about 2 days. It is interesting to note that close to half of the outages lasted at least 2 hours or more. The average duration of all the outages in the dataset (indicated with a dashed line in Fig. 4) is around 3.5 hours (213 minutes to be precise).

Fig. 5a shows the average outage duration by climatic description. Although calm weather had maximum number of outages, the average duration of the outages during calm weather is the lowest. The maximum average outage duration is during wind & precipitation and thunderstorm. This could be due to the amount and severity of damage caused by severe weather events, hence leading to longer repair times and outage duration compared to outages that occur in calm weather.

Fig. 5b shows the average outage duration per weekday. Maximum average outage duration occurs on Saturday. Fig. 5c shows the average outage duration per month of the year. March has the highest average outage duration whereas October has the least average outage duration. It is interesting to note that although the month of March ranked 6th in the frequency of outages, it has the highest average outage duration. Further investigation reveals that this is due to long duration outages occuring during thunderstorms as well as wind and precipitation in March as shown in Fig. 3b. September has the highest average outage duration during wind and precipitation.

Fig. 5d categorizes the average outage duration by the interrupted phase. Although outages affecting two phases simultaneously (AB, AC, BC) accounted for the lowest number of outages, they result in higher average outage durations as shown in the figure. On the other hand, outages affecting all three phases (ABC) account for the lowest outage duration.

Fig. 5e shows the average outage duration for each season. Spring and summer have higher average outage durations than the fall and winter. Fig. 5f shows the average outage duration per outage cause. Outages caused by trees have the highest average outage duration, followed by outages caused by lightning and equipment failure. Fig. 5g shows the average outage duration by voltage level in kV. It can be observed that the outage duration is higher for the 12 kV and 4 kV circuits compared to the 46 kV and 161 kV circuits. Fig. 5h shows the average outage duration by year. 2017 has the highest average outage duration.

Table II presents a summary comparison of outage frequency and average outage duration for the features considered in this analysis. The results in the table show that the class with the most number of outages does not automatically account for the highest average outage duration. For example, with respect to the month of the year, the highest number of outages occurred in June, but March had the highest average outage duration. On the other hand, with respect to interrupted phase, outages affecting phases A and C simultaneously accounted for the least number of outages, but had the longest average outage duration.

TABLE II
OUTAGE FREQUENCY AND DURATION FOR OUTAGE FEATURES

	Number of Outages		***Average Outage Duration***	
Features	***Highest***	***Least***	***Longest***	***Shortest***
Interrupted Phase	C	AC	AC	ABC
Month	June	December	March	October
Outage Cause	Tree	Third Party	Tree	Event Response
Season	Summer	Winter	Spring	Winter
Voltage Level	12 kV	161 kV	12 kV	161 kV
Climatic Description	Calm	Precipitation (Rain)	Wind and Precipitation	Calm
Weekday	Monday	Friday	Saturday	Friday

IV. FEATURE IMPORTANCE

In addition to exploring outage frequency and average outage duration, this study seeks to determine the features or variables in the dataset that affect average outage duration. To rank the importance of each variable, two machine learning-based approaches are used: Random Forest Regressor and Gradient Boosting. Both techniques are implemented using Python's Scikit-Learn library [12]. Feature importance is estimated by calculating the ratio of the number of samples that get through to a node to the total number of samples [13].

A. Random Forest Regressor

Random Forest is a tree-based supervised learning algorithm introduced in [14]. The random forest algorithm is a bagging-based algorithm that takes the ensemble of randomly sampled trees [15]. A random Forest regressor-based model is used to rank the various features based on their importance. The random forecast model is trained using the entire outage dataset and the importance of each feature is estimated. The random forest model is modeled with 150 trees. The number of trees was selected by performing a grid search, varying the trees from 30 to 300 trees and comparing their prediction score.

Fig. 6 shows the feature importance as estimated by the random forest algorithm. It is observed that climatic description has the maximum importance followed by failed equipment and wind speed. Interrupted phase, outage cause and temperature are moderately important while humidity, month, weekday, voltage level, year and season have very low importance. In general, calendar variables like year, month, season and weekday have very less importance, showing that there is no significant seasonal pattern in the dataset used in this study.

B. Gradient Boosting Regressor

Gradient boosting regressor is a supervised learning algorithm introduced in [16]. As the name suggests, gradient boosting is a boosting-based approach that uses decision trees and selects the best trees using a gradient loss function [17]. The gradient boosting model is trained using the entire outage dataset and the importance of each feature is estimated. The gradient boosting model is modeled with 100 trees. The

number of trees was selected by performing a grid search, varying the trees from 30 to 300 trees and comparing their prediction score.

Fig. 7 shows the feature importance according to the gradient boosting algorithm. Consistent with results from the random forest model, climatic description has the highest importance, however the magnitude of importance is more, followed by failed equipment and wind speed. Temperature ranks higher than interrupted phase and outage cause; this is different from the results of the random forest model. Season, year, voltage level and weekday are the features with the least importance.

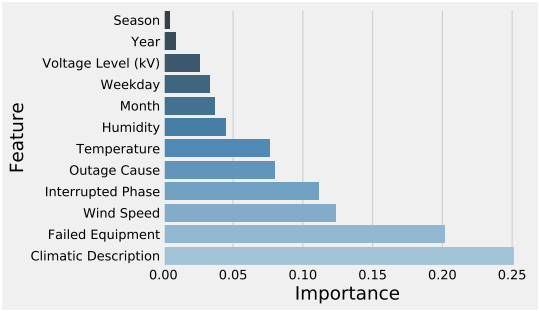


Fig. 6: Feature Importance using Random Forest Regressor

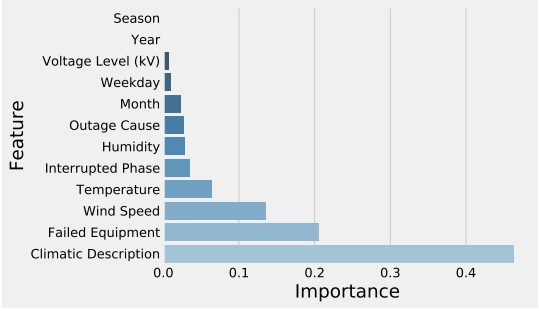


Fig. 7: Feature Importance using Gradient Boosting Regressor

V. Conclusion

This paper presented an analysis of the frequency and average duration of outages in a distribution network using the frameworks presented in [8] and [9]. Also, random forests and gradient boosting regression are used to rank the importance of several features in predicting outage duration. The results from both regressors show that climatic description is the most significant for explaining the variability of outage duration for the distribution network considered in this study. Other significant features include: failed equipment, wind speed and interrupted phase. Future work will focus on data-driven probabilistic outage prediction using weather data.

Acknowledgment

The authors would like to thank our utility partner, Electric Power Board of Chattanooga (EPB) for providing the data used in this study.

References

[1] A. Jaech, B. Zhang, M. Ostendorf, and D. S. Kirschen, "Real-Time Prediction of the Duration of Distribution System Outages," *IEEE Transactions on Power Systems*, vol. 34, no. 1, pp. 773–781, Jan 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8421644/

[2] M. Sullivan, J. Schellenberg, and M. Blundell, "Updated Value of Service Reliability Estimates for Electric Utility Customers in the United States," Lawrence Berkeley National Laboratory (LBNL), Berkeley, CA (United States), Tech. Rep. LBNL-6941E, Jan 2015. [Online]. Available: https://www.osti.gov/biblio/1172643

[3] M. Doostan and B. Chowdhury, "Statistical Analysis of Animal-Related Outages in Power Distribution Systems - A Case Study," in *2019 IEEE Power & Energy Society General Meeting (PESGM)*. Atlanta, GA: IEEE, Aug 2019, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/8973448/

[4] P. Kankanala, A. Pahwa, and S. Das, "Regression models for outages due to wind and lightning on overhead distribution feeders," in *2011 IEEE Power and Energy Society General Meeting*. Detroit, MI: IEEE, Jul 2011, pp. 1–4. [Online]. Available: https://ieeexplore.ieee.org/document/6039747/

[5] M. Gui, A. Pahwa, and S. Das, "Analysis of animal-related outages in overhead distribution systems with wavelet decomposition and immune systems-based neural networks," *IEEE Transactions on Power Systems*, vol. 24, no. 4, pp. 1765–1771, nov 2009. [Online]. Available: http://ieeexplore.ieee.org/document/5282388/

[6] L. Xu, M.-y. Chow, and L. Taylor, "Data Mining and Analysis of Tree-Caused Faults in Power Distribution Systems," in *2006 IEEE PES Power Systems Conference and Exposition*. Atlanta, GA: IEEE, 2006, pp. 1221–1227. [Online]. Available: http://ieeexplore.ieee.org/document/4075920/

[7] Mo-yuen Chow and L. Taylor, "Analysis and prevention of animal-caused faults in power distribution systems," *IEEE Transactions on Power Delivery*, vol. 10, no. 2, pp. 995–1001, Apr 1995. [Online]. Available: http://ieeexplore.ieee.org/document/400829/

[8] Mo-Yuen Chow, L. Taylor, and Mo-Suk Chow, "Time of outage restoration analysis in distribution systems," *IEEE Transactions on Power Delivery*, vol. 11, no. 3, pp. 1652–1658, Jul 1996. [Online]. Available: http://ieeexplore.ieee.org/document/517530/

[9] M. Doostan and B. H. Chowdhury, "A data-driven analysis of outage duration in power distribution systems," in *2017 North American Power Symposium, NAPS 2017*. IEEE, Sep 2017, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/8107170/

[10] OpenWeatherMap. Accessed: May 13, 2020. [Online]. Available: https://openweathermap.org/

[11] J. D. Glover, M. S. Sarma, and T. J. Overbye, "Unsymmetrical faults," in *Power System Analysis and Design*, 5th ed. Stamford, CT: Cengage Learning, 2012, ch. 9, p. 471.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[13] S. Ronaghan, "The mathematics of decision trees, random forest and feature importance in scikit-learn and spark," Nov 2019. [Online]. Available: https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3

[14] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[15] B. Wolff, O. Kramer, and D. Heinemann, "Selection of numerical weather forecast features for pv power predictions with random forests," in *Data Analytics for Renewable Energy Integration*, W. L. Woon, Z. Aung, O. Kramer, and S. Madnick, Eds. Cham: Springer International Publishing, 2017, pp. 78–91.

[16] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367 – 378, 2002, nonlinear Methods and Data Mining. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167947301000652

[17] S. B. Taieb and R. J. Hyndman, "A gradient boosting approach to the kaggle load forecasting competition," *International Journal of Forecasting*, vol. 30, no. 2, pp. 382 – 394, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169207013000812

[18] D. Hall, "Outage management systems as integrated elements of the distribution enterprise," in *2001 IEEE/PES Transmission and Distribution Conference and Exposition. Developing New Perspectives (Cat. No.01CH37294)*, vol. 2. Vancouver, BC: IEEE, Jul 2001, pp. 989–991. [Online]. Available: http://ieeexplore.ieee.org/document/970191/

[19] L. Lawton, M. Sullivan, K. V. Liere, A. Katz, and J. Eto, "A Framework and Review of Customer Outage Costs: Integration and Analysis of Electric Utility Outage Cost Surveys Environmental Energy Technologies Division," Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA, Tech. Rep. LBNL-54365, Nov 2003. [Online]. Available: https://emp.lbl.gov/sites/default/files/lbnl-54365.pdf

[20] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[21] M. Bessani, R. Z. Fanucchi, J. A. Achcar, and C. D. Maciel, "A statistical analysis and modeling of repair data from a Brazilian Power Distribution System," in *2016 17th International Conference on Harmonics and Quality of Power (ICHQP)*. Belo Horizonte: IEEE, Oct 2016, pp. 473–477. [Online]. Available: http://ieeexplore.ieee.org/document/7783446/