Prediction Model for the Distribution Transformer Failure using Correlation of Weather Data

Eun Hui Ko, Tatjana Dokic, Mladen Kezunovic
Department of Electrical and Computer Engineering Texas A&M University College Station, TX, U.S.A.
E-mail: kezunov@ece.tamu.edu

***Abstract*—Distribution Transformer (DT) is an integral component of a distribution network. Electric utilities have invested efforts in reducing DTs failure rates. This paper presents a method for prediction of probability of DT failure by analyzing a correlation between weather data and historical DT failures data. Logistic regression prediction model is used in order to predict DT failure, and to extract the correlation between parameters of weather and DT failure. Accuracy of prediction is reliable, which is presented using evaluation metrics. This method not only has a vital significance for the maintenance of DTs, but also improves the economic efficiency and reliability of distribution network operation.**
***Index Terms*— distribution transformer, failure, logistic regression, prediction model, weather data**

I. INTRODUCTION

The Distribution Transformer (DT) is a vital link in the chain of power apparatus supplying electric power to the customers. DT failures have a major financial impact on both utility company and customers. The utility companies invest large funds annually for maintenance of distribution transformers. Predictions of transformer failure rates is regularly assessed and DTs may be replaced according to set of criteria during the evaluation. However, there is still a limited capability to mitigate DT failures caused by weather. A practical utility example given in Table 1 illustrates that the rate of DT failures caused by weather is still high and presents the second highest cause after aging, also affected by weather.

The factors of DT failure consist of aging, weather, overloading, corrosion, out of maintenance, animal contact, installment error, people's error, etc. Among the causes, aging, corrosion and overloading constitute 47% (Table 1), which combined is the largest reason. In order to decrease this failure rate caused by aging, corrosion and overloading, utility company implements lots of methods including electric measurement methods for electrical insulation tests [1], Dissolved Gas Analysis (DGA) [1], Oil and paper tests and thermography for chemical and physical tests [1].

TABLE 1. THE CAUSES OF DT FAILURE IN JEONLLANAMDO OF SOUTH KOREA (2011-2018)

Cause	The number	Rate [%]
Aging	337	27.2
Weather	333	26.9
Corrosion	155	12.5
Animal contact	145	11.7
Out of maintenance	92	7.4
Overloading	87	7.0
Object contact	25	2.0
Tree	18	1.5
People error	16	1.3
Manufacture error	15	1.2
Flooding	13	1.0
Installment error	3	0.2
Fire	1	0.1
Total	1,240	100

While there are a quite a few papers that study the causes of DT failure, very limited research is done on prediction of weather-related distribution transformer failures. One of them [2] analyzes the root causes of failure, overvoltage and overloading, as well as the surge affecting the core and winding failure. The other study [3] shows the main causes (Tree, animal, contamination, nature disaster, and human) according to geographical regions. According to study about DT failure root causes in India, the reasons for failure are overload, lighting surge, moisture, and high voltage [4]. However, it does not deal with weather impact as a principal cause. Such analysis focuses on lightning-induced voltages and correlation between LV surge arresters and grounding resistance.

The utility companies need a practical method to predict weather-related failures and implement efficient methods for mitigation. Our paper provides such an approach by using logistic regression to analyze weather-related failures. The main advantage is that it can be used as an input to the replacement decision-making hence alleviating the limits of existing maintenance methods. It separates weather parameters causing failure, and expresses them with numerical values. In addition, it has high accuracy evaluated using objective evaluation metric.

The paper is organized as follows: Section II discusses the problem formulation. Section III demonstrates the data

source and processing. Section IV explains prediction model, and Section V describes evaluation. Section VI analyzes result, and Section VII contains conclusions.

II. PROBLEM FORMULATION

Power utility company are aware that there is a correlation between weather and DT failure according to their field experiences. However, it is typically not known which weather parameter is most relevant among different parameters, and such correlation is not quantified. In order to verify the correlation between weather parameters and failure, it is necessary to collect historical weather and DT failure data. Extracting DT failure data is the first step. In our case we are considering data from South Korea. The area where failure occurred the most in South Korea is selected and events are sorted by the dates of failure and areas. The next step is matching the areas of failure and area where weather stations are located. The weather stations are selected at one of the closest locations.

The last step is extracting the historical weather data according to dates and area of DT failures. The weather station provides weather data about Lightning, Average Temperature [°F] (AT), Highest Temperature [°F] (HT), Relative Humidity [%] (RH), Maximum Wind Speed [m/s] (MWS), Wind Gust [m/s] (WG), and Precipitation (mm) (PT). The information of DT failure and historical weather data is processed by using logistic regression. The goal of the logistic regression implemented in this study is prediction of future failures. We calculate coefficients of correlation between parameters and failure and apply to the test set next. Through this method, the degree of correlation between parameters of weather data and DT failure is measured and the probability of each failure is calculated.

III. DATA SOURCE

We studied step down transformers (22.9KV-220V) used in the distribution sectors in South Korea (see Table 2). The 237 events comprise the data set and the predictor variables (see Table 1) are used for our study for modeling purpose. We also applied preprocessing steps to extract the useful features and prepare data for prediction algorithm. We collected the data for modeling outage events used for prediction and analysis starting from year 2012 up to year 2018. Before applying the logistic regression, let us go through our adjusted modeling data in simple descriptive statistical measures. The historical outages are extracted for five causes; lighting, tree contact, snow, rain, and dust.

TABLE 2. DISTRIBUTION FACILITIES IN JEONLLANAM-DO (UNIT:1000) [9]

Transformer			Protective Devices		
Bank	Number	Capacity (kVA)	Breaker	Equipment	COS
104	243	9,252	12	1.4	72

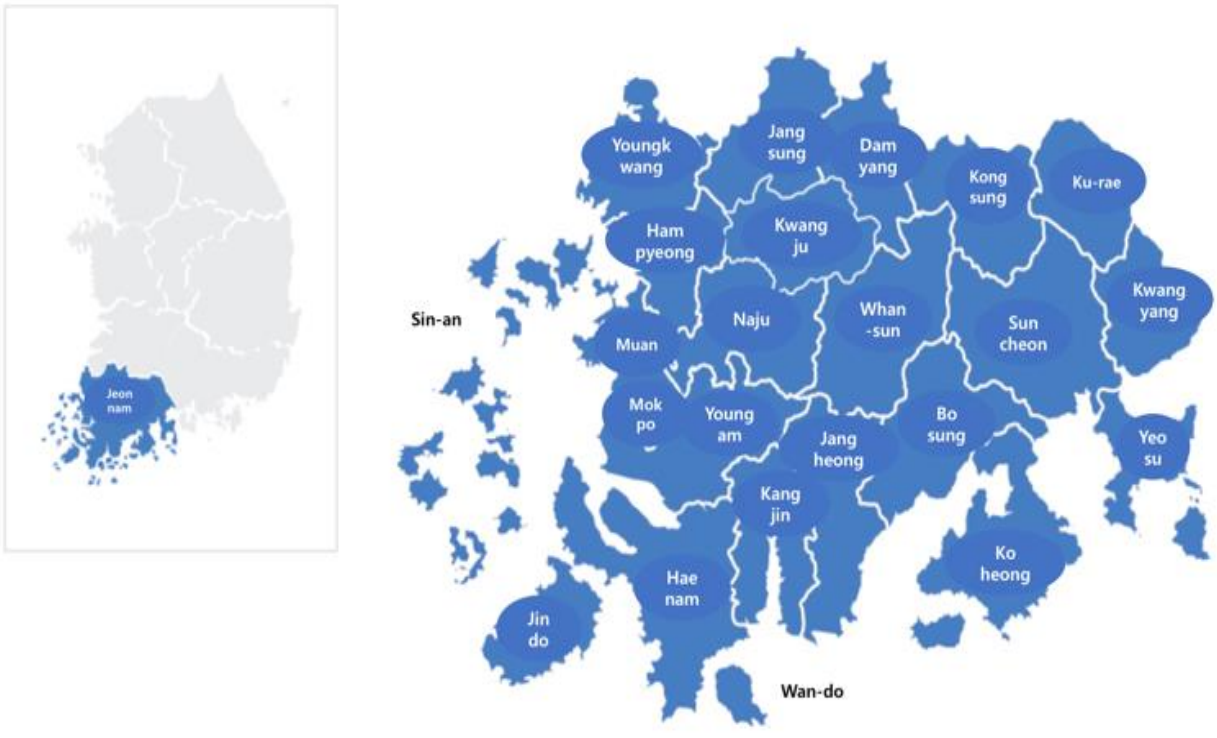


Figure 1. JeonllaNam-do area [8]

The region is JeonllaNam-do in South Korea and the size of the region is 4,729mi², the popularity is 190 million and there are 842,668 households. It consists of 22 cities as shown in Fig. 1, which account for 65% of the entire land area. The JeonllaNam-do area ranks first in the number of DTs in Korea as shown Fig 3. The area has more countryside and seaside than other states, therefore it is more vulnerable to weather impacts.

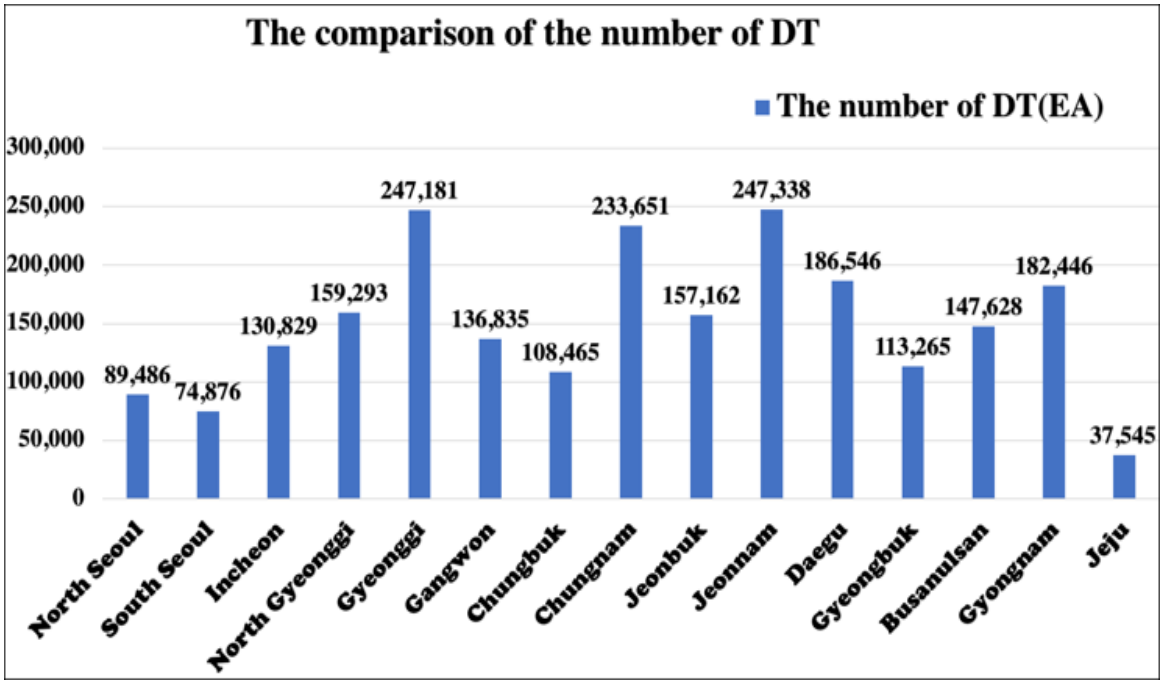


Figure 2. The comparison of the number of DT in Korea [8]

From 1/1/2011 to 11/2/2018 the number of total outages is 1,025 and failures caused by weather account for 237, which constitutes 24%. The causes of all outages include lightning, three contact, snow, aging, overload, bird contact, people fault, installation fault, manufacture fault, corrosion, fire, etc.

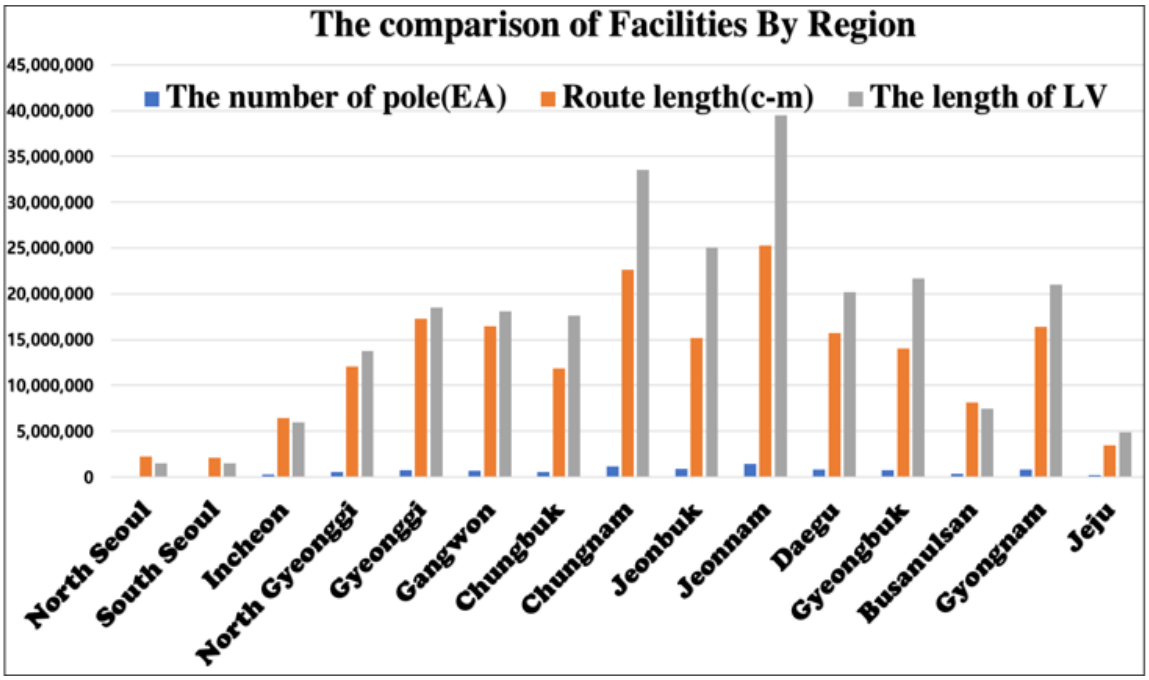


Figure 3. Distribution facilities in Korea [8]

The data for the logistic regression is extracted from outages caused by weather; lighting, rain, snow, dust. Aging and overload are related to temperature; however, those are not direct causes, thus these are excluded. The weather parameters that are taken into account are shown in Table 3.

TABLE 3. PARAMETERS OF WEATHER DATA

Lightning [0/1] (LI)	Average Temperature [˚F] (AT)	Highest Temperature [˚F] (HT)	Relative Humidity [%] (RH)
Maximum Wind Speed [m/s] (MWS)		Wind Gust [m/s] (WG)	Precipitation [mm] (PT)

The dates which have outages caused by weather are selected for Y=1 and the dates which don't have any outages are presented as Y=0 and historical weather are extracted.

IV. PREDICTION MODEL

Logistic regression is used for modeling a binary response (i.e., success/fail) in many applications [10, 11]. This model estimates the probability of the response occurring P(X) = Pr (Y=1 | X) through a linear function of explanatory variables X. In this study, it is natural that the response variable Y is a DT failure, i.e., 1 (failure) and 0 (no failure), and weather predictors like LT, AT, HT, RH, MWS, WG, PT are available for modeling logistic regression. Specifically, X is n × (p+1) design matrix where n is the number of observations and p is the number of weather predictors. Naturally, the number of coefficients is eight by seven predictors and an intercept. The corresponding coefficients β of predictors designate the effect of the weather predictors on the probability of DT failure. The basic intuition behind using maximum likelihood to fit a logistic regression model is as follows: we seek estimates such that the predicted probability of failure for each individual DT is most likely to agree with its observed failure. This intuition can be formalized using the mathematical Equations. (1), (2), and (3) as follows.

$$\boldsymbol{\beta} = [\beta_0, \ldots, \beta_7]^T \tag{1}$$

$$\ell(\boldsymbol{\beta}) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}} (1 - p(x_{i'}) \tag{2}$$

$$\hat{\boldsymbol{\beta}} = \max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) \tag{3}$$

Once the coefficients in Eq. (3) have been estimated, the probability of failure is given by

$$p(\boldsymbol{x}) = \frac{e^{\boldsymbol{x}^T\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}^T\boldsymbol{\beta}}} \tag{4}$$

V. EVALUATION

A. Evaluation Metric and Effect of Predictors

To evaluate logistic regression, Receiver Operating Characteristics (ROC) [12] graphs are useful for organizing classifiers and visualizing their performance. The Area Under Curve (AUC) [13] is used, and the estimated coefficients quantify the effect of weather on the probability of failure. AUC is the most popular metric for visualizing the performance and analyzing the coefficient serves as intuitive interpretation of the effects of predictors.

To distinguish between the actual class and the predicted class, we use the labels Y, N for the class predictions produced by a model as shown Fig 4.

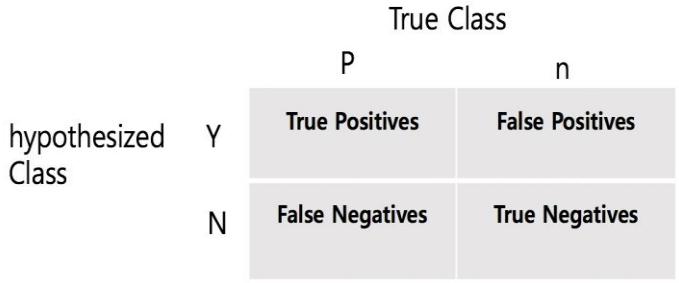


Figure 4. Confusion matrix and common performance metrics [12]

There are four possible cases. If the prediction is failure when real value is failure, it is true positive (tp), and if the prediction is failure when real value is no failure, it is false negative (fp). On the other hand, if the prediction is no

failure when real value is no failure, it is true negative, and if the prediction is no failure when real value is failure, it is false positive.

The true positive rate of a classifier is estimated as

$$\text{tp rate} \approx \frac{Positive\ correctly\ classified}{Total\ positives} \tag{5}$$

The false positive (fp) rate of the classifier is estimated as

$$\text{fp rate} \approx \frac{Negatives\ incorrectly\ classified}{Total\ negatives} \tag{6}$$

A ROC graph depicts relative tradeoffs between true positive and false positives. The ROC curve is a popular graphical method for simultaneously displaying the two types of errors for all possible thresholds. The AUC has an important statistical property. The overall performance of a classifier, which summarizes all possible thresholds, is given by the AUC. An ideal ROC curve will approach the top left corner, so the larger the AUC the better the classifier.

Coefficients show how the corresponding predictors have an impact on an outcome by describing the magnitude. Seven weather predictors are showed in X-axis and magnitude of the corresponding coefficient values are represented in Y-axis. It shows which weather parameter has the largest effect on DT failure.

B. Experimental setup

All DT failures have their own failure number and date, and the failure data spans from 2011 to 2018. The weather data is correlated through the date and location of failure [14]. The historical DT failure data is divided into the testing and training sets. The total number of DT failure is 237, and 90% of the total is selected for the training set. The remaining 10% of the data is used for the testing sets for model estimates. There is total of 148 of no failure cases used. The probability of the occurrence of 385 weather data is represented as a result of logistic regression. When a probability of DT failure is lower than 0.5, we show Y=0, and when a probability is above 0.5, we assign Y=1. The degree of high temperature (HT) is classified into three temperature thresholds such as 82.4°F, 86°F, and 89.6°F in order to make interpretation of HT coefficient precise. The basic loading of distribution transformer for a normal life expectancy is continuous loading under the operating condition in a constant 30°C (86°F) ambient temperature as discussed in [15].

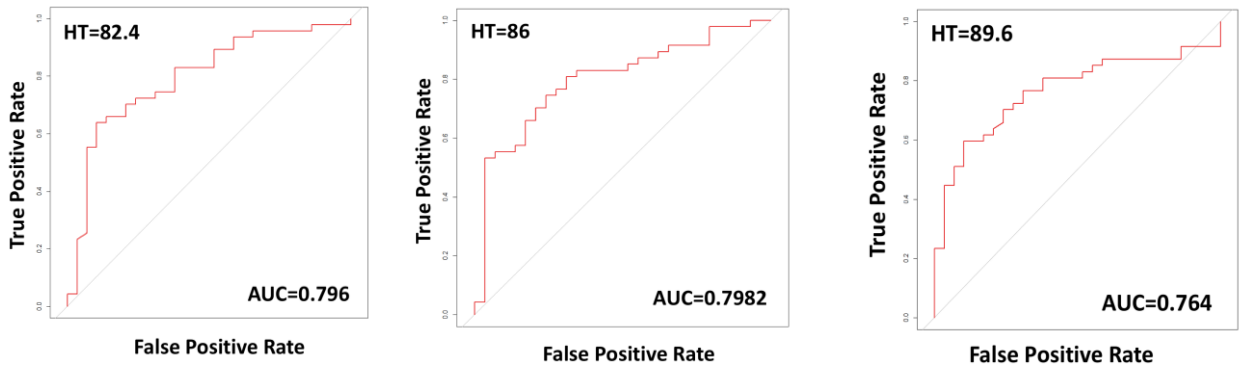
VI. RESULTS

The results are shown in Table 4. For the case of HT 82.4°F, 113 cases are predicted as no failure (i.e., Y=0), and 47 cases show prediction that there will be a failure (i.e., Y=1). For the case of HT 86°F and 89.6°F, 112 events and 111 cases are predicted as no failure respectively. On the other hand, 190 cases and 183 cases have prediction of failure (see Table 4). The accuracy of prediction is 0.7995 by calculating (113+190)/384=0.789. In the case of HT 86°F and 89.6°F, the accuracy of prediction =0.786((112+190)/384) and 0.766((111+183)/384) respectfully. For HT 82.4°F, the accuracy of prediction of probability is the highest.

TABLE 4. EVENT VS. PREDICTION OF FAILURE

Event	**Failure (Y/N)**	**Prediction**	
		Y=0	**Y=1**
HT 86°F or below	Y=0	113	47
	Y=1	35	190
HT 86°F - 89.6°F	Y=0	112	47
	Y=1	36	190
HT 89.6°F or above	Y=0	111	54
	Y=1	37	183

The AUC is 0.796, 0.798 and 0.764 respectively as shown Fig 5, which means the result can be considered as very good. The AUC of HT 86°F - 89.6°F have the highest values (see Fig 5, b)). HT is divided by three groups (0/1) according to the temperature. We assume that probability of DT failure would increase by HT. Three HTs such as 86°F or below, 86°F-89.6°F, and 89.6°F or above are selected. For example, for a certain temperature below 86°F, the probability is 0 and if the temperature is above 86°F, the value is 1.



a) HT: 86°F or below b) HT: 86°F - 89.6°F c) HT: 89.6°F or above

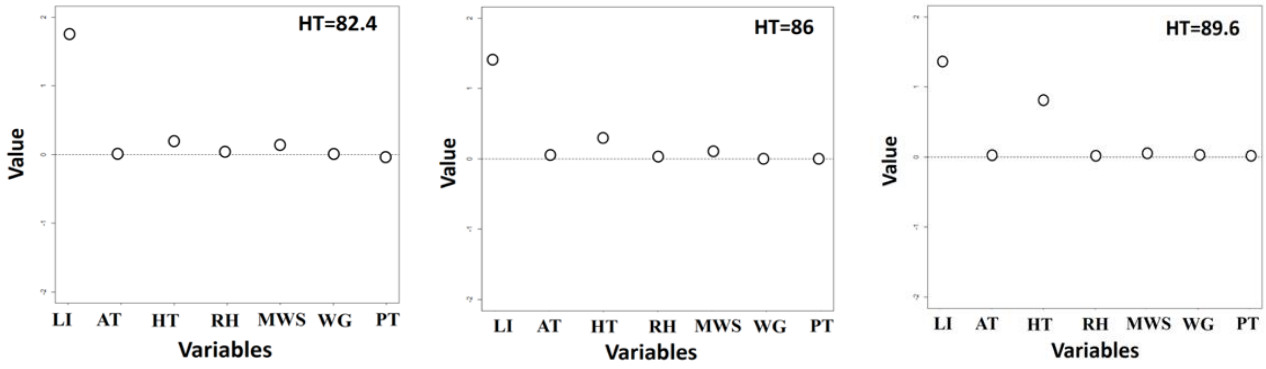
Figure 5. ROC for classification

The positive coefficient of the predictor indicates that the predictor increases the probability of DT failure, while the negative coefficient makes the probability of DT failure decrease. That is, predictor with the positive coefficient is the most relevant contributor to the failure, and predictor with the negative coefficient is the least relevant contributor to the failure. The positive coefficient is likely to have an effect on failure, and on the other hand, negative coefficient is less likely to have an influence on failure. The coefficients for DT failure shown in Table 5 and Fig 6, which are LI, AT, HT, RH, MWS and PT have all positive values. Lightning is the most influential coefficient, since it is the biggest one and HT is the second.

TABLE 5. COEFFICIENT VALUES

Degree of HT	LI	AT	HT
86°F or below	1.7455	0.003565	0.21771
86°F - 89.6°F	1.4329	0.007634	0.30125
89.6°F or above	1.3649	0.010839	0.78165

RH	MWS	PT
0.01176	0.13367	0.005994
0.008571	0.0823	0.01289
0.007447	0.060257	0.01922



a) HT: 86°F or below b) HT: 86°F - 89.6°F c) HT: 89.6°F or above

Figure 6. Coefficients value estimate

We realize that lightning and higher temperature, especially HT 89.6°F or above have the biggest effect on DT failure as shown Fig 6 c). For the coefficient value of HT, there is no huge differences between 86°F and 89.6°F from 0.21771 to 0.30125 in Table 5. However, the value increases in 89.6°F or above as 0.78165. The average Temperature (AT) has low relationship; however, it is positive. Therefore, in case AT is positive, HT coefficient turns into the important factor, which is in turn associated with higher probability of DT failure. On the other hand, RH, NWS, and PT have a low positive correlation.

VII. CONCLUSTION

The paper describes a logistic regression prediction model of distribution transformers failure by using correlation of weather parameters. The main contribution of our study are as follows: The variety of weather predictors causing DT failure has been identified including Lightning, Average Temperature [°F], Highest Temperature [F], Relative Humidity [%], Maximum Wind Speed [m/s], Wind Gust [m/s], and Precipitation (mm).

- The logistic regression model has been used to calculate the probability of DT failure caused by weather.
- The prediction model shows high-level accuracy, where the average AUC is 0.78.
- The coefficient values estimate show that the weather predictors have positive value indicating the Lightning and HT being the most important factor which affect the DT failure. HT has more effect on failure when it reaches high temperature such as 89.6℉ or above.
- The approach is a promising step to predicting not only DT failure but also other outages of power network.

VIII. ACKNOWLEDGEMENT

This work has been funded by Nationals Science Foundation under the project titled “BD Spokes: SPOKE: SOUTH: Collaborative: Smart Grids Big Data”.

IX. REFERENCE

[1] C.Rajotte et al., Guide for Transformer Maintenance, CIGRE, 2011
[2] Marín, OJ Soto, et al. "Causes of failure of Distribution Transformers in the East Zone of Caldas."
[3] Tippachon, W., et al. "Failure Analysis of Power Distribution Systems in Thailand." 2006 International Conference on Power System. Technology. IEEE, 2006.
[4] Pandit, Narasimha, and R. L. Chakrasali. "Distribution transformer failure in India root causes and remedies." 2017 International Conference. on Innovative Mechanisms for Industry Applications (ICIMIA). IEEE, 2017.
[5] Al-Arainy, A. A., N. H. Malik, and M. I. Qureshi. "A study of failure of pole mounted distribution transformers." 2012 International Conference on High Voltage Engineering and Application. IEEE, 2012.
[6] Piantini, Alexandre, et al. "Lightning-caused transformer failures in distribution systems." 2014 International Conference on Lightning Protection (ICLP). IEEE, 2014.
[7] De Carvalho, T. O., et al. "Analysis of lightning-caused distribution transformer failures." 2011 International Symposium on Lightning Protection. IEEE, 2011
[8] JeollaNamdo, JeollaNam Provincial Government, http://www.jeonnam.go.kr/contentsView.do?menuId=jeonnam0600000000
[9] EPSIS, Korea Electric Power Statistics Information system, https://epsis.kpx.or.kr/
[10] Adwere-Boamah, Joseph, and Shirley Hufstedler. "Predicting Social Trust with Binary Logistic Regression." Research in Higher Education Journal 27 (2015).
[11] Soule, Patrick. "Predicting Student Success: A Logistic Regression Analysis of Data From Multiple SIU-C Courses." (2017).
[12] Fawcett, Tom. "An introduction to ROC. analysis." Pattern recognition letters 27.8 (2006): 861-874.
[13] Gareth James, Daniela Witten, Trevor Hastie Robert Tibshiran, “An Introduction to Statistical Learning”, 2013
[14] Korea Meteorological Administration, KMA, http://www. kma.go.kr/eng/index.jsp
[15] Biçen, Yunus, et al. "An assessment on aging model of IEEE/IEC standards for natural. and mineral oil-immersed transformer." 2011 IEEE International Conference on Dielectric Liquids. IEEE, 2011.