DEVELOPING DATA-DRIVEN APPROACHES FOR ANALYZING, IDENTIFYING, AND PREDICTING POWER SYSTEM OUTAGES

by

Milad Doostan

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Electrical Engineering

Charlotte

2019

Approved by:

____

Dr. Badrul Chowdhury

____

Dr. Zia Salami

____

Dr. Valentina Cecchi

____

Dr. Churlzu Lim

____

Dr. Wenwen Dou

ABSTRACT

MILAD DOOSTAN. Developing Data-Driven Approaches for Analyzing, Identifying, and Predicting Power System Outages. (Under the direction of DR. BADRUL CHOWDHURY)

Outages that occur in power distribution systems can seriously endanger system operation in different ways. In recent years, with the explosion in data gathering within the smart grid framework, data analytics has emerged as a desirable tool in helping maintain power system security. In particular, with the deployment of an enormous number of intelligent electronic devices and various sensors, the data necessary for studying different outages have becoming available. Analyzing such data by employing analytical methods could shed light on understanding the characteristics of outages and could lead to developing models for analyzing, identifying, and predicting different outages.

In this dissertation, the focus is on developing various approaches for analyzing, identifying, and predicting outages in power systems by using rigorous statistical methodologies, data analytics, and machine learning algorithms. Developing such approaches, however, requires addressing various practical challenges. As a result, those challenges are discussed throughout the dissertation and workable solutions are provided.

The proposed approaches have the potential to provide not only a succinct view of the current system status but also more meaningful knowledge such as outage risks, and locations of potential problems, as well as suggestions on remedial actions. As such, this dissertation envisions a transformative framework that will demonstrate the use of innovative data analytics technologies for early warning of degrading operating conditions that may imperil system operation and/or quality.

DEDICATION

This is for you, Mom. No words can describe my feeling towards you and how much I love you.

ACKNOWLEDGEMENTS

I would like to extend my thanks to my Ph.D. advisor, Professor Badrul Chowdhury. Badrul is the most knowledgeable advisor and one of the smartest people that I know. His constant support throughout My Ph.D. was invaluable. He gave me the freedom to pursue various projects without objection and always provided insightful discussions about my research.

I will forever be thankful to my former advisor, Professor Zia Salami. Zia provided me with tremendous support during my admission and has been always helpful throughout my studies.

I, also, would like to thank my family, Zohreh, Alireza, Iman, Mojdeh, Sarah, Sadaf, Safoora, and Behzad and my friends, Reza Sohrabi, Saeed Mohajeryami, Iman Mazhari, Iman Naziri-Moghaddam, Mehrdad Biglarbegian, Behdad Vatani, Saman Mostafavi, Amirreza Sahami, Roozbeh Karandeh, and Mahboubeh Yazdanifar for their massive support. Achieving my goals would be impossible without your continued help and support.

Moreover, I would like to thank my other Ph.D. committee members, Dr. Valentina Cecchi, Dr. Churlzu Lim, and Dr. Wenwen Dou for their insightful suggestions and comments which helped me shape my Ph.D. research in a more realistic way.

In the end, I am forever grateful for the opportunity provided by the Electrical and Computer Engineering Department and the Graduate Assistant Support Plan (GASP) for the financial support over the past few years, which allowed me to pursue my Ph.D. with ease of mind.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Motivation

In power distribution systems, providing customers with the most reliable supply of electricity in the form of uninterrupted service is usually the most important mission of electric utility companies. Reliability plays a crucial role in reducing the cost of electricity outage and bringing customers' satisfaction. While utility companies essentially aim at maintaining the reliability at its highest level, outages caused by various reasons pose serious challenges to attaining this goal.

In order to mitigate the loss of load from outages, utilities are primarily interested in preventing avoidable outages. This includes designing more fault-resilient distribution systems, conducting more frequent inspections, installing protective devices, and scheduling regular maintenance operations.

One main course of action to effectively identify and carry out the necessary preventive actions is to acquire knowledge from past outages. For instance, if it turns out that a distribution system had experienced most of its vegetation-related outages during the months of June and July, then the utility will recognize the need to carry out essential preventive maintenance for those two months. In addition to modifying outage management practices, such knowledge is highly beneficial for improving the design of existing distribution systems to reduce the number of future outages. In order to gain this knowledge, it is required to carry out an in-depth root cause analysis for different outages.

Although taking these pro-active measures may reduce the potential risk for outages, it is often impossible to eliminate such risks. As a result, in addition to implementing preventive measures, it is important for utilities to take appropriate responses in dealing with these outages, either by identifying them immediately after they occur or by predicting them in advance. Outage prediction and identification are extremely useful as 1) they enable the utility company to better manage the outages that are unavoidable, and 2) they give the operation and maintenance crew the chance to find the potential would-be source of the problem and fix it faster.

Over recent years, with the deployment of an enormous number of intelligent electronic devices and various sensors, necessary data for studying different outages have become available. Analyzing such data by employing analytical methods could shed light on understanding the characteristics of outages and could lead to developing models for analysis, prediction, and identification of different outages. Despite the fact that there exist some studies on this important subject, due to numerous existing challenges, practical approaches that can mine the outage data to obtain a considerable amount of knowledge or approaches that can predict and identify the outages with acceptable accuracy have not yet been developed. The shortcomings of the existing approaches are discussed throughout the dissertation.

1.2 Objective and scope

In this dissertation, the focus is on developing various approaches for predicting, identifying, and analyzing outages in power systems by using data analytics and machine learning algorithms. Various practical challenges that are faced in outage-related problems are also discussed and workable solutions to address those are provided. The proposed approaches have the potential to provide not only a succinct view of the current system status, but also more meaningful knowledge, such as outage risks, and locations of potential problems, as well as suggestions on remedial actions, ultimately helping to improve system operations.

The proposed approaches may be categorized into three main groups of 1) predicting unavoidable outages 2) preventing avoidable outages by obtaining insights from the data, and 3) managing the situation during the outages. In what follows, the rationale behind each category, as well as approaches that have been proposed, are discussed.

1.2.1 Predicting unavoidable outages

Outage prediction has been mostly neglected in power distribution systems. Although predicting outages will bring enormous benefits to power companies, developing mathematical models for fulfilling this task is extremely challenging. This practical difficulty lies in the lack of necessary data, complex nature of outages, and the fact that a substantial number of factors have a strong influence on the occurrence of outages. However, as mentioned earlier, over recent year,s due to the collection of various types of data by utility companies, and advancements in data analytics methodologies, creating powerful tools that can improve utilities' predictive abilities on the outages has become possible.

Vegetation, animals, and thunderstorms are three main causes of outages. Models that already exist to predict the occurrence of these sources of outage have various shortcomings. These shortcomings will be discussed in great details in future chapters. Due to the importance of these outages and the lack of a practical predictive models, three data-driven approaches are proposed and developed to predict the occurrence of these outages in different horizons.

In order to build predictive models for outages, the general framework is illustrated in Fig. 1.2

- Understanding: Based on the domain expertise and the need of the utility companies, the problem is defined and the objective is specified. A literature review is then carried out to make sure the work is up-to-date and existing work is not repeated.

- Data acquisition: Based on the problem and objective, various types of data are collected.

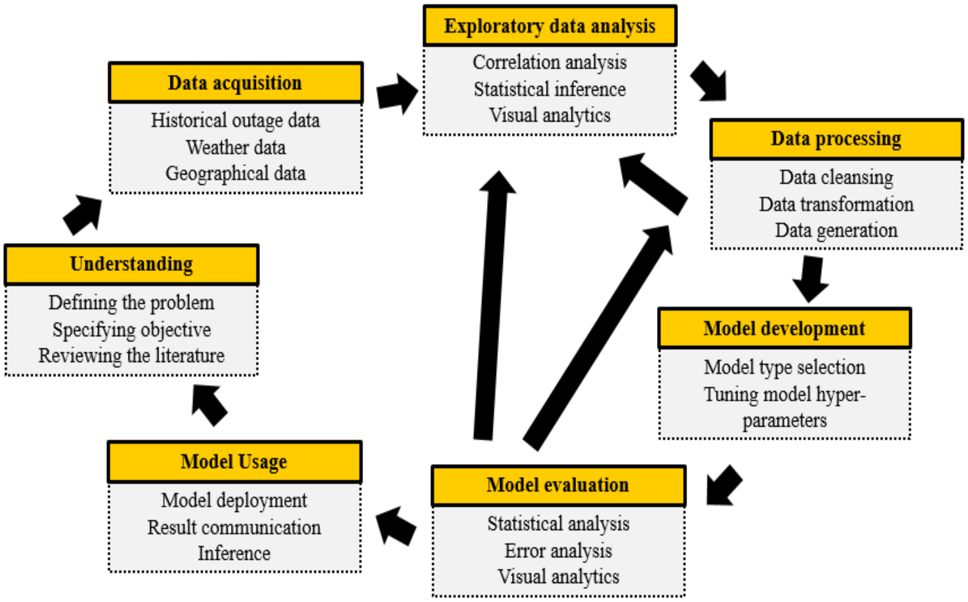


Figure 1.1: General framework for building predictive models

- Exploratory Data Analysis (EDA): By using EDA, the data is summarized and its main characteristics are found, often with visual methods. Primarily, EDA is for seeing what the data can tell beyond the formal modeling or hypothesis testing tasks.

- Data processing: After insights from data based on an initial EDA are obtained, the data cleansing (handling missing values, outliers, errors) is carried out, and new features are created or new data is generated. Then, the EDA may be repeated to explore the data, get further insight and find new relationships. This procedure is repeated to get progressively more insights, and to build more informative features.

- Model development: After the data is ready and informative features have been created, a model (SVM, NN, RF, linear, time series, etc.) is built. The hyper-parameters are tuned by k-fold cross-validation.

- Model evaluation: After the model is developed, the error of the model is investigated, visually and statistically, to see if there is any bias or unusual patterns.

If the errors satisfy several requirements, then the model is validated; otherwise, there could be a problem (outliers are not handled properly, transformations are needed, etc.). Therefore, some of the previous steps are repeated to make sure the errors look satisfactory.

- Model usage: After the model is evaluated, it will be used to do a task (prediction, etc.) or answer a specific question (e.g., why something happened).

The predictive models will be fully explained in the upcoming chapters; however, a brief description is provided below:

1.2.1.1 Predicting vegetation-related outages in power distribution systems

This study presents a novel data-driven approach for predicting the number of vegetation-related outages that occur in power distribution systems on a monthly basis. In order to develop an approach that is able to successfully fulfill this objective, there are two main challenges that ought to be addressed. The first challenge is to define the extent of the target area. An unsupervised machine learning approach is proposed to overcome this difficulty. The second challenge is to correctly identify the main causes of vegetation-related outages and to thoroughly investigate their nature. In this study, these outages are categorized into two main groups: growth-related and weather-related outages, and two types of models, namely time series and non-linear machine learning regression models are proposed to conduct the prediction tasks, respectively. Moreover, various features that can explain the variability in vegetation-related outages are engineered and employed. Actual outage data, obtained from a major utility in the U.S., in addition to different types of weather and geographical data are utilized to build the proposed approach. Finally, by utilizing various time series models and machine learning methods, a comprehensive case study is carried out to demonstrate how the proposed approach can be used to successfully predict the number of vegetation-related outages and to help decision-makers to detect vulnerable zones in their systems.

1.2.1.2 Predicting lightning-induced outages in power distribution systems: a statistical approach

This study presents a novel data-driven approach for predicting lightning-induced outages that occur in power distribution systems on a daily basis. In order to develop an approach that is able to successfully fulfill this objective, there are two main challenges that needs to be addressed. The first challenge is to define the extent of the target area. An unsupervised machine learning approach is proposed to overcome this difficulty. The second challenge is to adequately identify characteristics of lightning-induced outages and to explore the relationship between these outages and weather-related variables (thunderstorm events). In this study, these outages are clustered into a few manageable groups. Then, a probabilistic model is presented to estimate the likelihood of each group of outages. Finally, a machine learning classification algorithm that can handle the imbalanced problem is developed to predict whether the event will lead to zero outage, one outage, or two or more outages on a specific day in a specific area of the system under study. Actual outage data, obtained from a major utility in the U.S., in addition to radar weather forecast data are utilized to build the proposed approach. Also, three case studies are provided to show several issues associated with predicting lightning-induced outages, and to demonstrate how the proposed approach can address those problems adequately.

1.2.1.3 Statistical analysis of animal-related outages in power distribution systems - a case study

This study presents a data-driven approach for exploring animal-related outages in power distribution systems. The main objectives are to uncover the underlying structure of animal-related outages, identify variables that strongly influence their frequency, and build a model to predict their occurrence on a weekly basis. To carry out this study, actual outage data obtained from a major utility company in the U.S., in addition to different types of wildlife and weather-related data are utilized. Results of this study could be very informative to utility companies in maximizing insight into their animal-related outage problems, and to build parsimonious models for predicting them.

1.2.2 Preventing avoidable outages by obtaining insights from the data

As mentioned earlier, in recent years, with the increasing requirements on power distribution utilities to ensure system reliability, utilities seek to find practical solutions that enable them to restrict specific outages or to better manage their responses to unavoidable power outages. For achieving either, it is crucial to acquire a profound understanding of different outages by exploring their underlying causes and identifying key variables related to those causes. Currently, statistical models, as well as advanced data analytics techniques, are common tools to gain such understanding. Although basic statistical analysis provides a general knowledge of the primary causes of outages; nevertheless, it falls short of describing nuanced conditions that lead to an outage. On the other hand, applying sophisticated algorithms can produce deeper insight into the main causes; however, it would be computationally burdensome and might require a tremendous amount of running time. In order to overcome these problems, a novel approach for outage root cause analysis by using association rule mining has been proposed. The brief description of the proposed approach is as follows.

1.2.2.1 Power distribution system outage root cause analysis by using association rule mining

The primary goals of this study are to characterize outages according to their underlying causes and to identify important variables that strongly impact outage frequency. This study proposes a step-by-step procedure, which deals with data preparation, practical issues associated with outage data sets, and implementation of association rule mining. The procedure is followed by a comprehensive case study to demonstrate how the proposed approach can be used to mine for causal structures and identify frequent patterns for vegetation, animal, equipment failure, public accident, and lightning-related outages.

1.2.3 Managing situation during outages

After an outage occurs, utility companies are to produce an appropriate response to make sure that the restoration process is fast and smooth and also the customers are well-informed about the situation. Identifying the cause of the outages could be considered as an important task for reaching the aforementioned objectives. As a result, a model has been developed to identify equipment failures which are one of the most common source of outage in power distribution systems.

1.2.3.1 Power distribution system equipment failure identification using machine learning algorithms

In this study, an approach for identifying equipment failure outages in distribution systems is explored. This task is considered as a binary classification problem in which outages are categorized into two classes of equipment failure and non-equipment failure types. To carry out this study, actual outage data collected by Duke Energy are utilized. First, different variables that make contributions to equipment failures are described and their relationships are examined. Afterward, the presence of imbalanced classes, as a common issue in outage data set, is addressed. Then, to assure that all features are relevant, their importance is examined by employing a novel feature selection algorithm. At the end, three classification algorithms, namely decision tree, logistic regression, and naive Bayesian classifier are trained and tested and their performances are evaluated.

1.3 Outage Statistics

The power systems outage data used in this dissertation was collected by Duke Energy - a major investor-owned utility company in the US. The data includes information on outages such as time, location, duration, impact, protective device, cause, action taken to restore the system, to mention a few. The outages occurred in the states of North Carolina and South Carolina between the years 2011 and 2014.

Before starting to build sophisticated models, a fresh look at some characteristics of the outages might be beneficial. Some interesting statistics about the outages are briefly discussed in what follows.

1.3.1 Cause of outages

Figure 1.2 shows the cause of different outages and their frequency (percentage). As seen in the figure, the cause for a majority of outages is vegetation. The second most frequent cause is labeled as "unknown", meaning that the crew could not find any evidence. The least frequent cause is the loss of transmission, which occurred only on very rare occasions.

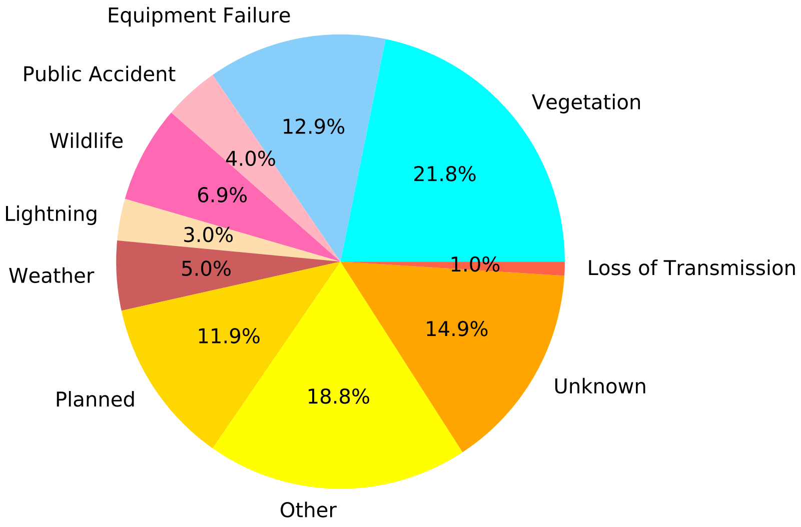


Figure 1.2: Frequency of outages for different causes

1.3.2 Protective device

Figure 1.3 shows different protective devices that were activated to clear the fault and their frequency (percentage). As seen, around 30% of the outages were cleared by line fuses. Transformer fuses were also activated on a considerable number of instances.

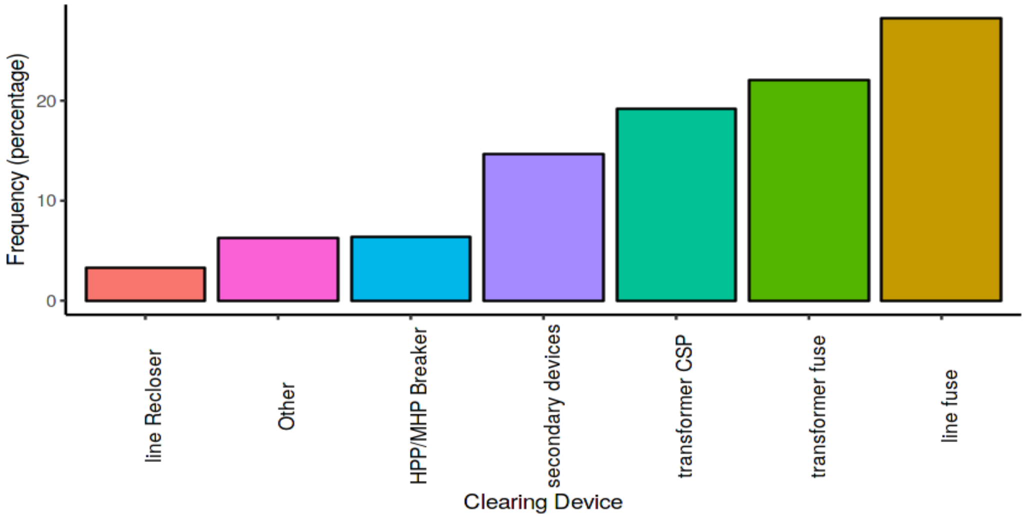


Figure 1.3: Frequency of clearing devices (percentage)

1.3.3 Duration of outages

Figure 1.4 shows the distribution of the duration of outages with respect to each cause. It is worth mentioning that there is a considerable number of outliers that exist in the duration variable (i.e., some outages would last for a few days). For the sake of illustration, these outliers are removed and only outages with a duration of fewer than 500 minutes are examined. As seen, the duration seems to differ for each cause. For planned outages, the duration is highly likely to be short. However, for weather-related events, the probability of having a long duration is greater compared to other causes (i.e., fatter tail).

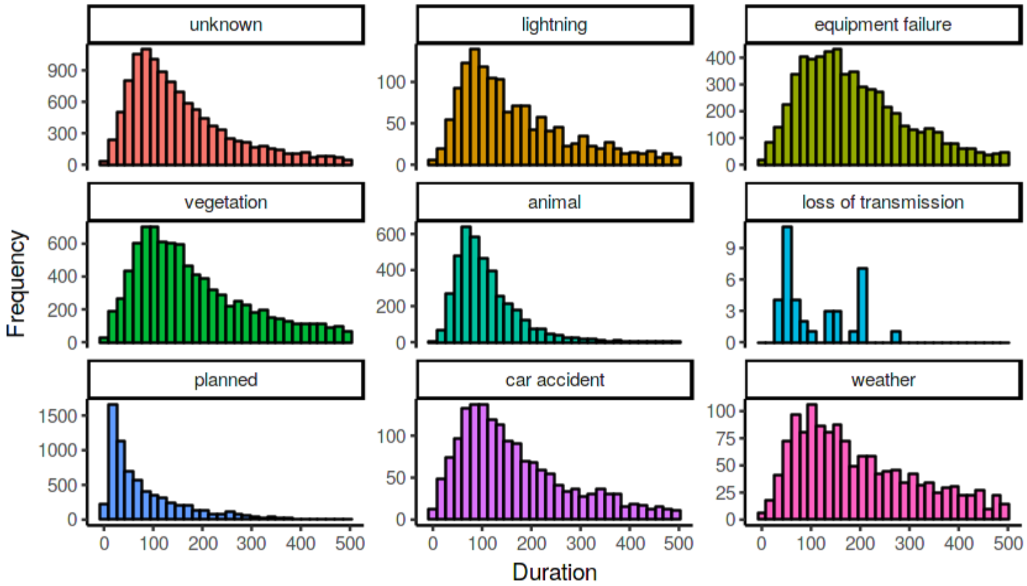


Figure 1.4: Distribution of outage duration for different causes

1.3.4 Month

Figure 1.5 shows the relationship between the cause of outages, month, and frequency of outages. The frequency is normalized, i.e. for each cause, summing the values over all months will lead to a value of 1. Based on the figure, it can be inferred that during summer, especially during the months of June and July, a peak on outages caused by different reasons occurs.

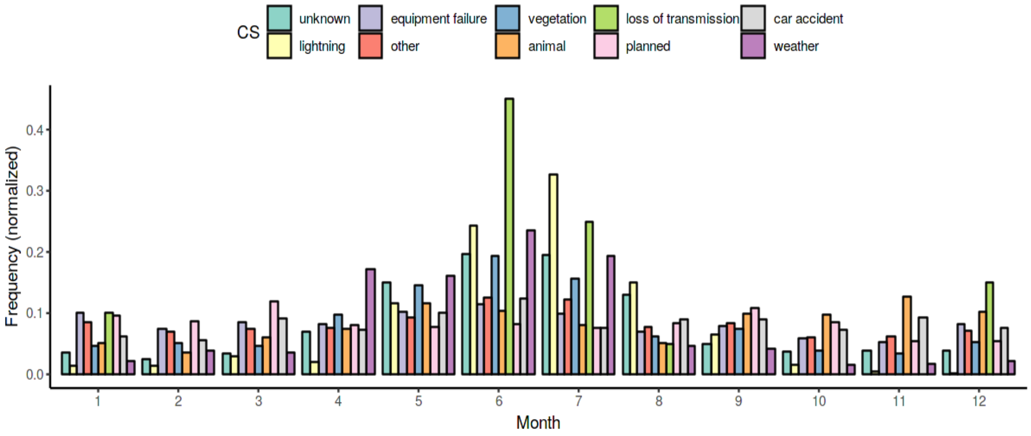


Figure 1.5: Frequency of outages on different months for different causes (normalized)

1.3.5 Hour

Figure 1.6 shows the relationship between the cause of outages, hour, and frequency of outages. The frequency is normalized, i.e. for each cause, summing the values over all hours will lead to a value of 1. Based on the figure, it can be inferred that during midnight, the frequency of outages decreases. Planned outages have mostly occurred during morning time between hours 8:00 am and 11:00 am. Lightning-related outages are mostly observed during evening time between hours 16:00 pm and 19:00 pm. Animal-related outages were observed during the morning time between hours of 7:00 am and 10:00 am.

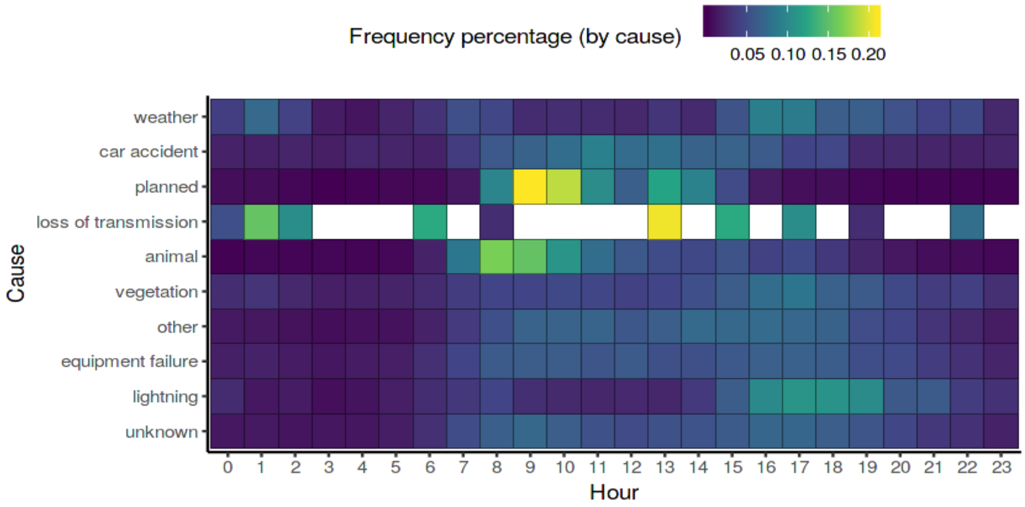


Figure 1.6: Frequency of outages on different hours for different causes (normalized)

1.3.6 Frequency of outages over time

Figure 1.7 shows the frequency of outages (aggregated on a monthly basis) over time for different causes.

The frequency is relative meaning that the values are divided by the first value that was recorded for each cause. For example, assume that during the first month available in the dataset, the number of vegetation-related outages was calculated to be 100. In this case, a relative frequency of 4 for a specific month means that the number

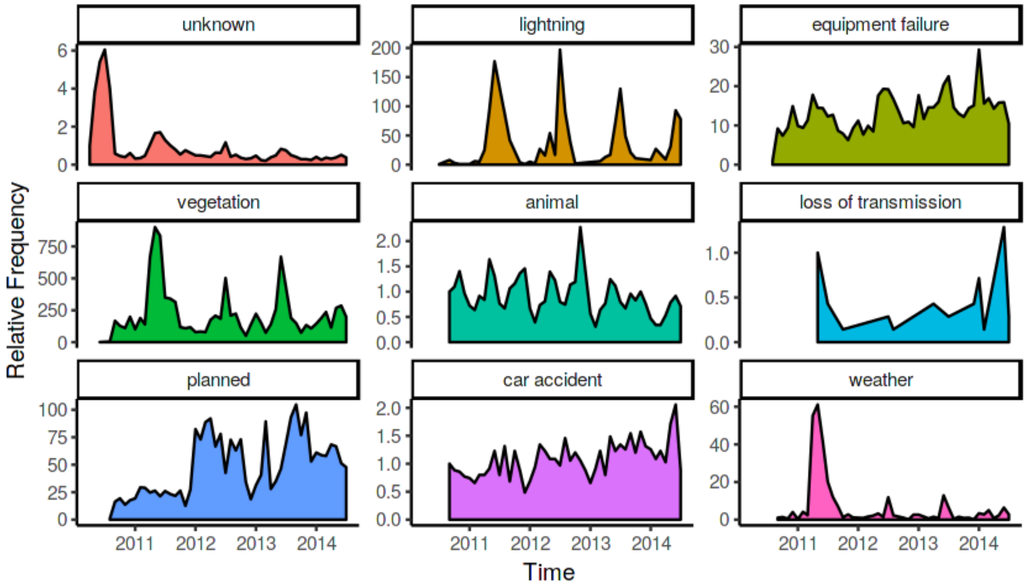


Figure 1.7: Relative frequency of outages over time for different causes

of outages was 400 in that month. These plots show the trend and seasonal behavior of the number of outages. As seen, the number of "unknown" outages decreases over time. A reason for that might be that the utility company was able to better find the cause of outages over the years. The trend for some causes is going down; however, for some causes including equipment failure and a car accident is increasing.

1.3.7 Frequency of outages based on location

Figure 1.8 shows the frequency of outages based on latitude and longitude. As seen, some areas in the system experienced a considerably higher number of outages. The substation located close to the city of Winston-Salem experienced around more than 1500 outages over the span of 4 years that the data was collected.

1.3.8 Counties with decreasing or increasing number of outages

An interesting question would be whether the outage frequency for a specific county remained the same or not during the span of years under study. For example, if the county of Greenwood is explored, did the outage frequency remain constant (or shows small variation) during the course of this time? In order to answer this question, the

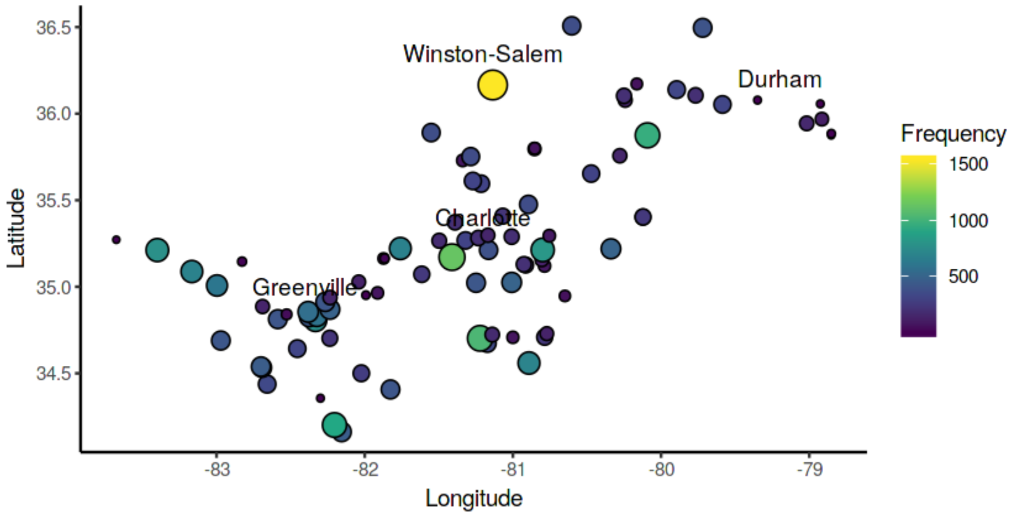


Figure 1.8: Frequency of outages based on location

data is grouped based on counties and months, and a simple linear model is fitted to each group (x would be month number ranging from 1 to 52, and y would be the number of outages). Then, the slope and intercept of the fitted lines are explored. Figure 1.9 shows the distribution of the slopes as well as the relationship between slopes and intercepts.

As seen, the slope values are mostly below zero. This shows that for the majority of counties, the number of outages decreased over the span of 4 years. The most extreme case was Greenwood county. On the other hand, there are multiple counties with the slope of zero or greater than zero. The number of outages remained constant or even increased during the span of 4 years under study for those counties. Cleveland county was one of those. Figure 1.10 shows the number of outages for the county of Greenwood (aggregated monthly). Figure 1.11 shows the number of outages for the county of Cleveland over time.

1.3.9 Long-duration outages

Duration of outages was shown for different causes in section 1.3.3. However, the duration was limited to 500 minutes. A critical issue for power utilities is, of course, long-lasting outages. In this dissertation, outages with a duration greater than 500

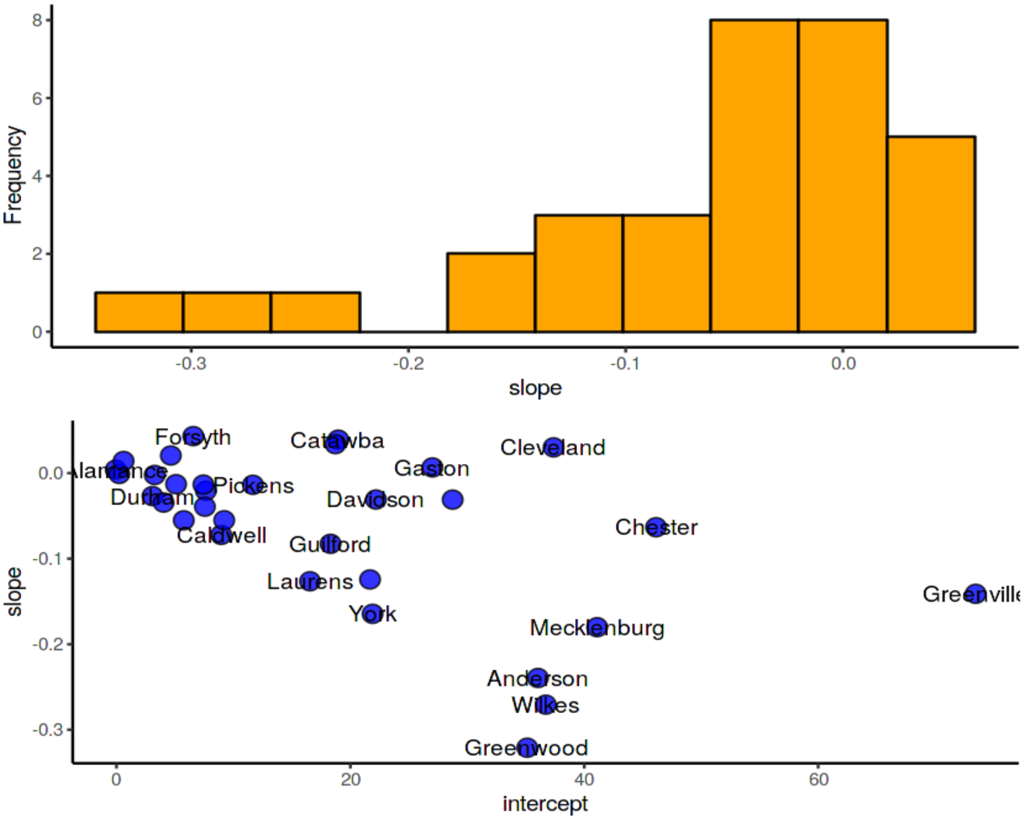


Figure 1.9: Statistical analysis of slope and intercept values for linear models fitted to different counties

minutes are considering to be long-lasting. A way to analyze this type of outages would be to investigate what percentage of outages for each cause belonged to the long duration variety during the course of different years. Figure 1.12 shows the aforementioned statistics.

According to the Figure 1.12, it can be inferred that for example, during year 2011, around 50% of weather-related outages resulted in long-lasting outages. This fact highlights how critical weather-related outages can be.

Another way that one can investigate the duration of outages (both long-lasting and short outages) is to explore the average duration for different causes over the course of time (in this case, weeks). Figure 1.13 demonstrates the aforementioned statistics.

It is worth mentioning that the duration is weekly average and is normalized (i.e.,

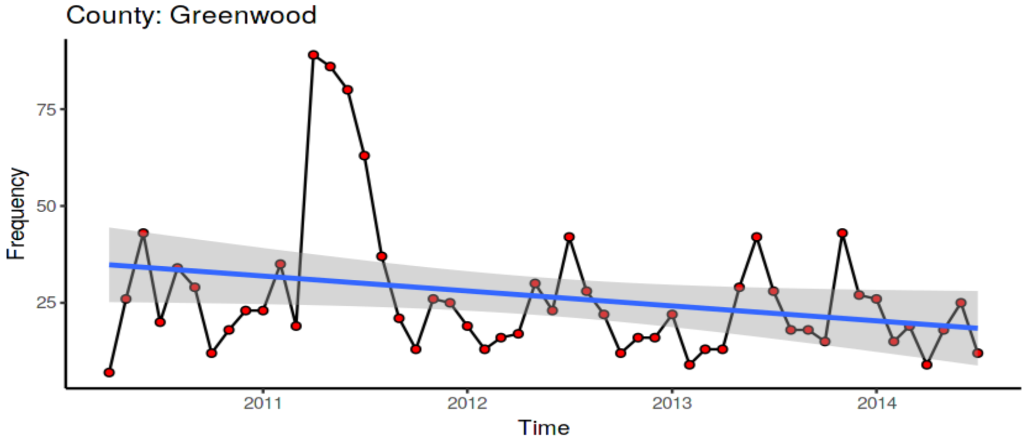


Figure 1.10: Frequency of outages for Greenwood county over time

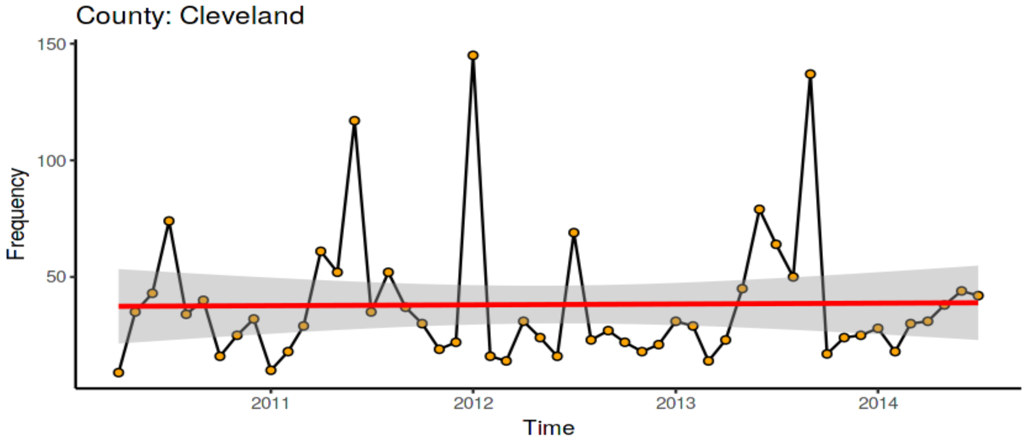


Figure 1.11: Frequency of outages for Cleveland county over time

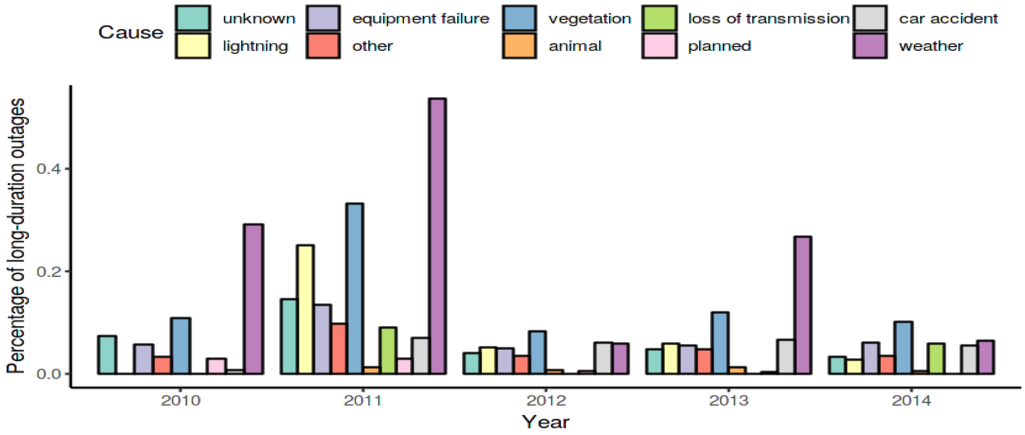


Figure 1.12: Percentage of outages resulted in long-duration outage

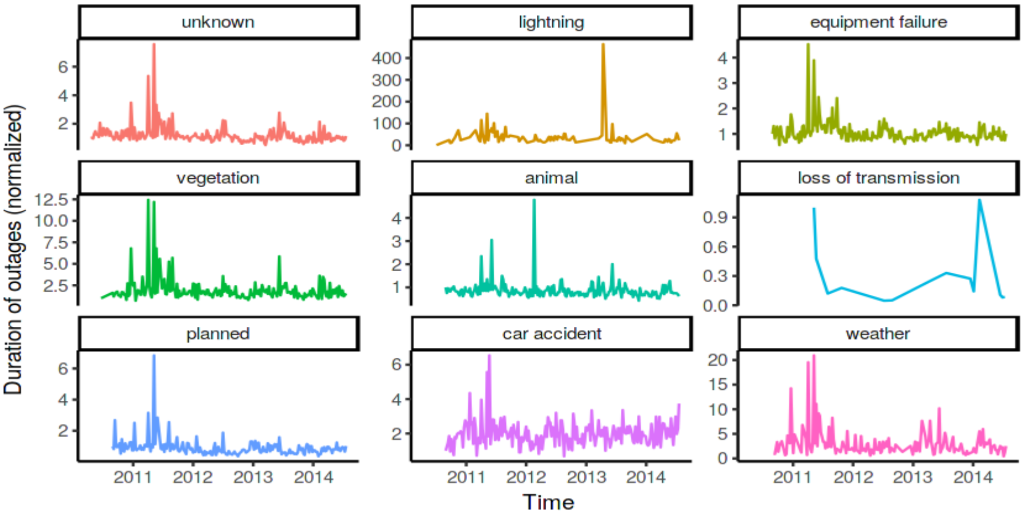


Figure 1.13: Normalized duration of outages over time (weekly average)

values are divided by the first value for each cause). An important observation from Figure 1.13 could be how long could a lightning-induced outage get. In fact, in a specific week, a lightning-induced outage lasted almost 450 times that a lightning-induced outage lasted in the first week in the dataset. Weather-related outages also frequently led to considerable duration. On the other hand, planned outages seem to last for a short duration except on a few rare occasions.

1.4 Organization

The rest of this dissertation is organized as follows. In Chapter 2, a data-driven approach for predicting vegetation-related outages in power distribution systems is proposed. In Chapter 3, a statistical approach is proposed for predicting lightning-induced outages. In Chapter 4, a statistical analysis of animal-related outages is presented and a predictive model for those is proposed. In Chapter 5, a novel approach by using association rule mining is developed for outage root cause analysis. Chapter 6 contains an approach for identifying equipment failure by using machine learning algorithms. Finally, in Chapter 7, conclusions, as well as recommendations for future works, are provided.

CHAPTER 2: A Data-Driven Approach for Predicting Vegetation-Related Outages in Power Distribution Systems

2.1 Introduction

2.1.1 Motivation

Many power utility companies recognize vegetation as the primary cause of outages in their power distribution systems. As a matter of fact, sustained or momentary outages caused by vegetation present serious challenges to maintaining adequate power quality and pose substantial risks to the reliability of the system [1]. In addition, vegetation growing near power lines can cause wildfires, creating major hazards to human and wildlife resources. In particular, over recent years, with more frequent extreme weather conditions and adverse natural phenomena caused by climatological factors, the risk of destructive interaction between vegetation and power lines has increased, making this source of outage a growing concern for power utilities [2, 3, 4].

In order to mitigate the damaging effects of vegetation on power systems, utilities are primarily interested in taking preventive measures. This includes designing more resilient distribution systems, conducting more frequent inspections, and scheduling regular tree-trimming operations [5]. Although taking these pro-active measures may reduce the potential risk of vegetation-related interactions with power systems, it is often impossible to eliminate vegetation-related outages under adverse weather conditions. Moreover, depending on the species, vegetation can sprout and grow very quickly, diminishing the impacts of trimming operations. As a result, in addition to implementing preventive measures, it is important for utilities to take appropriate responses in dealing with these outages, either by identifying them immediately after they occur or by predicting them in advance.

Electric utilities have shown significant interest in vegetation-related outage, and in general, outage cause identification, especially in the smart grid era. In fact, identifying the root cause of outages soon after they occur will enable utilities to accelerate the restoration process, ultimately improving the reliability of their systems [6]. This problem - specifically for vegetation-related outages - has received a great deal of attention, leading to the development of various models [6, 7, 8, 9]. On the other hand, outage prediction has been historically neglected. Although predicting outages, particularly of the vegetation-related category in advance, will bring enormous benefits to power companies, developing mathematical models for fulfilling this task is extremely challenging. This practical difficulty lies in the complex nature of vegetation-related outages and the fact that a substantial number of factors have a strong influence on the occurrence of these outages.

Nevertheless, in recent years, with the explosion of data gathering efforts within the smart grid framework and massive improvements in weather forecasting models, necessary data for studying vegetation-related outages has become available. Furthermore, mathematical methodologies are now being combined with advanced data analytics techniques, creating powerful tools that can improve utilities' predictive abilities on the aforementioned outages. Adopting these predictive approaches will enable effective and timely decision-making actions by operators as well as planners, ultimately improving operational integrity and resiliency [10].

2.1.2 Literature Review

Several studies have aimed at exploring the underlying causes of vegetation-related outages that occur in power distribution systems and developing analytical approaches to predict some characteristics of these outages.

The authors in [11] propose an approach for predicting the rate (number of outages per mile-year) of vegetation-related outages that are caused due to vegetation growth on an annual basis. They develop and evaluate four different models, vis-$\grave{a}$-vis, linear regression, exponential regression, linear multivariate regression, and artificial neural network. The main inputs to their models are historical outage data and climatic variables that affect the vegetation growth.

Another study [12] proposes a statistical approach to predict vegetation-related outages under normal (non-storm) operating conditions. In particular, the authors carry out an investigation on the impact of tree trimming on the frequency of vegetation-related outages. They utilize historical outage, geographical, and tree-trimming data in their approach. The data is fed into three statistical models, namely Poisson generalized linear model, negative binomial generalized linear model, and Poisson generalized linear mixed model, and the performance of each model is evaluated.

In [13] the authors present a study to evaluate the impact of various factors on predicting vegetation-related outages under hurricane conditions. In particular, they use LiDAR data to derive variables that can model the height and location of trees in the system under study. By utilizing the aforementioned data along with vegetation management and system infrastructure information, they develop an ensemble machine learning algorithm to predict whether or not a specific area will experience vegetation-related outages under an extraordinary weather condition.

In [14], the authors develop a data-driven approach to conduct a comprehensive root cause analysis of outages that occur in distribution systems. Their primary goals are to characterize outages according to their underlying causes and to identify important variables that strongly impact the outage frequency. By applying the proposed approach on vegetation-related outages, they demonstrate the importance of climatological and geographical factors for predicting these outages.

Despite the fact that the aforementioned studies attempt to predict vegetation-induced outage rate or its characteristics (albeit some of them take advantage of advanced data gathering tools and machine learning methods), an approach that has the potential for implementation to predict the anticipated number of these outages on a monthly basis for a specific location, has not yet been developed. In fact, the existing approaches are either not designed for this purpose, or have practical shortcomings. These shortcomings include delivering a low degree of accuracy when the number of outages is large, and lacking the ability to make the prediction for a specific location in the system within a short-term horizon. Moreover, existing models merely focus on one cause of vegetation-related outages and therefore, fail to provide a comprehensive approach that takes various sources of such outages into consideration. It is evident that the main reason for the perceived lack of studies by the research community on this critical problem is the limited access to a sufficient amount of outage data. In fact, a majority of utility companies do not publish their outage data in great detail, and therefore this problem has not been explored to the fullest extent yet.

2.1.3 Contributions

In this chapter, a novel approach to predict the number of vegetation-related outages in power distribution systems on a monthly basis is proposed. Utilizing statistical and machine learning predictive models, the proposed approach is an intelligent solution that combines weather and geographical data with past vegetation-related outage information and makes a prediction for the anticipated number of future outages.

The proposed approach provides a meaningful knowledge about risks and locations of vegetation-related outage problems. This presents a succinct view of the current system status to the operators, which enables effective and timely decision-making actions with regards to vegetation-related problems. By providing a preliminary but accurate prediction, the proposed approach allows operators to take high-resolution imagery of areas with high risk of an outage, or utilize LiDAR data, or dispatch a crew to find the exact locations in the system that a vegetation-related outage could occur.

What distinguishes this approach from the other studies in the literature may be summarized in the following four statements:

- The proposed approach takes different characteristics of vegetation-related outages into consideration, and successfully categorizes them according to their underlying causes. Moreover, the approach provides a specialized predictive model for each category.

- The approach offers a workable solution to address the challenges brought about by the extent of the prediction's target area.

- The approach takes advantage of a considerable amount of vegetation-related outage data that is provided by a major utility company.

- All the advantages of the proposed approach are built upon generic outage data collected by utilities (if available), and typical daily weather forecast data, which is publicly available. This fact makes the implementation of the approach easily attainable within a reasonable level of accuracy.

- Several considerations taken in the formulation of the proposed approach enable its deployment to be highly flexible to a variety of different settings and objectives such as prediction horizon.

2.1.4 Study Organization

The rest of this chapter is organized as follows. In Section 2.2, a detailed problem description is provided where the objectives and approach are explained in details. In Section 2.3, the data is described and a discussion of the data pre-processing procedure is provided. In Section 2.4, the proposed data-driven methodology is explained. In Section 2.5, a comprehensive case study is carried out, and the results are presented and discussed. Finally, in Section 2.6, conclusions, as well as recommendations, are provided.

2.2 Problem Description

In this chapter, a data-driven approach is proposed to predict the anticipated number of sustained vegetation-related outages for a particular month (preferably the month ahead) in a given area within a power distribution system. In order to develop a practical approach for this purpose, two major challenges have to be addressed comprehensively.

The first challenge is to define the extent of the prediction's target area. In fact, if one's intent is to point-predict the number of vegetation-related outages at the location of any given substation, one might encounter serious difficulties. These difficulties lie in the fact that the degree of randomness for the number of outages that occur for each substation is relatively large. Moreover, accurate radar weather forecasts may not be available at the exact location of each substation as a result of the weather station being far from the substation. Consequently, developing an approach that is capable of predicting the number of vegetation-related outages at an exact given location in the system is neither realistic nor practical.

To the best of the author's knowledge, this problem has not been investigated in detail yet. As mentioned before, one reason could be the limited access to a sufficient amount of data. However, thanks to the considerable amount of data provided by a large power company in the southeastern US, this problem could be addressed in this study. In order to deal with this issue, the proposal is to aggregate substations and to build larger areas, in which each area includes multiple substations and local weather stations. As a result of doing so, the randomness in the number of outages would be harnessed considerably, leading to greater predictive capability. Moreover, carrying out this task will produce more comprehensive weather forecast since for each area multiple weather stations will be considered, making sure that at least one of the stations captures the major weather events.

The second challenge is to thoroughly investigate the nature of vegetation-related outages. In fact, vegetation-related outages occur due to various reasons; however, they could be categorized into two coherent groups. The first group includes outages that are caused by vegetation that naturally grow into and make contact with distribution lines. The second group consists of those ones that are caused by vegetation falling into or making contact with distribution lines due to weather-related factors.

Vegetation-related outages that are caused by the natural growth of vegetation mainly depend on the time of the year and manifest a strong seasonal pattern. Moreover, their occurrence is highly affected by vegetation management operations such as regular trimming. On the other hand, outages that are caused by weather-related factors do not show an apparent trend. Although time could play a role in the frequency of such outages, there are various other climatological and geographical factors that make strong contributions to the occurrence of such outages.

In this study, in order to fully consider the aforementioned characteristics of vegetation-related outages, developing two different models for the prediction purposes is proposed: (i) using a statistical approach based on time series algorithms to predict the outages that fall into the first group, and (ii) employing machine learning based approach for the second group. This separation will significantly increase the predictive capability.

The technical flow-chart of the proposed approach is depicted in Fig. 2.1. Each block of the flow-chart will be discussed in the following sections.

2.3 Data description and pre-processing

2.3.1 Data description

As mentioned, the input data for the proposed approach are obtained from three main sources: 1) historically recorded outages, 2) geographical information, and 3) radar weather forecasts. The power systems outage data is collected by Duke Energy - a major investor-owned utility company in the US. The data includes information on the exact time and responsible substations of sustained vegetation-related outages

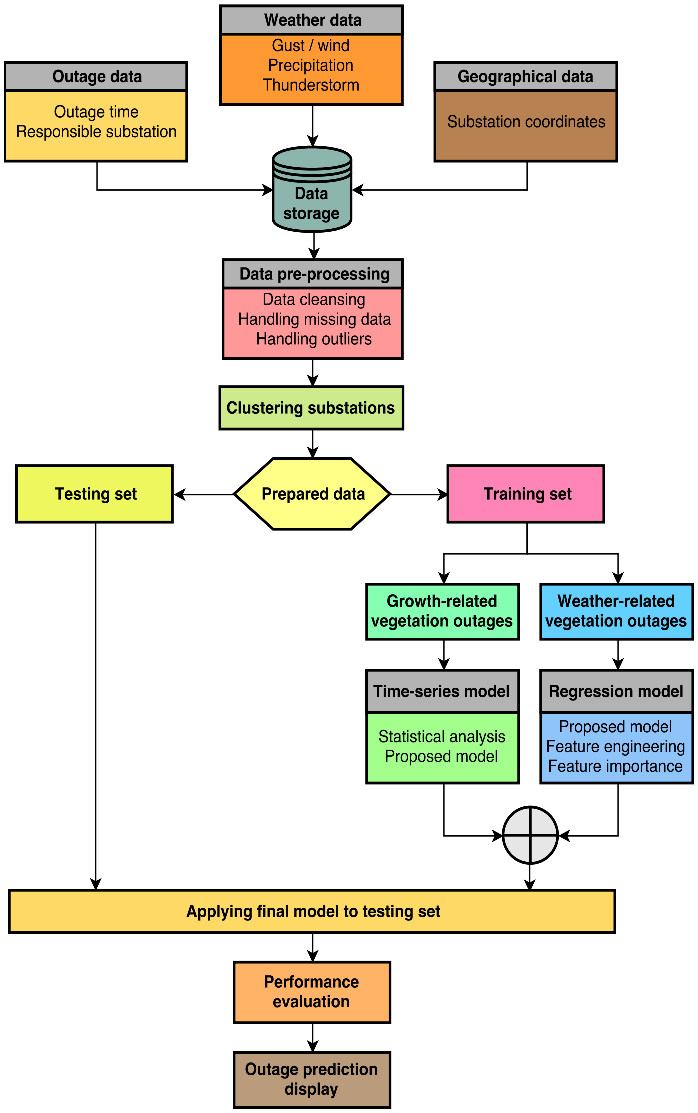


Figure 2.1: Technical flow-chart of the proposed approach

that occurred around approximately 85 substations located in the states of North Carolina and South Carolina between the years 2011 and 2014. It is worth noting that the models for the proposed approach are developed based on the data gathered in years 2011 to 2013. The year 2014 will be used as a case study to show the performance of the proposed approach. The geographical data is also provided by Duke Energy, and contains information about the exact location of different substations. The hourly radar weather data are collected from several external sources for all substations over the span of the aforementioned years. The data will be explained in more details in the upcoming sections.

2.3.2 Data pre-processing

Real-world outage, weather, and geographical datasets usually contain various input errors, duplicate data, extreme and unexpected values, missing values, and outliers. These types of data can result in misleading representations and interpretations of the collected data and may skew the operation of data-driven models. Therefore, the raw data should be explored and appropriate actions ought to be taken in order to avoid potential problems. In this study, after the data is cleansed, two major data pre-processing tasks, i.e. handling missing values and outliers are performed and discussed below. Providing this discussion is important as the data pre-processing task could have a considerable impact on the performance of models that will be developed later on.

2.3.2.1 Dealing with Missing Values

There is a relatively small number of missing values in the weather dataset. For example, on some occasions, the information about gust intensity - one of the main inputs to the proposed regression model, for a specific area at a particular timestamp is missing. In order to properly handle this issue, the missing data is filled in by drawing values from non-missing values utilizing a linear interpolation technique. The reason behind selecting this method is that the nature of the variables that contain missing value is such that they can reasonably be expected to change linearly over short time intervals. Therefore, interpolation should be a rational approach.

2.3.2.2 Dealing with Outliers

Outliers, being the extreme values that deviate markedly from the other observations, could be present in real-world datasets. As a matter of fact, the outage datasets used in this work includes a small number of samples that show an unexpectedly large number of outages recorded for a specific area at a particular time-stamp. Moreover, the weather dataset contains some disproportionately extreme values with regard to weather-related factors.

To deal with the existing outliers in both outage and weather datasets, first, the outliers are detected by using the Inter-Quartile Range measure, and then removed them from the dataset. It is worth mentioning that there are various other strategies to detect and accordingly handle the outliers [15], [16]. Selecting the most effective strategy is not a straightforward task, and depends on the situation and the dataset. In this study, the presence of the outliers is mainly due to incorrectly entered or measured data. Moreover, different analyses are performed and it was concluded that the selected approach is the most appropriate strategy for this study. In particular, taking the aforementioned action does not make any significant impact on the distribution of the variables.

2.4 Proposed data-driven methodology

2.4.1 Aggregating substations and creating different areas

As mentioned, one of the main challenges with developing the proposed approach is defining the extent of prediction target area. To deal with this issue, the proposal is to aggregate substations and to build larger areas, in which, each area includes multiple substations and local weather stations, where the weather forecast for each substation is obtained from its closest weather station. In order to define the aforementioned areas, k-means clustering algorithm is utilized. This algorithm is a widely used unsupervised machine learning algorithm, which aims at categorizing a given dataset into a certain number (k) of clusters. In this study, this algorithm is used to group substations into different clusters, where each cluster represents an area. The main idea of this algorithm is to define k centroids at random, one for each cluster, and then to minimize the squared error function represented in (3.1) [17].

$$J(r, \mu) := \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{k} r_{ij} \left\| x_i - \mu_j \right\|^2 \tag{2.1}$$

where m is the number data points, k is the number of clusters, r_{ij} is an indicator, which is 1 if, and only if, x_i is assigned to cluster j, x_i is data point, μ_j is the centroid for cluster j, and $\|.\|^2$ denotes the Euclidean distance. In this study, data points are locations of substations (approximately 85 data points), which are represented by latitude and longitude in a two-dimensional space.

One major challenge with this algorithm is to specify the number of clusters. In fact, there is no global theoretical method to find the optimal value of this parameter; however, a few approaches are common among data scientists for dealing with this problem. One workable approach is to run k-means clustering for a range of different k values and to calculate the aforementioned squared error function for each value. In this case, the error tends to decrease toward zero as k increases; however, after a certain k value, the decrease would be very gradual. Therefore, analyzing different values of k and finding the aforementioned threshold could help on deciding a reasonable number of clusters [18].

Applying k-means clustering algorithm and using the aforementioned method to calculate an appropriate number of clusters for grouping substations will result in 14 areas for this study. In order to solve the k-means clustering problem, the Lloydâs algorithm is utilized in this study. Fig. 2.2, illustrates these geographically dispersed areas.

It is worth mentioning that although there are other clustering approaches, the approach adopted here is the most suitable for this study. In fact, since the goal is to cluster a two-dimensional data (i.e., latitude and longitude) and to work with distances, the k-means approach makes the most sense. Moreover, considering the size of clustering data which is small, there would be no need for utilizing more sophisticated algorithms.

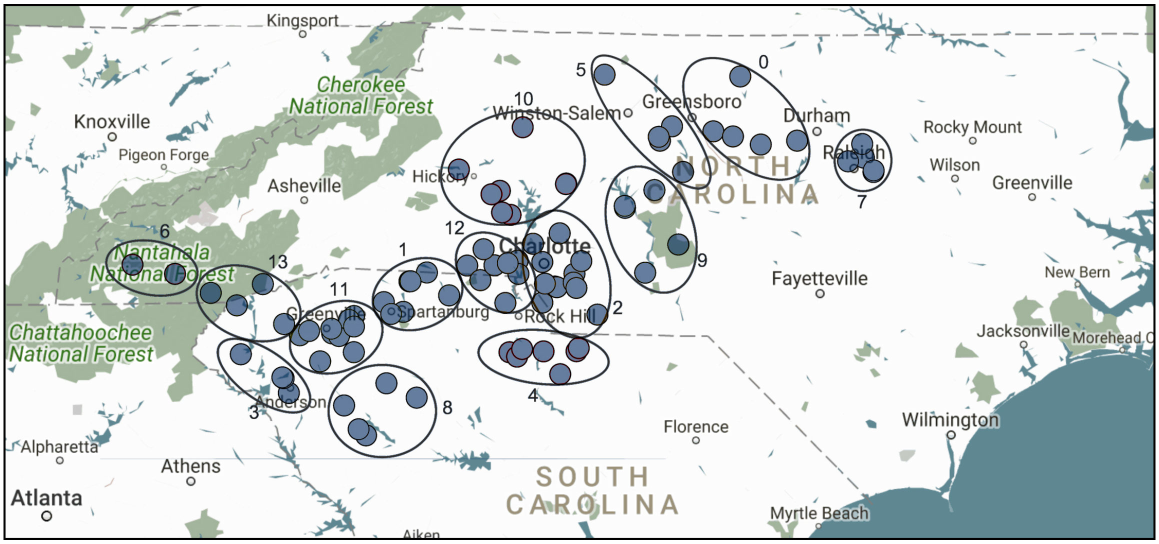


Figure 2.2: Demonstration of different areas

2.4.2 Proposed model for predicting growth-related vegetation outages

As discussed earlier, a category of vegetation-related outages is comprised of those events that are caused by vegetation naturally growing into, and making contact with distribution lines. This type of outage will be called growth-related vegetation outage from this point. In order to identify this type of outage, weather and outage datasets were analyzied carefully, and those outages for which no weather-event was recorded by any weather station within the stipulated area for a few hours prior to the occurrence of the outages were selected. The fact that no weather-related event was recorded for those outages ensures that such outages have occurred due to natural reasons. It is also worth noting that gaining access to a sufficient amount of vegetation data could further help in identifying this type of vegetation-related outages.

The main objective of this part is to propose a type of model that is able to successfully predict the growth-related vegetation outages for a specific area in the system during a specific month. With this context, the target for prediction is defined as Growth-related Vegetation Outage Count Index (GVOCI), which for area a and month m can be formulated as in (2.2).

$$GVOCI_{am} = \sum_{s \in a} \sum_{d \in m} \sum_{h=1}^{H} GVOC_{sdh} \tag{2.2}$$

where s, d, and h represent the substation, day in month, and hour in the day, respectively. The $GVOC$ is hourly growth-related outage count and belongs to $GVOC \in 0, 1, 2, 3, ...$ These parameters will allow the utility to have the option to predict the number of outages that occur during specific hours in the day, specific days in the month, and for specific substations. In order to predict all potential outages, these values should encompass all substations in the target area, 28-31 days (depending on the month) and 24 hours. Fig. 5.3, top plot, demonstrates the $GVOCI_{am}$ aggregated over all areas for all months starting from 2011 to the beginning of 2014 represented in a series.

To clearly describe the patterns in growth-related outage series, the data can be decomposed into three main components, namely trend, seasonality, and residual. These components explain the large-scale increase or decrease, the variations that periodically repeat, and the variations that is random in the series, respectively. Fig. 5.3 illustrates these components.

As observed in figure, the yearly trend shows that the number of outages has decreased. This phenomenon may be mainly due to performing regular tree-trimming operations. It should be noted that tree-trimming operations will result in a smaller number of vegetation-related outages that are caused by natural reasons. The model that will be employed for predicting this type of vegetation-related outages is able to capture and adopt this pattern for predicting the number of future outages. Also, it should be noted that depending on the species, vegetation can sprout and grow quickly, diminishing the impacts of trimming operations. Moreover, based on the figure, it can be observed that, on average, the number of such outages in the summer months, especially in the months of June and July, is much higher than during other months. By applying further statistical analyses, it is understood that even though

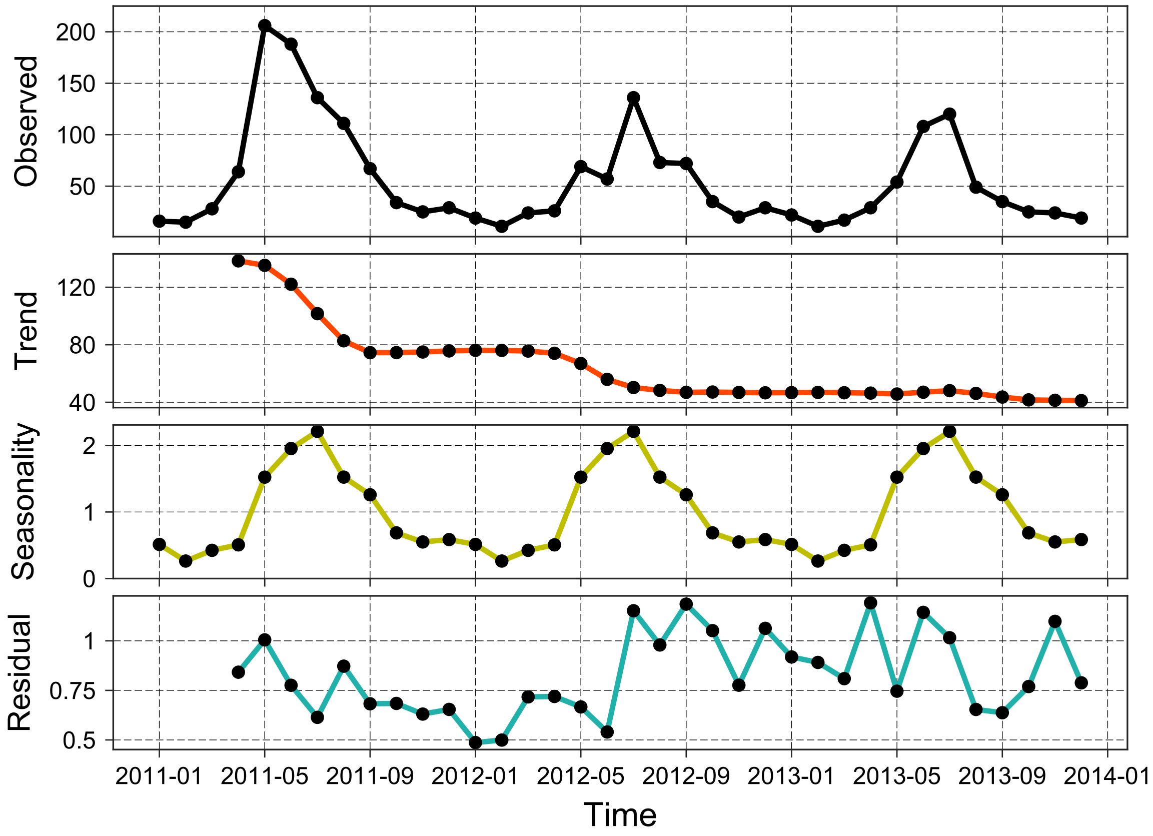


Figure 2.3: Decomposition of growth-related vegetation outages

the number of outages occurring in each month is considerably different, their variance is relatively small, suggesting that there is a strong seasonal effect.

Since such outages show obvious trend and seasonal structures over time and it seems highly unlikely that different outside variables insert influence on these outages, the proposal is to utilize a time series analysis to identify their nature and to predict their future values. In fact, time series analysis is a well-established statistical analysis and due to the aforementioned reasons, appears to be the most sensible approach.

There are various models to analyze time series; however, the selection of the most appropriate model mainly depends on the type of data and application. In the case study section that follows, two well-established time series models are explored and utilized to demonstrate how growth-related vegetation outages can be predicted.

2.4.3 Proposed model for predicting weather-related vegetation outages

2.4.3.1 Proposed type of the model

Another type of vegetation-related outages is that results from vegetation contacting or damaging power distribution system infrastructure due to severe weather-related factors. For simplicity, these types of outages will be called weather-related vegetation outages from this point. In order to identify these outages, weather and outage datasets were explored, and those outages for which a major weather-event was recorded by any weather stations within the stipulated area during the time of their occurrence, and a few hours prior to that were selected. The reason for considering a margin of few hours is to account for potential errors in weather forecasts.

As opposed to the growth-related vegetation outages, weather-related vegetation outages do not show a clear trend over time. Although, as will be shown, there could be a correlation between time-related factors and the frequency of such outages, the occurrence of these outages is more dependent on climatological and geographical factors. As a result, the problem of predicting the future values of such outages cannot be simply defined as a time series problem.

Therefore, to successfully carry out the prediction task for weather-related vegetation outages, the proposal is to define the problem as a supervised machine learning problem and to develop regression models that can handle the non-linearity in the data. One crucial point to argue with regards to the type of regression models is that linear models do not deliver a satisfactory performance for this problem. As a matter of fact, linear regression models assume a linear relationship between the input variables and the output variable. More specifically, such models assume that the output can be calculated from a linear combination of the input variables. However, various statistical analyses have been conducted (i.e., fitting linear models and analyzing residuals and checking linear model assumptions) and it has been concluded that linear models are not suitable for this problem. Therefore, utilizing non-linear regression models will be proposed. It will be demonstrated that how well-established non-linear models perform through a comprehensive case study.

In order to develop highly capable data-driven models for predicting weather-related vegetation outages, an essential step is to gather as much information as possible about the factors that influence the occurrence of these outages. These factors can be categorized into three main groups of climatological, geographical, and time-related variables. By utilizing these variables, different aspects of aforementioned outages including the severity of weather events, the characteristics of the distribution system infrastructure, the vegetation density, and the potential correlation between time and extreme weather conditions can be explained.

After the necessary data is collected, a comprehensive process needs to be carried out to transform the raw data into the features that accurately describe the inherent structures within the data. This process, which is known as feature engineering, will result in significant improvement in the accuracy of the model particularly when the scale and size of data are not considerably large. In what follows, the data that was collected is explained, the target variable is defined, the feature engineering task is performed, resultant features are explained, and their importance is evaluated.

2.4.3.2 Defining target variable and conducting feature engineering

As explained, the main objective in this part is to predict the number of weather-related vegetation outages that occur in a specific area within the system on a particular month. Similar to $GVOCI_{am}$, an index can be created for weather-related vegetation outages to represent the target variable. This index is defined as the Weather-related Vegetation Outage Count Index ($WVOCI$). The mathematical formulation of the $WVOCI$ for area a and month m is expressed in (2.3).

$$WVOCI_{am} = \sum_{s\in a} \sum_{d\in m} \sum_{h=1}^{H} WVOC_{sdh} \tag{2.3}$$

Where s, d, and h represent the substation, day in month, and hour in the day, respectively. The $WVOC$ is hourly weather-related vegetation outage count and belongs to $WVOC \in 0, 1, 2, 3, ...$

In order to successfully predict the $WVOCI_{am}$, various types of information, namely weather, geographical, and time-related factors are gathered and processed. With regards to the weather-related factors, as frequently reported, the occurrence of weather-related vegetation outages is highly influenced by the frequency of gust (extremely windy) conditions, the intensity of the gust events ($mile/hour$), and the occurrence of heavy precipitation and thunderstorm conditions. Therefore, in order to include these factors into the model and represent them, necessary hourly weather data are collected and three indices of Gust Count Index (GCI), Gust Intensity Index (GII), and Storm Count Index (SCI) are created. The mathematical formulation of these indices for area a and month m is presented in (2.4) to (2.6).

$$GCI_{am} = \frac{1}{S_a} \sum_{s \in a} \sum_{d \in m} \sum_{h=1}^{H} GC_{sdh} \tag{2.4}$$

$$GII_{am} = \frac{1}{S_a} \frac{1}{D_m} \frac{1}{H} \sum_{s \in a} \sum_{d \in m} \sum_{h=1}^{H} GI_{sdh} \tag{2.5}$$

$$SCI_{am} = \frac{1}{S_a} \sum_{s \in a} \sum_{d \in m} \sum_{h=1}^{H} SC_{sdh} \tag{2.6}$$

where GC, GI, and SC demonstrate the hourly number of gust events, gust speed, and number of storm events and satisfy the following conditions: $GC, SC \in 0, 1$, and $GI > 8$mph.

Moreover, S_a represents the number of substations in the area, D_m shows the number of days that are investigated in the target month, and H demonstrates the number of hours that are considered for each day. With regards to the aforementioned indices, certain factors need further clarification:

- The windy condition is considered as gust when the speed exceeds the value of 8 *mph*. This threshold is selected based on [19], where the authors discuss the impact of different wind speed values on vegetation. It is necessary to mention that the value selected in this study is slightly smaller than the minimum value presented in [19] to make sure that no gust event is missed. If the weather forecast shows any gust event, the GC will become one; otherwise, zero.

- In order to investigate the impact of gust intensity, different variables such as average, maximum, and minimum of gust speed were created. After a careful analysis, it was observed that these variables show a considerable correlation with each other. Also, the average gust speed showed a stronger relationship with the target variable. Hence, only the average variable, represented by GII_{am}, is utilized in this study.

- The storm refers to both heavy precipitation conditions and thunderstorms. Almost all publicly available weather forecast sources report these conditions on an hourly basis. If the weather forecast shows any of these event, the SC will become one; otherwise, zero.

- Since obtaining accurate values for the intensity of precipitation and thunderstorms is difficult, it is decided not to include such information in the model; however, if accurate values were available, their inclusion may be useful.

- For the specific distribution system under investigation, snow is not a common weather-related phenomenon because of its geographical location; hence, no factors related to snow was considered in this study. However, in areas with heavy snowfall potential, it is highly recommended to include a variable to describe the count and intensity of the snow as an input.

In order to demonstrate the relationship between the proposed variables and the $WVOCI_{am}$, these variables are evenly categorized into eight different categories, and the relative frequency of the $WVOCI_{am}$ is calculated for each category and plotted. The selection of the number of categories depends on the analyst's preference. Fig. 2.4 illustrates such relative frequency for these variables.

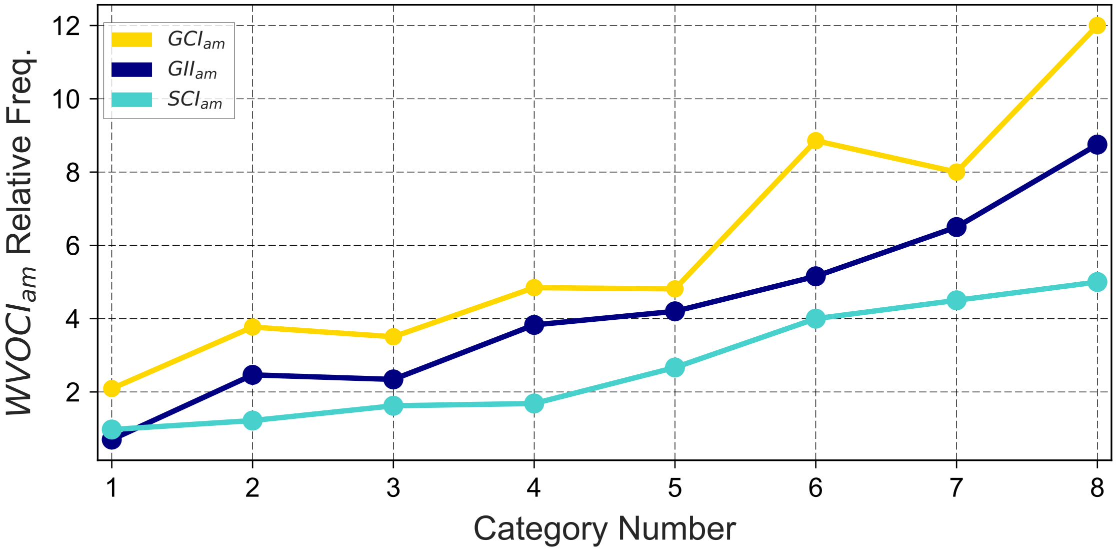


Figure 2.4: Weather-related factor's impact

According to the figure, it can be realized that as the value for the proposed variables increases, the $WVOCI_{am}$ increases. This observation is not surprising because it is expected as the number and intensity of gust and storm events increases, more vegetation-related outages occur. However, this observation confirms that the proposed variables can successfully explain the variability of the target value and capture the statistics of the data.

The second types of variables that have significant impacts on weather-related vegetation outages are geographical variables. In fact, these variables can describe the infrastructure of the distribution system under investigation, how prone the system is to the outage, how much interaction the system has with the vegetation, etc. These factors differ from area to area. A major challenge with these factors is that obtaining information about them could be extremely difficult. As a matter of fact, the data that contain such information typically is either not available or is kept confidential.

Hence, to address this problem, and to consider the effect of geographical factors, the proposal is to create a variable based on historical outages and weather-related factors to explain the geographical information. The proposed variable is defined as Area Outage Index (AOI) which, for area a, is formulated as in (2.7).

$$AOI_a = \frac{\sum_{y=1}^{Y}\sum_{m=1}^{M} WVOCI_{yam}}{\sum_{y=1}^{Y}\sum_{m=1}^{M}(GCI_{yam} + SCI_{yam})} \tag{2.7}$$

where Y represents the number of years for which outage and weather data are investigated (i.e. 3 in this study). The AOI shows the ratio of the number of all weather-related vegetation outages that occurred in an area over the number of extreme weather conditions that area experienced.

Fig. 2.5 is provided to better illustrate the AOI variable. As seen in the figure, bottom plot, during years 2011 to 2013, each area experienced a different number of gust and storm conditions. It turned out that the average gust intensity for all areas is almost the same. In the top plot of Fig. 5, the green bars show the number of

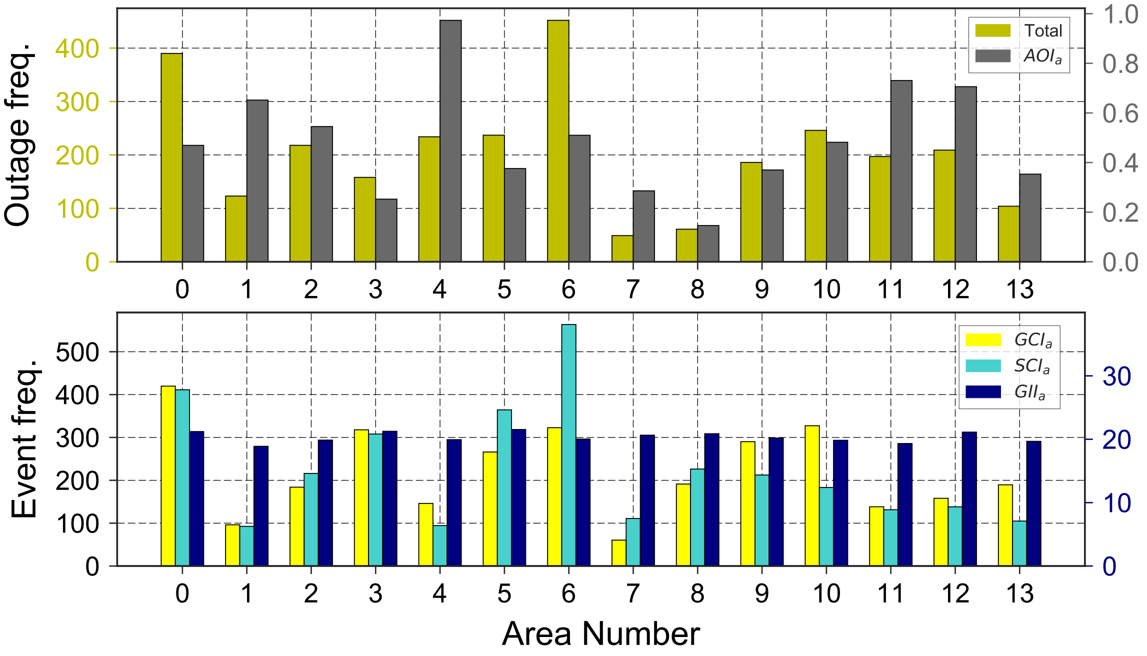


Figure 2.5: AOI_a representation

outages that each area experienced during the same time span. It can be understood that although some areas experienced a smaller number of weather-related events, the occurrence of outages was more frequent for them. Therefore, the geographical characteristics of these areas are such that they are more prone to outage. The gray bar plots representing the AOI variable clearly demonstrate this effect. Areas with large AOI value, have high potential for weather-related vegetation outages. Besides being of use as an input, this variable can inform utility asset owners about the vulnerable zones in their systems. It is worth noting that in case that any geographical information is available, it is highly recommended to include it into the model; however, the proposed variable can successfully capture the geographical statistics in the data.

Finally, as mentioned before, there is a correlation between time-related factors and the $WVOCI_{am}$. In fact, since the time of the season has a significant effect on vegetation-growth and density, it can affect the amount of influence weather-related factors can have on the vegetation. In order to capture this correlation, a variable for each month is created. This variable is called MOI which, for month m, is formulated as in (2.8).

$$MOI_m = \frac{\sum_{y=1}^{Y}\sum_{a=1}^{A} WVOCI_{yam}}{\sum_{y=1}^{Y}\sum_{a=1}^{A}(GCI_{yam} + SCI_{yam})} \tag{2.8}$$

This variable shows the ratio of the number of all weather-related vegetation outages that occurred in each month over the number of extreme weather conditions that month experienced over the span of all available years. Fig. 2.6 shows the relationship between month, weather-related factors, and outage.

As seen in the figure, bottom plot, winter months experienced a larger number of weather-related events; however, the occurrence of weather-related vegetation outages (top plot, pink bars) is not as frequent as in the summer months. This can be explained by the lower density of vegetation during these months. In fact, the MOI

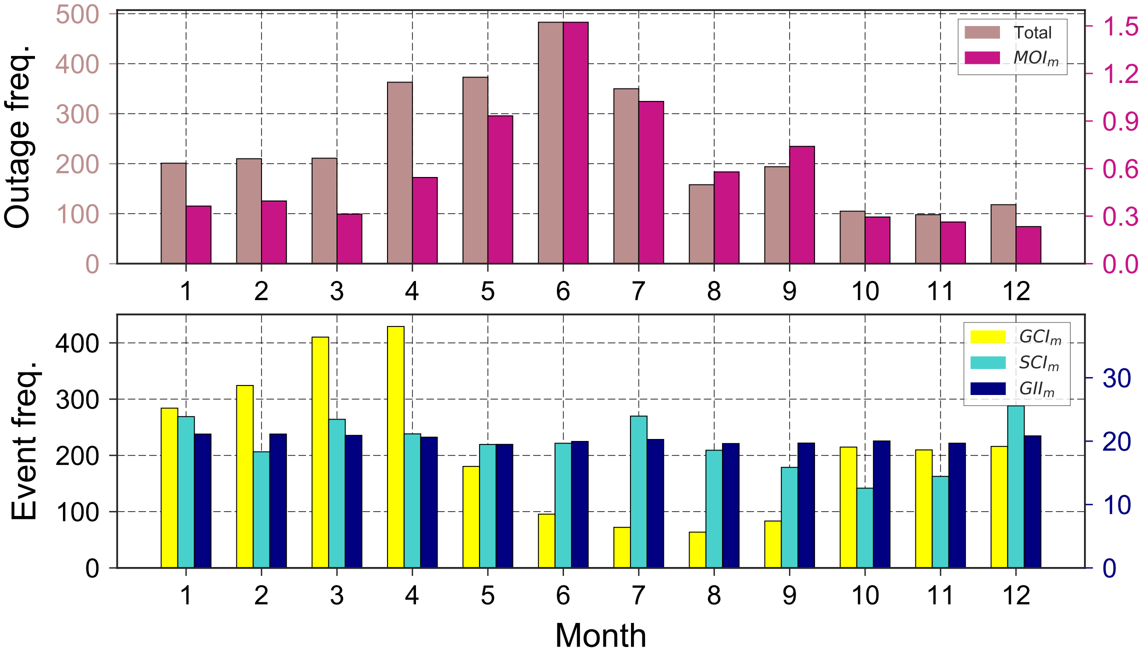


Figure 2.6: MOI_m representation

variable, which is shown in the top plot with purple bars, demonstrates that the summer months, especially June, are more critical since vegetation density reaches its maximum and therefore weather-related factors are more influential. It is worth mentioning here that the data suggest the average gust intensity is almost the same in different months; hence, this variable is not included in the formulation of MOI.

2.4.3.3 Conducting feature importance analysis

After the necessary features are created and explained, it is important to assess their significance in explaining the variability of the vegetation-related outages caused by weather-related factors. As demonstrated, such variability could be explained by various features, in which some take on a high importance, and some may be less significant. Conducting the feature importance analysis is particularly vital when the number of features is considerably large and obtaining some of them is difficult. In order to find the importance of each feature, the problem is formulated as an all-relevant feature selection problem in this study.

Since the number of features used in this study is relatively small, and these features are not sparse, a random forest-based algorithm, introduced in [20], is used to perform the all-relevant feature selection analysis. To find the relevant features and their importance, this algorithm calculates the sensitivity of the estimation model to random permutations of feature values. The rationale behind the permutation is that altering values for features that are not useful for estimation does not lead to a significant reduction in the model's performance. However, important features are usually more sensitive to the permutation, and will therefore gain more importance [21].

By implementing the random forest algorithm, all features discussed in this study are confirmed to be important. The importance of all features is demonstrated in Fig. 2.7. According to this figure, it can be understood that climatological factors take higher importance compared to geographical features; however, their difference is not considerable. It is worth noting that such conclusion could not be generalized and should be investigated for new datasets.

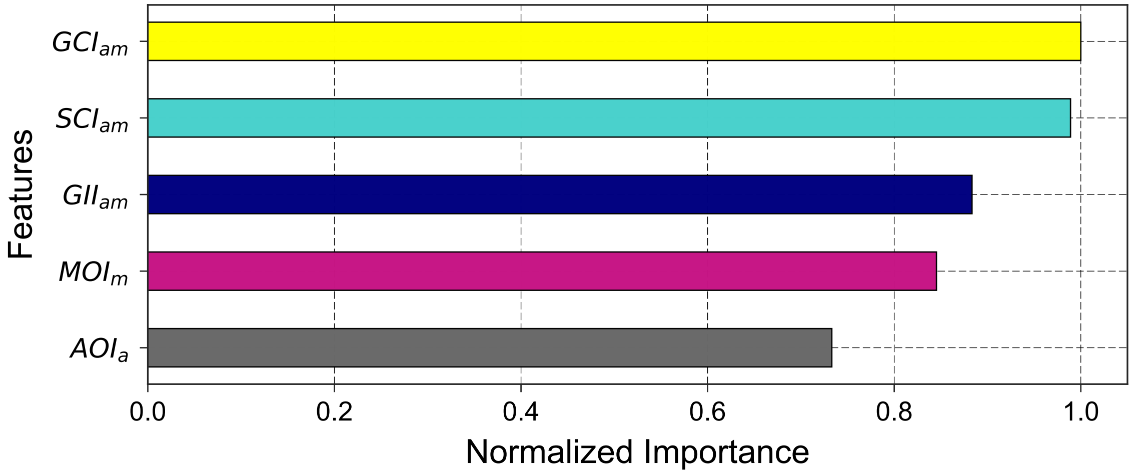


Figure 2.7: Feature importance analysis

2.5 Case study

2.5.1 Description

In order to demonstrate the effectiveness of the proposed approach for predicting the anticipated number of vegetation outages, a comprehensive case study is carried out in this section. As mentioned earlier, the proposed models and features are derived from the data that was obtained during the years 2011 to 2013. To conduct the case study, data from the year 2014 is utilized. During 2014, the outage information is available only for the first seven months; hence, the case study is carried out for these months only.

In order to implement each step of the proposed approach different statistical and machine learning algorithms are utilized. To compare the performance of these algorithms two different strategies are employed. The first strategy is to define an error metric and to analyze it. This metric is defined by the authors as Normalized Mean Absolute Error ($NMAE$), which for area a, is formulated as in (2.9).

$$NMAE_a = \frac{1}{M} \sum_{m=1}^{M} \frac{\left| \hat{OCI}_{am} - OCI_{am} \right|}{{OCI_a}^{max} - {OCI_a}^{min}} \tag{2.9}$$

where M is the total number of months included in the test dataset (seven in this case study), $\hat{OCI}_{am}$ is the predicted target variable, OCI_{am} is the actual target variable, and ${OCI_a}^{max} - {OCI_a}^{min}$ represents the range of outages that occurred for area a in the test dataset. It is necessary to mention that OCI_{am} is a general term and can represent any of $GVOCI_{am}$, $WVOCI_{am}$, and $TOCI_{am}$, which will be discussed later. This metric demonstrates the rate of prediction error considering how large the number of outages is in each area. As a matter of fact, including the range of outages in the formula is essential as it allows to normalize the error value for different areas and therefore show the result without any bias, making it easy to compare and evaluate. The usage of other metrics, which could possibly show better results, would be misleading.

The other strategy to compare the performance of the adopted algorithms is to statistically and visually investigate the predicted versus actual values of the target variable for each month aggregated over all areas. It is necessary to mention again that the ultimate goal of the study is to predict the number of outages in each month for each area; however, to gain a comprehensive understanding about the performance of the models, the aforementioned aggregated representation is highly beneficial. In what follows, the proposed approach is implemented, the error analysis is conducted, and the results are presented and discussed.

2.5.2 Predicting growth-related vegetation outages

As mentioned earlier, in order to predict the number growth-related vegetation outages, it is proposed to utilize time series analysis. There are various models to analyze and predict time series data. Among them, Auto-Regressive Integrated Moving Average (ARIMA) and Holt-Winters exponential smoothing methods are two of the most widely known algorithms. Both of these algorithms have different parameters that make them capable of modeling the major aspects of a times series data. Hence, these methods are employed in this case study to conduct the prediction task.

A brief explanation of these two algorithms is provided below [22].

ARIMA: This model is an extension of the auto-regressive moving average model so that non-stationary time series can also be handled. The AR part of ARIMA refers to the fact that the target value is regressed on its own lagged values. The I refers to the feature that the data values have been replaced with the difference between their values and the previous values (possibly more than once). The MA part indicates that the regression error is actually a linear combination of error terms in the past. Also, seasonal ARIMA (SARIMA) can take into account the seasonal component of the time series. The general forecast formula of a SARIMA$(p, q, d)(P, Q, D)_m$ is as

(10)

$$\hat{y}_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \theta_i e_{t-i} + \sum_{i=1}^{P} \Phi_i y_{t-im} + \sum_{i=1}^{Q} \Theta_i e_{t-im} \tag{2.10}$$

where $\hat{y}_t$ and y_t are the differenced (according to the values of d and D) versions of the predicted target values and actual target values at time t, respectively. Also, $e_t = y_t - \hat{y}_t$, and μ, ϕ_i, Φ_i, θ_i, Θ_i are the parameters that have to be estimated based on the training data.

Holt Winters Seasonal (HWS): This algorithm is based on decomposing the univariate target value into three different components and applying exponential smoothing to these components over time. These components can be combined in an additive or multiplicative way. In the additive form (used in this study), if the objective is to predict the value y at time $t + h$ given the data for all the times up to and including time t which is denoted as $\hat{y}_{t+h|t}$, then the following components are calculated based on the previous values for these components as:

$$L_t = \alpha(y_t/S_{t-m}) + (1-\alpha)(L_{t-1} + B_{t-1}) \tag{2.11}$$

$$B_t = \beta(L_t - L_{t-1}) + (1-\beta)(B_{t-1}) \tag{2.12}$$

$$S_t = \gamma(y_t/(L_{t-1} + B_{t-1})) + (1-\gamma)(S_{t-m}) \tag{2.13}$$

and after that $\hat{y}_{t+h|t}$ which is the prediction for time-step $t + h$ can be derived as:

$$\hat{y}_{t+h|t} = (L_t + hb_t)S_{t+h-m(k+1)} \tag{2.14}$$

where m is the length of the seasonal cycle, y_t is the actual value of the target at time t, and $k = \lfloor (h-1)/m \rfloor$. Also, $\alpha,\ \beta,\ \gamma$ (all belonging to the interval $[0, 1]$) are the tunable parameters of the algorithm, and there are effective methods to calculate their optimal values based on the data in the training dataset.

As demonstrated, the growth-related vegetation outages show a strong seasonal pattern; therefore, it is necessary to model the seasonality component. As a result, Seasonal ARIMA model (SARIMA) and a Holt-Winters Seasonal (HWS) method are utilized. It is worth noting that since these are well-established methods whose formulation and parameters are readily available [23], the further mathematical details of these models are not discussed in this study.

In order to build effective SARIMA and HWS models, it is necessary to choose the optimal values for their parameters. To conduct this task, after a careful data pre-processing, rolling based training and testing sets by utilizing the data obtained from years 2011 to 2013 are created. At each step, different sets of parameters are considered, a grid search is performed, the $NMAE_a$ on the validation set is calculated, the combination of parameters that deliver the lowest error is selected ans used to make the prediction on the test set. At the end, the combination of parameters that results in the lowest error value on average in validation sets is selected as the optimal parameters. Afterward, the model is built based on the selected set of parameters and is apply to make predictions for the year 2014. This procedure, which is known as k-fold cross validation, prevents over-fitting problem and helps demonstrate how the model results could be generalized. Fig. 2.8 shows this procedure.

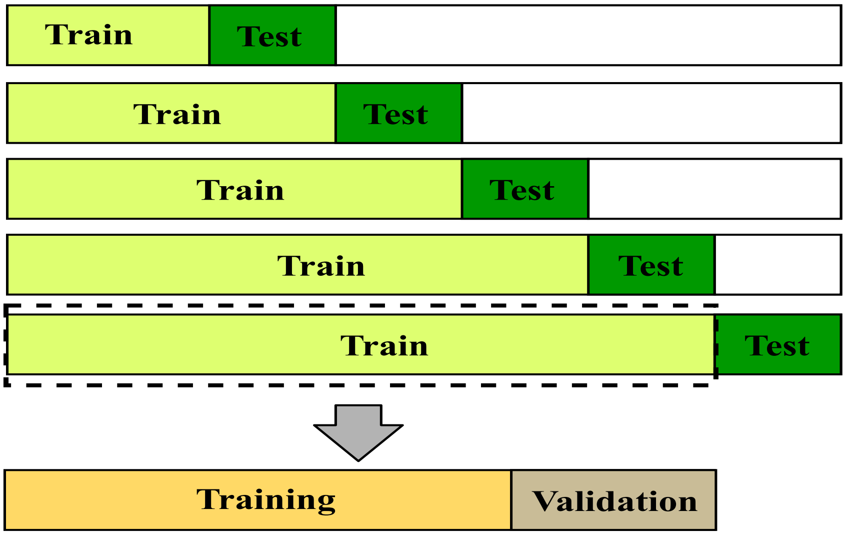


Figure 2.8: $k-$fold cross validation procedure

Fig. 2.9 demonstrates the $NMAE_a$ of all areas for SARIMA and HWS for the year 2014.

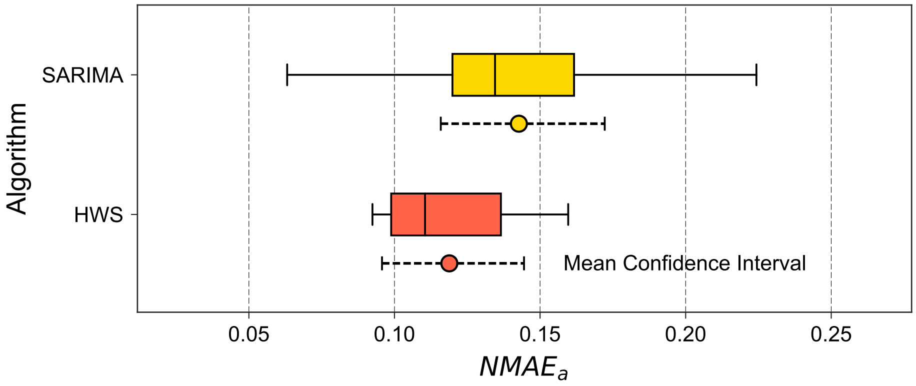


Figure 2.9: $NMAE_a$ for time series algorithms

The plot includes two types of statistical analyses, namely box plot and confidence interval of the average error. The box plot demonstrates the distribution of the error metric for all areas. The confidence interval provides a range of values which is likely to contain the population error value. The confidence interval is constructed based on the t-distribution, as well as a confidence interval of 95%.

As observed in Fig. 2.9, the HWS method delivers a narrower distribution of error values. Moreover, the average error for HWS is smaller compared to SARIMA. However, as seen, the confidence intervals for the methods overlap, suggesting that there is no convincing evidence of the difference between the population average error for these methods. The reason for HWS outperforming SARIMA for this case study could be explained from Fig. 2.10.

Fig. 2.10 demonstrates the predicted versus actual target values for each month aggregated over all areas. Moreover, the error for each algorithm is shown with dotted lines. According to this figure, SARIMA delivers a better performance for the winter months; however, during summer months, especially month July, its performance is poorer than HWS. The month of July for the year 2014 is a special month as the number of outages that occurred in this time period is considerably lower compared

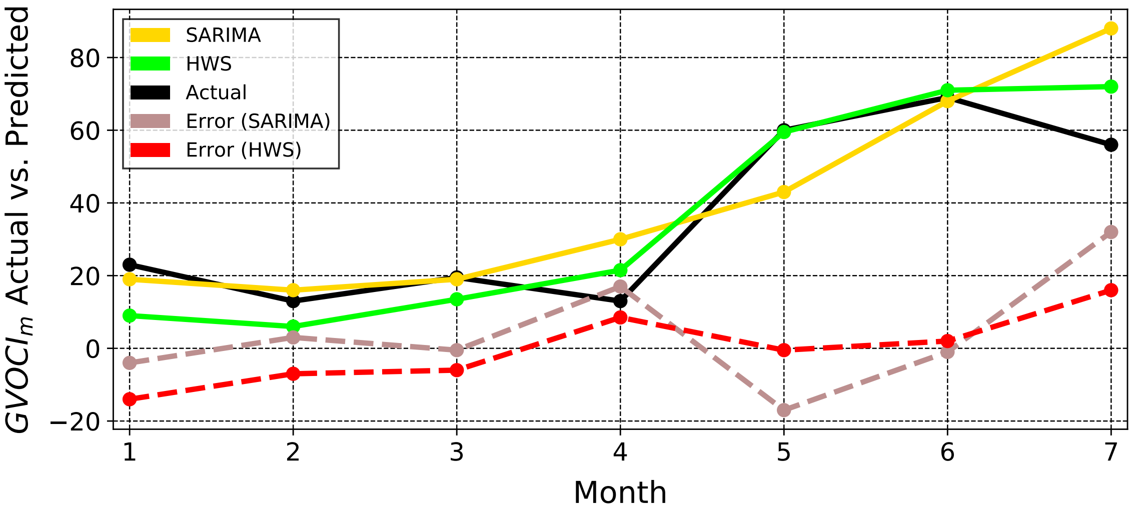


Figure 2.10: Actual vs prediction for time series algorithms

to previous years. This could be because of error in recording outages. Therefore, based on the error values observed from the figures, the authors believe that for this specific case study the HWS outperforms SARIMA, however, this argument cannot be generalized and should be tested for new outage datasets.

2.5.3 Predicting weather-related vegetation outages

In order to predict the number of weather-related vegetation outages, as proposed, machine learning regression models that can handle the non-linear interactions within the data, seem to be appropriate choices. As a result, three well-known models, namely Random Forest (RF), Support Vector Machine (SVM), and Neural Network (NN) were selected to carry out the prediction task in this case study. It is worth mentioning that several other machine learning methods such as KNN and Naive Bayes were also investigated; however, the aforementioned models (RF, SVM, and NN) delivered the best performance. It also should be noted that since the above-mentioned methods are all well-established methods whose formulation and descriptions are readily available, the details of these methods are not discussed here.

In order to train the aforementioned models, the data obtained from years 2011 to 2013 is utilized. Moreover, the hyper-parameters associated with each model are tuned by using the k-fold cross-validation technique, where k is selected to be five. Considering the fact that the size of the data used in this study is not considerably large, k-fold cross-validation appears to be essential, particularly to avoid the over-fitting problem. Hyper-parameters that, on average, performs best on cross-validation are selected as the optimal set of parameters to train models. Afterwards, the trained models are applied to the test data, the year 2014, and predictions are produced. The optimal hyper-parameters are as follow:

1. RF: trees = 20, mean sample leaf = 1, depth = 4
2. SVM: kernel = rbf, degree = 4, $C = 1$
3. NN: hidden layers = 1, hidden units = 120

Fig. 2.11 illustrates the distribution of the error as well as the confidence interval of the average error for each algorithm. As seen in the figure, the inter-quartile range for SVM is narrower, suggesting the variability of error for this algorithm is lower in this case study. Also, it can be observed that the maximum error for SVM is smaller compared to the same statistic from two other algorithms. However, overlapping confidence intervals indicate that there is no convincing evidence for the superiority

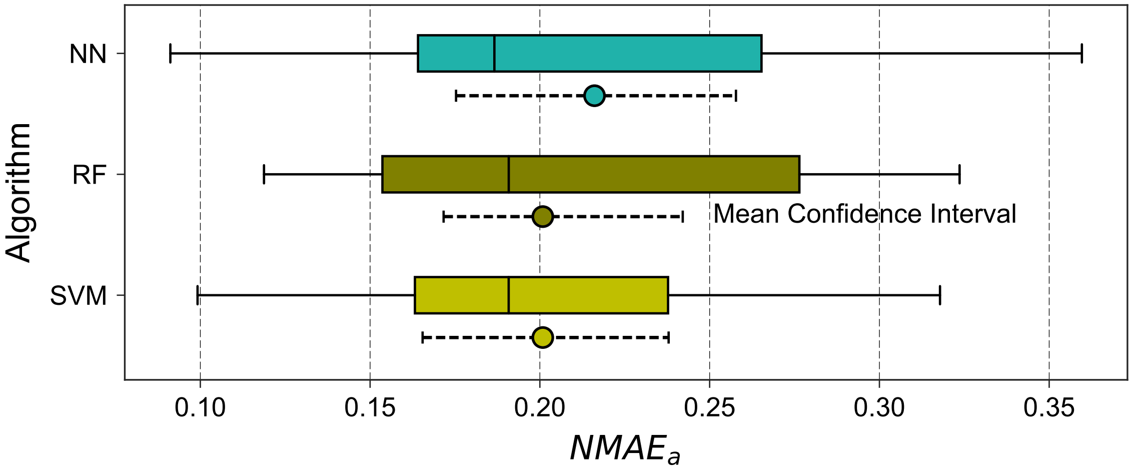


Figure 2.11: $NMAE_a$ for machine learning algorithms

of any algorithm in terms of the population average error. The performance of these algorithms can also be evaluated by analyzing the predicted versus actual target values for each month aggregated over all areas shown in Fig. 2.12.

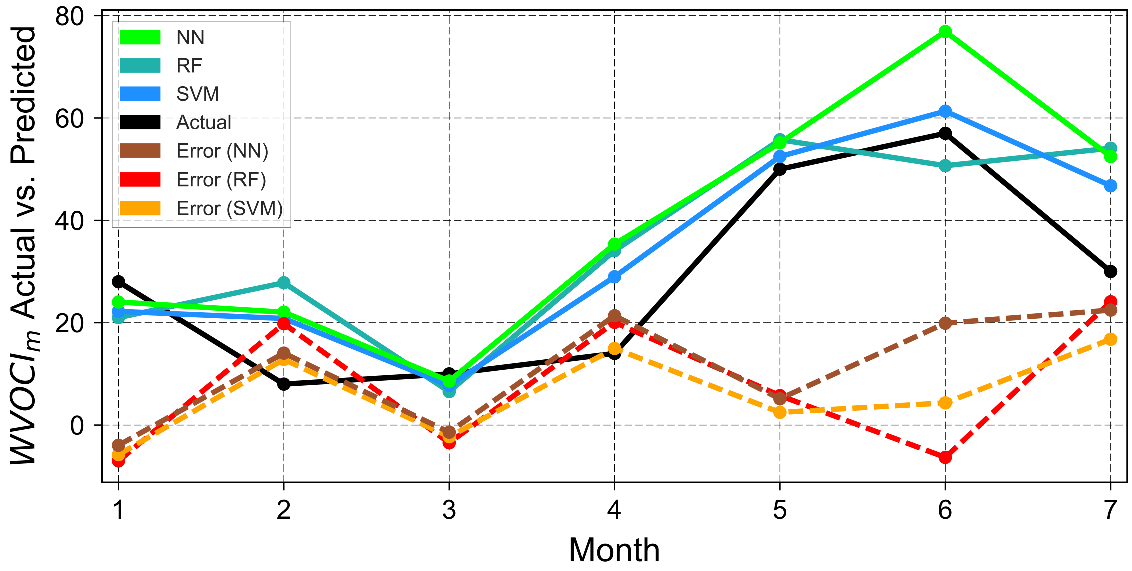


Figure 2.12: Actual vs. prediction for machine learning algorithms

Based on the figure, it can be realized that the SVM follows the actual target value better, delivering smaller errors (shown in dotted line), particularly in the summer months. It is worth noting that while the SVM results in better performance for this case study, such superiority cannot be generalized, and therefore should be explored for each dataset separately. Moreover, considering the fact that the scale of data used in this case study is not significantly large, it would be difficult to state an obvious reason as to why the SVM works better.

2.5.4 Producing final predictions

The total number of vegetation-related outages can be defined as Total Outage Count Index ($TOCI$). The prediction of this index for area a and month m may be calculated using (2.15).

$$TO\hat{C}I_{am} = GV\hat{O}CI_{am} + WV\hat{O}CI_{am} \tag{2.15}$$

where ˆ denotes the predicted value.

The HWS and SVM deliver the best performances on cross validation procedure; hence, the final prediction is produced by using these algorithms. Fig. 2.13 demonstrates the actual number of vegetation-related outages that occurred in the year 2014 versus the prediction generated by the proposed approach for each month aggregated over all defined areas. In order to show the effectiveness of the proposed approach and to compare it with very simplistic approaches, a naive model have been created where its result are presented in the figure as well. In the naive model, the outages that occurred in the past for each area on each month are considered and the average value is calculated.

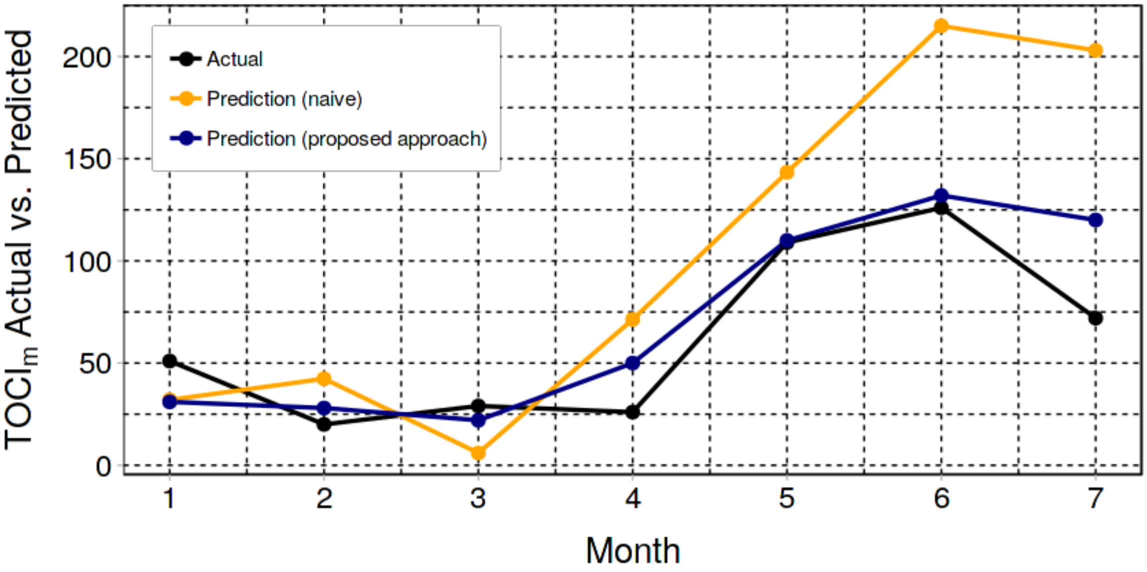


Figure 2.13: Final actual vs. prediction

Based on the figure, it can be understood that the proposed approach, in general, is able to generate prediction values that closely follow the actual value and its trend. In particular, as seen, for months February, March, May, and June, the proposed approach delivers excellent performance. However, differences between the prediction and actual counts for the months January, April, and July may be observed. As explained, the months of April and July in the year 2014 were special months since the number of weather-related vegetation outages that occurred during this time period were significantly lower than during the same time in previous years (note that this may be due to recording errors). It is also worth mentioning that although the proposed approach can be used to make the prediction for any month in the future, the best performance is expected to be achieved when the model is utilized for one month ahead. This is because the most accurate weather forecasts are available for one month ahead. Moreover, time series models are particularly suitable for short-term predictions.

Compared to the proposed approach, the naive model delivers a poor performance. As seen in the figure, especially for months April to July, the naive model over-predicts the outages by a substantial margin of error. This demonstrates that simply calculating the average number of outages by using the historical data and using those for making the prediction does not lead to satisfactory results. Moreover, this confirms that efforts devoted to develop the proposed approach result in significant improvements in producing a prediction for vegetation-related outages.

The performance of the proposed approach can also be evaluated by exploring the normalized error values. In order to demonstrate the superiority of the proposed approach with regards to error values, two other models will be utilized to generate the final prediction and the results of those models will be compared with the proposed approach. The first model is the naive model (i.e., simple average), which was introduced earlier. The second model is a benchmark machine learning model. In the benchmark model, as opposed to the propose approach in which the outages are categorized into two groups and separate models are built for them, all available inputs are fed into one model and the prediction is generated. The inputs to the benchmark model are all features shown in Fig. 7, as well as area number (dummy variable) and various time-related features. These inputs are fed into three machine learning algorithms of SVM, NN, and RF to generate the final prediction. By investigating the performance of the aforementioned algorithms, it was realized that combining the predictions produced by those (i.e., averaging the predictions of all three algorithms) leads to obtain the best performance for the benchmark model. It is worth-mentioning that the hyper-parameters of those algorithms are tuned properly using 10-fold cross-validation procedure. The values are as follow:

1. RF: trees = 10, mean sample leaf = 1, depth = 4
2. SVM: kernel = rbf, degree = 3, $C = 1$
3. NN: hidden layers = 1, hidden units = 100

Fig. 2.14 presents the $NMAE$ for each area and each model shown with bars. As seen, the performance of the proposed approach for some areas is remarkable with a minimum error of 0.12. Also, on many occasions, especially for areas 7 and 8, the proposed approach outperforms the benchmark machine learning model considerably. The naive model, on the other hand, delivers large errors on most cases. The average error of all areas for each model is presented in Table 2.1. From the table, it could

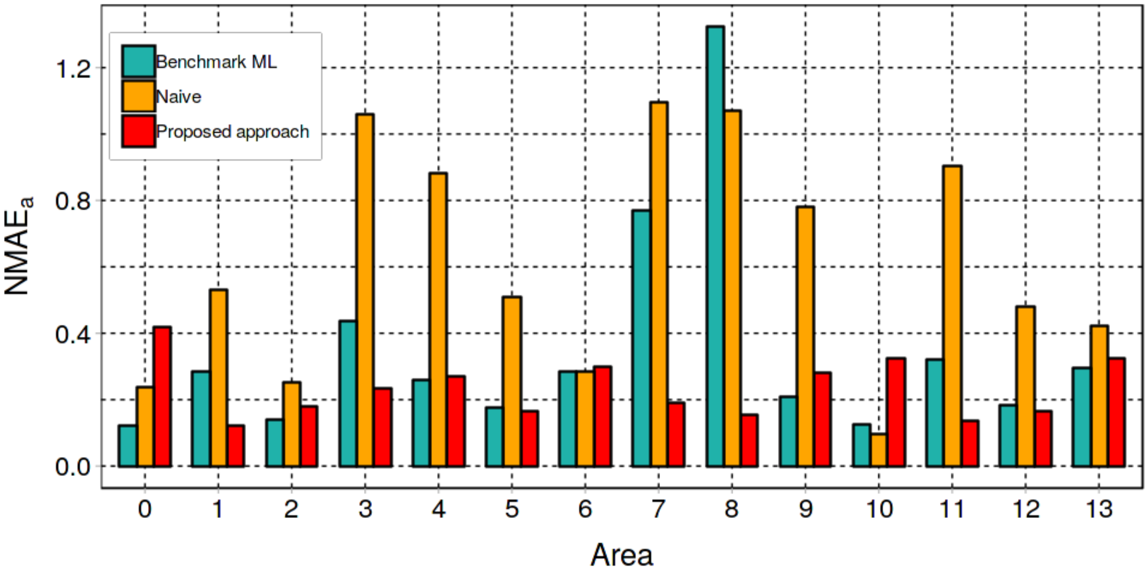


Figure 2.14: $NMAE_a$ for final prediction for different areas

Table 2.1: AVERAGE NMAE FOR DIFFERENT MODELS

Model	**Mean NMAE**
Naive	0.61
Benchmark ML	0.35
Proposed approach	0.23

be understood that the proposed approach results in the best prediction performance by delivering an average error of 0.23. The error for the benchmark machine learning model is 0.35, which is considerably higher. The results of the naive model, as expected, is the worst. From the results, it can be inferred:

1. The proposed approach outperforms the naive model, demonstrating that efforts devoted to develop a comprehensive approach leads to obtaining significant improvements compared to very simplistic approaches;

2. The proposed approach outperforms the benchmark machine learning model, demonstrating that categorizing vegetation-related outages and building separating models for them leads to improvement compared to building a single model without differentiation between categories of outages

A visual demonstration of what the proposed approach can offer to the decision makers is given in Fig. 2.15. Here, the prediction for the month of May in the year 2014 for each area is presented. Since the number of outages that occur in the system is predicted, it provides great flexibility because the utility companies can decide whether or not the number of outages is critical based on their criteria and subsequently can categorize outages based on their severity. For the sake of illustration, the number of predicted outages are categorized into four distinct categories and color-coded, accordingly in this study. For example, one can realize that areas 4 and 10 have a high number of outages. On the other hand, areas 3, 7, and 8 will experience a smaller number of outages in the aforementioned month. Using this platform, the operators can quickly identify vulnerable zones, and take necessary actions.

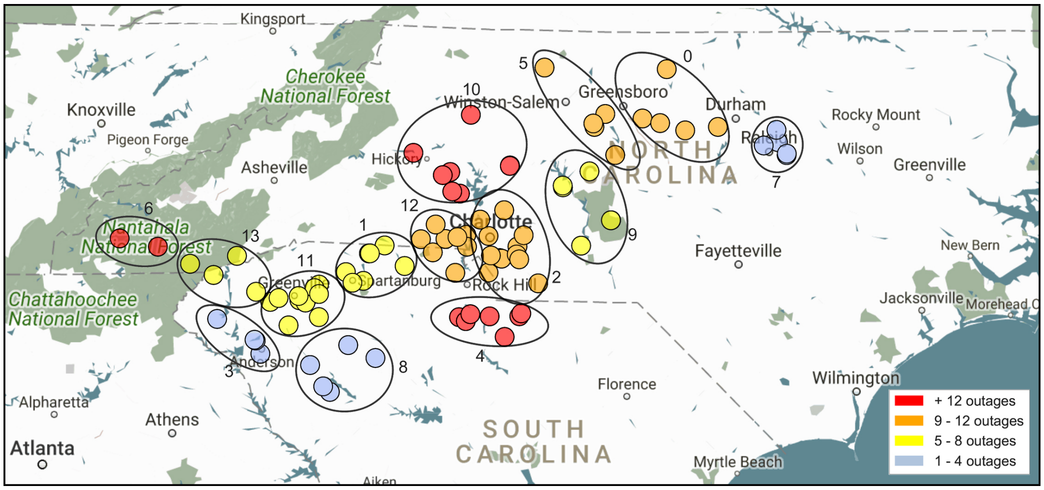


Figure 2.15: A demonstration of the application of the proposed approach for different areas

2.6 Conclusions

A data-driven approach was proposed for predicting the number of vegetation-related outages in power distribution systems on a monthly basis. Based on this study, the following conclusions can be drawn.

1. In order to develop a practical approach, various types of information including historical records of outages, climatological, and geographical variables should be obtained and processed.

2. A successful approach requires a great deal of attention to the data pre-processing task. In particular, conducting a comprehensive outlier detection and handling process is essential.

3. A key step in building a realistic approach is to adequately define the extent of the predictions' target area. Aggregating substations and creating broader geographical areas by using clustering algorithms seem like a workable solution for this purpose. This helps to harness the randomness of the number of outages and to obtain more accurate weather information.

4. Vegetation-related outages occur due to various reasons; however, they could be

categorized into two main groups: growth-related and weather-related outages. Each category of the vegetation-related outage reveals a different pattern, and therefore requires a different treatment. Such categorization is necessary to build an accurate model.

5. Due to the complex nature of vegetation-related outages and several factors that influence their occurrence, utilizing simplistic approaches such as calculating the average number of outages based on historical data does not lead to obtaining an accurate predictive model; therefore, more sophisticated models are required. Moreover, the occurrence of these outages depends on various factors, which are subject to change with time; hence, models that take these factors into considerations are required.

6. Time series models can successfully explain the patterns existing within growth-related vegetation outages, and are able to produce convincing predictions.

7. Machine learning regression models that can handle the non-linear interaction in the data are effective tools for predicting weather-related vegetation outages.

Although many different pieces of information pertaining to vegetation-related outages were considered in this study, all possible factors were not accounted for due to lack of access to related data. As a result, the performance of the proposed approach may be improved by the inclusion of additional climatological and geographical information (e.g., satellite images for more vegetation data). In fact, as mentioned earlier, all the advantages of the proposed approach are built upon generic outage data collected by utilities, and typical daily weather forecast data, which is publicly available. This fact makes the implementation of the approach easily attainable within a reasonable level of accuracy. However, the approach provides the flexibility to be improved by utilizing various other sources of data.

Moreover, obtaining a considerably larger scale of outage data may open up unique opportunities for utilizing the most advanced predictive models including deep learning algorithms for predicting vegetation-related outages. It is worth mentioning that the main contribution of this chapter is not to compare the performance of different predictive algorithms, but rather providing an approach for building robust predictive models for vegetation-related outages and creates solid foundations for it. Hence, utilizing more sophisticated time series models and AI approaches within the proposed approach is encouraged, and could result in more accurate predictions.

Results of this study could be very informative to utility companies in gaining insight about their vegetation-related outage problems, and to build better models for predicting them. Especially, by providing a preliminary but accurate prediction, the proposed approach enables operators to take high-resolution imagery of areas with high risk of an outage, or utilize LiDAR data, or dispatch a crew to find the exact locations in the system that a vegetation-related outage could occur. Moreover, in this study, workable solutions to some existing problems were proposed and several new features that can be used by researchers to improve their outage predictive models were generated.

CHAPTER 3: Predicting Lightning-Induced Outages in Power Distribution Systems: A Statistical Approach

3.1 Introduction

3.1.1 Motivation

Lightning is a major cause of outages in power distribution systems [24]. Transient over-voltages caused by direct or indirect lightning strikes may inflict severe damages to the susceptible equipment and can produce detrimental effects on power quality and reliability of the system. With upward trends in extreme weather and climate events in recent years, the intensity and frequency of the lightning activities are expected to increase, leading power utilities to be confronted by a growing problem with regards to this weather-related phenomena [25], [26], [27].

In order to reduce the destructive effects of lightning on distribution systems, a common strategy is to implement a proper lightning protection design, i.e. installing surge protective devices and shielding wires [24]. While taking such preventive actions appear to be effective for protecting the system against severe damages, momentary outages caused by the activation of these protective devices can exert secondary effects on the system. On the other hand, should the lightning effect exceeds the designed levels of the protection system, permanent faults may occur, leading the system to experience a sustained interruption [28]. Therefore, besides implementing preventive measures, it is important for utilities to take an appropriate response to the lightning-induced outages either by identifying them immediately after they occur or by predicting them in advance.

By looking at historical outage data, analyzing electrical characteristics of the system, and tracking lightning activity in an area immediately after an outage occur, the utility companies are able to identify and distinguish lightning-induced outages from other causes. Moreover, multiple studies have been carried out to develop models for identifying this source of outages [29], [30]. Opposed to that, predicting lightning-induced outages has been left relatively unattended. This might stem from the complex nature of this type of outages, lack of enough information pertaining to them, shortcomings of classical mathematical models, and the fact that a multitude of factors influences the occurrence of these outages.

However, with the technological advancement in data gathering through the smart grid framework and due to tremendous improvements in weather forecasting efforts, a massive amount of data has become available in recent years. This could shed light on the different characteristics of lightning-induced outages. Moreover, advanced data analytics techniques are now being developed and combined with mathematical methods, creating powerful tools that can noticeably enhance predictive abilities. Adopting these predictive approaches will enable effective and timely decision-making actions by operators as well as planners, ultimately improving operational integrity and resiliency of the system.

3.1.2 Literature Review

By this time, several studies have been carried out to explore different aspects of lightning-induced outages and to ultimately predict some characteristics of these outages.

The authors in [31] propose an approach for estimating the number of wind and lighting-related outages combined together on a daily basis. In order to carry out this task, they develop a machine learning regression model based on an ensemble boosting algorithm. In [32], the authors propose a method for forecasting the cumulative number of outages during a storm condition. They first create empirical models for different types of storms and then develop an exponential model for the forecasting purpose. They employ the model to predict lightning-induced outages that may have occurred during several summer storms.

Another study [33] presents a Monte Carlo simulation model to study the reliability indices under lightning storm condition. Their model is based on the storm parameters and the outage rate. Therefore, first, the authors build a statistical model to explain the storm intensity and duration and then utilize this model as well as a data-driven approach to calculate the lightning-induced outage rate. The performance of the proposed model is evaluated by conducting a case study using data collected from lightning storm weather conditions that occurred in an area in the Midwest United States in the span of five years. In [24], the authors carry out an experimental study to investigate significant factors that influence the frequency of lightning strike flashovers. Moreover, they develop a probabilistic model for estimating the number of lightning-induced outages on an annual basis.

Even though some approaches are proposed to predict the rate or trend of lightning-induced outages in the aforementioned studies, there are major problems associated with them, making them impractical for being implemented by utility companies. As a matter of fact, these approaches are mostly developed based on the combination of weather-related outages and therefore cannot to be used for predicting only lightning-induced outages. Moreover, they consider the entire distribution system under study for the prediction task; hence, do not provide the ability to make the prediction for a specific area within the system. Furthermore, a majority of these approaches make the prediction for a long-term horizon (i.e., yearly) and consequently cannot be utilized for making predictions on a short-term horizon (i.e., daily or weekly). Last but not least, such approaches are expected to deliver a low degree of performance when the number of outages is large. This argument would be fully supported later on.

3.1.3 Contributions

In this study, a novel approach to predict lightning-induced outages that occur in power distribution systems is proposed. In particular, an approach is built that is able to predict whether zero outage, one outage, or two or more outages will occur on a specific day that experiences thunderstorm events in a particular area in the system. Utilizing statistical and machine learning predictive models, the proposed approach crafts an intelligent solution that combines weather forecast data with past lightning-induced outage information and makes a prediction for future outages. The proposed approach overcomes the aforementioned shortcomings and provides the ability to make the prediction on a short-term horizon (i.e., daily basis) for a specific area within the service territory.

The proposed approach provides a meaningful knowledge about risks and locations of lightning-induced outage problems. This presents a succinct view of the current system status to the operators, which enables effective and timely decision-making actions with regards to lightning-induced problems. By providing a preliminary but accurate prediction, the proposed approach allows operators to utilize satellite imagery or sophisticated lightning detection systems to find the exact locations in the system that could have high risk of a lightning-induced outage.

The main contributions of the proposed approach are summarized below:

1. A workable solution to address the challenges brought about by the extent of the prediction's target area is offered.

2. It is demonstrate that to obtain the best possible predictive performance, lightning-induced outages should be clustered into a few manageable groups, which in this study, are three main groups, namely zero outage, one outage, and two or more outages.

3. A probabilistic model is provided to calculate the likelihood of each group of

outages and validate the model using statistical tests.

4. A machine learning classifier is developed that can handle the imbalanced problem and utilize it to predict the specific group of outages that will likely occur.

3.1.4 Study Organization

The rest of this study is organized as follows. In Section 3.2, a detailed problem description is provided where the objectives and the proposed approach are explained in detail. In Section 3.3, the outage and other data used in the analysis is described. In Section 3.4, an approach is presented to address the challenges brought about by the extent of the predictionâs target area. In Section 3.5, a statistical analysis is provided to cluster the outages. In Section 3.6, a probabilistic model is provided for calculating the likelihood of outages. In Section 3.7, a machine learning approach is presented for predicting the outages. In Section 3.8, three case studies are provided to show some practical issues associated with predicting lightning-induced outages, and to demonstrate the effectiveness of the proposed approach to address those issues. Finally, in Section 3.9, conclusions are provided. The flow of the study discussed above is illustrated in Fig. 3.1.

3.2 Problem Description

3.2.1 Objectives and Approach

In this study, the main objective is to predict the occurrence of lightning-induced outages on a daily basis in a particular area within the power system. In order to carry out this task, there are two main challenges that should be addressed.

The first challenge is defining the extent of the prediction's target area. As a matter of fact, if one intends to point-predict the number of lightning-induced outages at the location of any given substation, one might encounter serious difficulties. These difficulties lie in the fact that even though the weather forecasts may show the possibility of thunderstorm events at the location of the substation for a given day, the lightning

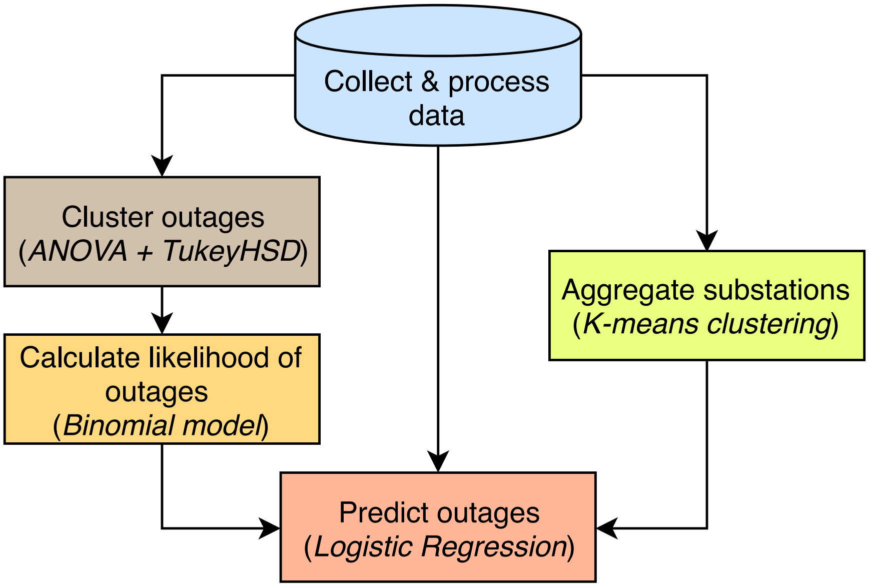

Collect & process data
Cluster outages (*ANOVA + TukeyHSD*)
Calculate likelihood of outages (*Binomial model*)
Aggregate substations (*K-means clustering*)
Predict outages (*Logistic Regression*)

Figure 3.1: Technical flow-chart of the proposed approach

may travel and hit a location that is at a large distance from the substation. On the other hand, lightning flash can occur at a considerable distance from the substation but hits a feeder connected to the substation (In the latter case, a substation weather forecast does not show any thunderstorm event). Moreover, accurate radar weather forecasts may not be available at the exact location of each substation as the weather station could be far from the substation. Furthermore, the degree of randomness for the number of outages that occur for each substation is relatively large, making it difficult to predict. As a result, developing an approach that is capable of predicting lightning-induced outages at an exact given location in the system is neither realistic nor practical. Hence, an appropriate extent for the prediction's target area should be defined.

To the best of this author's knowledge, this problem has not been investigated in detail yet. As mentioned before, one reason could be the limited access to a sufficient amount of data. However, thanks to the considerable amount of data provided by a large power company in the southeastern US, this problem could be addressed in this study. In order to deal with this issue, it is proposed to aggregate substations and build larger areas. This will be explained in Section 3.4.

The second challenge is to adequately investigate the characteristics of lightning-induced outages and to explore the relationship between these outages and weather-related variables (thunderstorm events). In order to take the aforementioned factors into consideration, and to build a realistic predictive model, the following chronological steps are suggested:

1. First, it is demonstrated that the outages should be clustered into a few manageable groups. In this study, three groups are zero outage, one outage, and two or more outages are created.

2. Next, a probabilistic model for estimating the likelihood of each group for a given area and given daily weather condition is presented.

3. Finally, a machine learning classifier that can handle the imbalanced problem to predict what group (zero, one, two or more) will the outage belong to on a specific day at a specific area in the system is built. The imbalanced problem and a workable solution to handle that problem are discussed in Section 3.7.

The aforementioned steps, as well as supplementary arguments, are provided in Sections 3.5 through 3.7.

3.3 Data Description

The input data for the proposed approach is obtained from two main sources: 1) historically recorded outages, and 2) radar weather forecasts. Outage data is collected by a major investor-owned utility company serving the southeastern US. The data includes information on the time and locations of lightning-induced outages that occurred around approximately 85 substations located in the states of North Carolina and South Carolina between the years 2010 and 2014. The data is comprised of almost 800 samples of outages. The radar weather forecast data is collected from several external sources for weather stations located close to power substations over the span of the aforementioned years and includes the number of thunderstorm events that occurred on a daily basis for each weather station. In order to calculate the number of thunderstorm events, the hourly weather forecast for the entire 24 hours is considered and the summation value of the logical variable that shows whether or not each hour might experience a thunderstorm is calculated. The logical variable is available in almost any weather forecast platform.

3.4 Aggregating Substations

As mentioned, one of the main challenges with developing the proposed approach is defining the extent of the prediction target area. In order to deal with this issue, the proposal is to aggregate substations and to build larger areas, in which, each area includes multiple substations and local weather stations, where the weather forecast for each substation is obtained from its closest weather station. Such clustering can effectively solve challenges discussed in Section 3.2. This is because by aggregation, instead of examining a single substation and a single weather station, a broader area that contains multiple weather stations is examined, and the average number of thunderstorm events in the area is calculated. By doing this, a better weather forecast is obtained. Moreover, if the flash is created within the area, even if it travels, it is highly likely that it ultimately hits a point within the area. Also, aggregation reduces the considerable randomness in the data and helps to see the patterns more clearly.

In order to define the aforementioned areas, k-means clustering algorithm is utilized in this study. This algorithm is a widely used unsupervised machine learning algorithm, which aims at clustering a given dataset into a certain number (k) of groups. In this study, this algorithm is used to cluster substations into different groups where each group represents an area. The main idea of this algorithm is to define k centroids at random, one for each cluster, and then to minimize the squared error function represented in (3.1) [34].

$$J(r, \mu) := \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{k} r_{ij} \left\| x_i - \mu_j \right\|^2 \tag{3.1}$$

where m is the number of data points, k is the number of clusters, r_{ij} is an indicator, which is 1 if, and only if, x_i is assigned to cluster j, x_i is data point, μ_j is the centroid for cluster j, and $\|.\|^2$ denotes the Euclidean distance. In this study, data points are locations of substations (approximately 85 data points), which are represented by latitude and longitude in a two-dimensional space.

One major challenge with this algorithm is the need to specify the number of clusters. In fact, there is no global theoretical method to find the optimal value of this parameter; however, a few approaches are common to deal with this problem. One workable approach is to run k-means clustering for a range of different k values and to calculate the aforementioned squared error function for each value. In this case, the error tends to decrease toward zero as k increases; however, after a certain k value is reached, the decrease in error would be very gradual. Therefore, analyzing different values of k and finding the aforementioned threshold could help in deciding a reasonable number of clusters [35].

Applying k-means clustering algorithm and using the aforementioned method to calculate the proper number of clusters for grouping substations will result in 14 areas. These areas define the extent of the prediction. It is worth mentioning that Lloydâs algorithm is utilized in this study to solve the k-means clustering problem. Fig. 3.2, illustrates these areas. It is worth mentioning that although there are other clustering approaches, the approach adopted here is the most suitable for this study. In fact, since the goal is to cluster a two-dimensional data (i.e., latitude and longitude) and to work with distances, the k-means approach makes the most sense. Moreover, considering the size of clustering data which is small, there would be no need for utilizing more sophisticated algorithms.

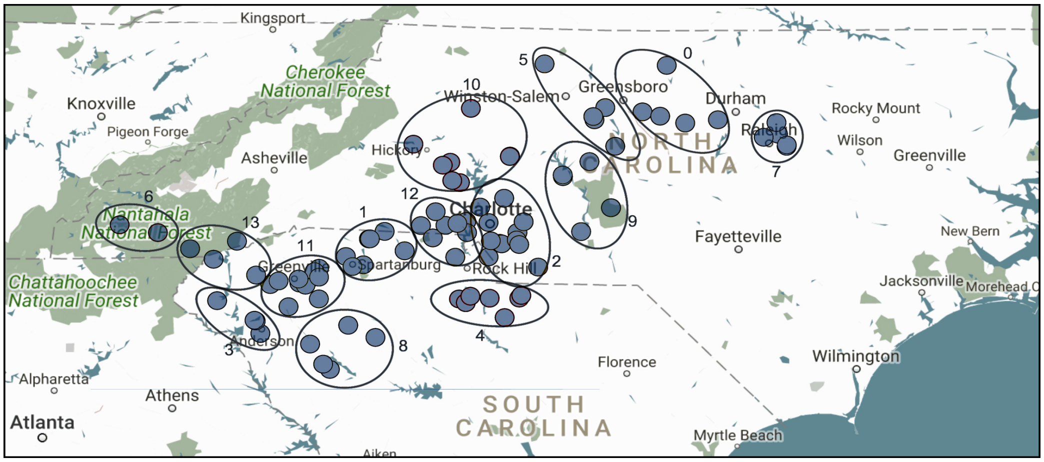


Figure 3.2: Demonstration of different areas

3.5 Clustering The Outages

By conducting various analyses, it is concluded that predicting the exact number of outages in days that have thunderstorm events may not be possible. In fact, it can be argued that an increase in the number of thunderstorm events does not necessarily translate into an increase in the number of outages. In what follows, statistical supports to these argument will be provided. In particular, it is postulated that lightning-induced outages may be clustered into a few manageable groups. In this study, these groups are 1) zero outage, 2) one outage, and 3) two or more outages.

Before providing the rationale behind this argument, a fresh look at a sample of the dataset under study is necessary. Table 3.1 shows a sample of the dataset that includes six observations. As seen, each observation is associated with four attributes of time, area, number of thunderstorm events, and the number of outages. The complete dataset contains the aforementioned information for all fourteen areas shown in Fig. 3.2 for the years 2010 to 2014 on a daily basis.

The research question that will be explored is the relationship between the number of thunderstorm events and the number of outages. The maximum number of outages which has been recorded in the available data is five. As a result, the number of

Table 3.1: A SAMPLE OF THE DATASET UNDER STUDY

Time	Area	Thunderstorm Events	Number of Outages
2011-08-13	5	0	0
2012-08-22	3	3	1
2012-05-22	12	30	2
2012-09-02	8	26	3
2013-07-17	8	50	4
2014-07-03	12	38	5

outages has a limited number of outcomes and therefore could be considered as an ordered categorical variable. This consideration is reasonable because it would be highly unlikely that an area experiences a large number of lightning-related outages on a specific day. As a result, a sensible approach to investigate the relationship between the aforementioned variables is to analyze the spread of the number of thunderstorm events for each possible value of outages through a box plot, as illustrated in Fig. 3.3.

From the plot, it may be inferred that for occasions with zero outage, the number of thunderstorm events is considerably smaller than for days with any number of outages. This can be realized by observing the median value (solid horizontal line in the middle of the box). In fact, the median of thunderstorm events for zero outages is significantly smaller compared to other outage instances (i.e., it is close to zero).

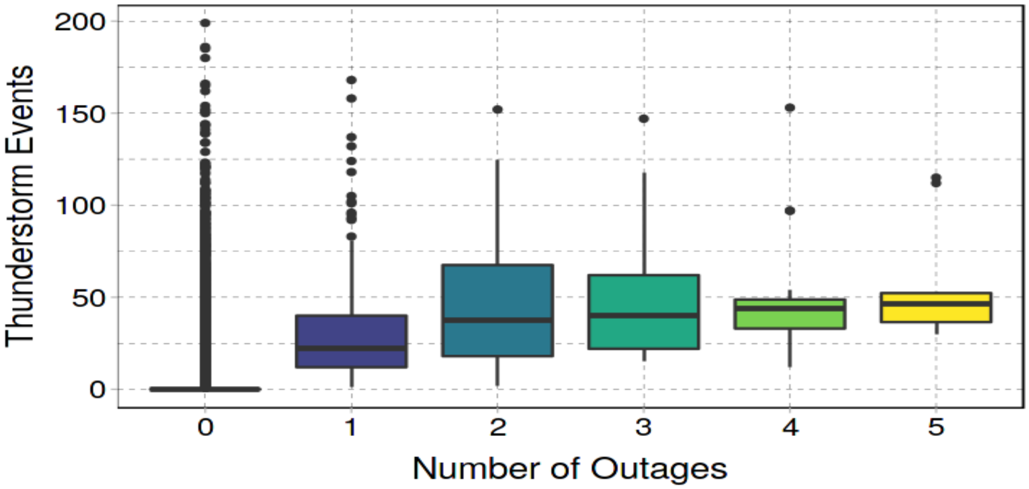


Figure 3.3: Number of thunderstorm events for each number of outages

The number of thunderstorm events for occasions with one outage also seems smaller compared to days with two or more outages. On the other hand, the median value of the number of thunderstorm events for occasions with two or more outages seems larger and close to each other. It is worth mentioning that dot points in the plot demonstrate the outlier values.

In order to quantitatively explore the relationship between the number of thunderstorm events and outages, and especially to realize whether or not there is a difference between the number of thunderstorm events among different values of outages (including days with zero outage), a one-way ANOVA test is carried out. The ANOVA may be used to determine whether there are any statistically significant differences between the means of two or more independent groups regarding a specific explanatory variable [36]. In this study, the null hypothesis in the test would be that the mean number of thunderstorm events is the same across different values of outages. On the other hand, the alternative hypothesis would be that at least one pair of means is different from each other.

This analysis is conducted and the results are provided in Table 3.2. It is worth mentioning that the necessary analysis is carried out to make sure that the ANOVA assumptions [36] hold for this study and the available data can be analyzed using this test. According to the table, the p-value is almost zero; hence, the null hypothesis can be rejected in favor of the alternative hypothesis and it can be concluded that the average number of thunderstorm events is not equal for different groups of outages. This is not a surprise because significant differences were observed in the boxplot. However, a follow-up question, which could shed more light on the differences between

Table 3.2: RESULTS OF ANOVA

Source	**DF**	**SS**	**MS**	**F**	**P**
Regression	5	389018	77804	403	<2e-16
Residuals	20126	3885367	193		

values of outages with regards to the number of thunderstorm events is to investigate the difference in a pairwise manner and to quantify it. In order to carry this out, a post-hoc test, Tukey's HSD, is conducted. This test allows answering which means are different and by how much and whether or not the difference between outages in a pairwise manner is statistically significant. It is worth mentioning that since the ANOVA and Tukey's HSD are very well-established methods whose formulation and descriptions are readily available [36], their details are not discussed in this study.

The results of the Tukey's HSD are provided in Table 3.3. The pair column shows the combination of days with two different numbers of outages. The difference column represents the differentiation between the average number of thunderstorm events between pairs. The lower and upper columns demonstrate the limits of the 95% confidence interval for the difference in the average value, and finally, the p-value shows the results of the hypothesis that there is not any statistically significant difference in the population of the average number of thunderstorm events for each pair. A p-value less than 0.05 shows that the results are significant.

Table 3.3: RESULTS OF TUKEY'S HSD TEST

Pair	**Difference**	**Lower**	**Upper**	**p-value**
1-0	28.8	26.2	31.5	0.0
2-0	40.9	35.5	46.3	0.0
3-0	43.7	36.3	51.0	0.0
4-0	45.6	36.3	55.0	0.0
5-0	51.7	39.1	64.2	0.0
2-1	12.1	6.1	18.1	0.0
3-1	14.9	7.0	22.7	0.0
4-1	16.8	7.1	26.5	0.0
5-1	22.8	10.0	35.6	0.0
3-2	2.7	-6.4	11.9	0.9
4-2	4.7	-6.1	15.5	0.8
5-2	10.7	-2.9	24.4	0.2
4-3	1.9	-9.9	13.8	0.9
5-3	8.0	-6.5	22.5	0.6
5-4	6.0	-9.6	21.7	0.9

According to the table, it can be inferred that for days with zero outage, compared to days with one or more outages, the average number of thunderstorm events is smaller and the difference is statistically significant. This is because the confidence interval does not include zero (i.e., equivalently, the p-value is zero). The same argument could be made for days with one outage compared to other days. However, for days with two or more outages, as seen in the table, the confidence interval ranges from negative to positive values (i.e., it includes zero), and the p-value is greater than 0.05. As a result, we fail to reject the hypothesis that there is no difference between these pairs with regards to the number of thunderstorm events and therefore can conclude that there is sufficient evidence that the average number of thunderstorm events is the same for days that have two or more outages.

The aforementioned analysis creates the foundation for clustering outages. Considering the facts that 1) there is a causal relationship between thunderstorm events and lightning-induced outages and that 2) there is a significant difference between the days with zero outage compared to other days, and days with one outage compared to other days, and 3) the observation that days with two or more outages do not share distinguishable characteristics with each other with regards to the number of thunderstorm events, outages are clustered into three groups: 1) zero outage, 2) one outage, 3) two or more outages.

To further confirm that these groups show distinguishable characteristics with regards to the number of thunderstorm events, the empirical cumulative distribution function of the number of thunderstorm events is provided in Fig. 3.4. From the figure, it may be seen that days with two or more outages show a higher number of thunderstorms compared to days with one and zero outage, and days with one outage show higher number of thunderstorm events compared to days with zero outage.

Considering the aforementioned clustering, the ultimate objective would be to predict which group of outages will occur on a certain day in a specific area. Such clustering is necessary to obtain the best possible predictive performance. Therefore,

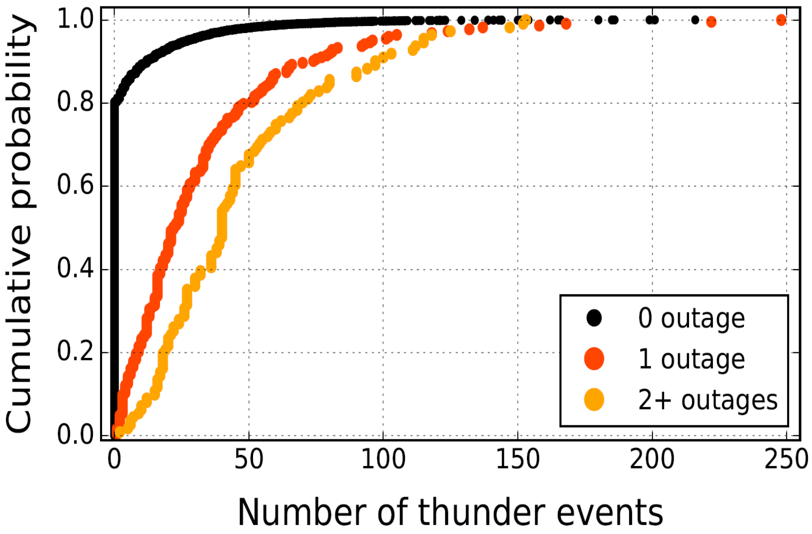


Figure 3.4: Empirical cumulative distribution function of number of thunderstorm events for groups of outages

it can be said that models that attempt to predict the exact number of outages (some of which were mentioned in the literature review), are expected to deliver a low degree of accuracy especially when the number of outages is large. A quantitative result is provided in Section 3.8 for this argument.

3.6 Likelihood of Outages

Based on the preceding methodology, it was concluded that during days with thunderstorm events, there could be three possible outcomes with regards to the number of outages. The next step would then be to determine the likelihood of the occurrence of each group of outages on a given day with a given number of thunderstorm events at a specific area. For this purpose, the proposal is that the binomial distribution model would be an appropriate model to calculate the likelihood of lightning-induced outages. Statistically speaking, the binomial distribution model allows computing the probability of observing a given number of successes when a process is repeated a specific number of trials and the outcome for a given trial is either a success or a failure. In this study, success is the outage occurrence and the trials are the set of thunderstorm events.

The rationale behind the proposed model is three-fold:

1. Each thunderstorm event results in one of two possible outcomes (outage or no outage). This is confirmed by the data.

2. The probability of the outage is the same for each thunderstorm event (because it depends on the geographical characteristics of the area which is expected to more or less remain the same)

3. The thunderstorm events are independent, meaning that the fact that a thunderstorm event results in an outage does not impact the probability of an outage in another thunderstorm event. It is assumed that after an outage, the responsible dispatched crew is able to repair the protective devices that were impacted and therefore they will operate as expected.

The aforementioned properties indicate that the assumption of a binomial model holds; therefore, it is valid candidate model for the purpose at hand.

Using the binomial model, one may calculate the likelihood of the occurrence of each group of outages. In order to clarify this, suppose that the weather forecast for the next day for a specific area demonstrates a total number of n thunderstorm events. Let's assume, by using the historical data, it has been realized that the probability that a thunderstorm event leads to an outage in that area is p. Considering this information, the likelihood of having no outage, having exactly one outage, and finally having two or more outages can be calculated as illustrated in Fig. 3.5.

In order to examine whether or not the assumption of the binomial model is correct, a hypothesis test is devised. The null hypothesis would be that the occurrence of outages arises from a binomial model with the probability of p. The alternative hypothesis would be that the data does not come from a binomial distribution. A test could be based on the differences between the observed and expected numbers of outcomes. If those differences are all small, the data is consistent with the null

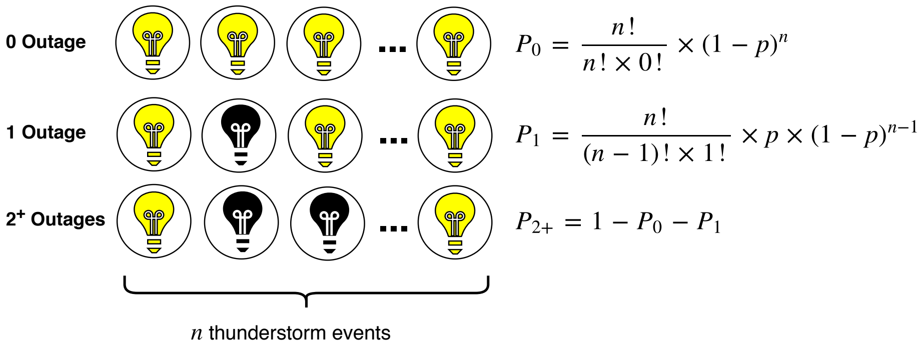

0 Outage
1 Outage
2+ Outages
$P_0 = \frac{n!}{n! \times 0!} \times (1-p)^n$
$P_1 = \frac{n!}{(n-1)! \times 1!} \times p \times (1-p)^{n-1}$
$P_{2+} = 1 - P_0 - P_1$
n thunderstorm events

Figure 3.5: Calculating the likelihood of groups of outages using binomial probability model

hypothesis. If those differences are sufficiently large, either the null hypothesis is false, or an event has occurred that has a small probability.

This could be carried out by the Chi-square test. The Chi-square statistic is a summary measure of how well the observed frequencies of categorical data match the frequencies that would be expected under the null hypothesis that a particular probability model for the data is correct [37]. It is worth mentioning that the critical value of the test is chosen based on 99% confidence interval.

The Chi-square values for all possible values of thunderstorm events are calculated and plotted in Fig. 3.6. For a significant majority of the number of thunderstorm events, the Chi-square values fall below the critical value. This demonstrates that the assumption of the binomial probability model is reasonable. In fact, only 12.8% data points fall out of the range for the 99% confidence level and especially the cases with a high number of thunderstorm events are all represented accurately, with only one data point with over 50 thunderstorm events falling out of the 99% confidence level area. These cases are of special interest for this study, since they also often lead to a case of severe damage in the power system. Now, with great confidence, it can be argued that the binomial distribution would be an adequate model to find the likelihood of any groups of outages (i.e., 0, 1, 2+) occurring given a certain number of thunderstorm events. In order to calculate the likelihood values, the procedure

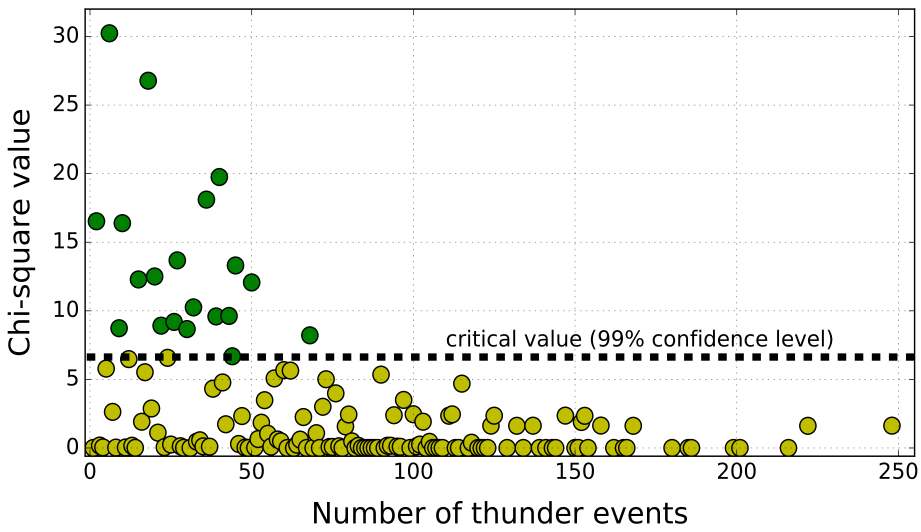


Figure 3.6: Results of the Chi-square test for goodness of fit for binomial probability model

illustrated in Fig. 3.7 is followed.

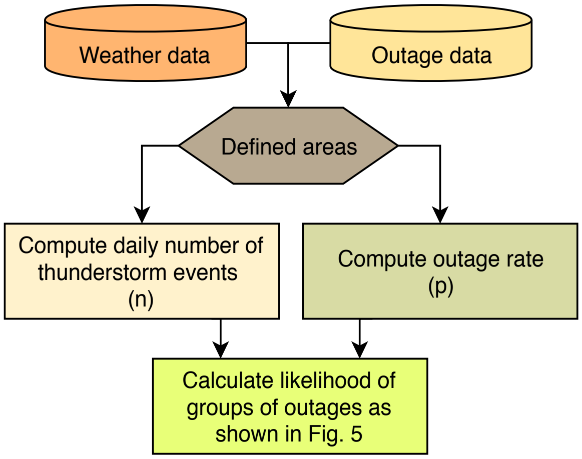

Weather data
Outage data
Defined areas
Compute daily number of thunderstorm events (n)
Compute outage rate (p)
Calculate likelihood of groups of outages as shown in Fig. 5

Figure 3.7: Flowchart of calculating likelihood of outages

For a given day and given area, the number of thunderstorm events is calculated, n, from the weather data. This was explained in Section 3.3. Then, by using the historical data, the probability that a thunderstorm event leads to an outage, p, for each area is calculated. In order to compute this value, the total number of outages that occurred in each area divided by the total number of thunderstorm events experienced by that area is calculated. This probability value is called the outage rate from this point and is demonstrated in Fig. 3.8. As seen in the figure,

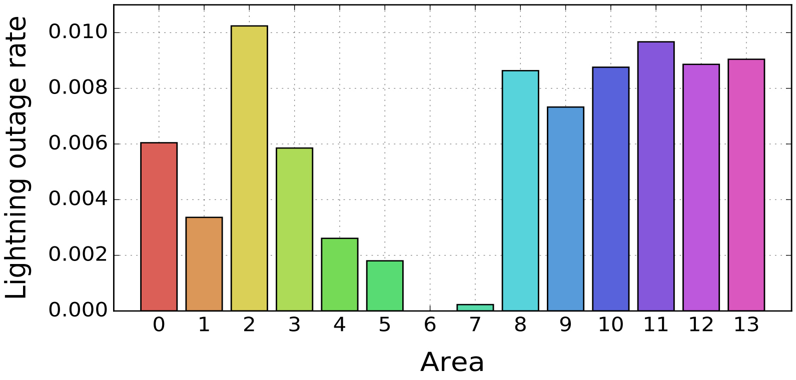


Figure 3.8: Outage rate values for different areas within the system

the outage rate for some areas is considerably higher compared to others. This may be explained by the geographical characteristics of that area and its exposure to lightning strikes. By knowing these values and employing the binomial probability model as illustrated in Fig. 5, the likelihood of each group of outages can be calculated.

3.7 Outage Prediction

By calculating the likelihood of outages for a given day, weather condition, and a given area in the system using the aforementioned methodology, one may make a final prediction on what group of outages will occur. In order to provide a practical means of predicting which outage group will occur, the problem is defined as a machine learning multiclass classification problem. In fact, by using the binomial model, three likelihood values for zero outage, one outage, and two or more outages can be calculated. The actual group of outages is also known from the historical data. As a result, we would have a supervised machine learning problem, in which the variables are the likelihood values, and the label is the class of outages (i.e., 0, 1, 2+). It is worth mentioning that there could be other approaches, such as setting some threshold values and finding whether or not the likelihood values exceed those thresholds to make the prediction; however, those approaches are beset by a number of practical constraints.

With the aforementioned context, the main objective here would be to predict different classes of outages correctly while the minimum number of alarms is issued. An alarm is issued when the model predicts either one outage, or two or more outages. Since the occurrence of lightning-induced outages is not very frequent, the majority of the alarms turn out to be false. As a result, it is crucial to build a classifier that minimizes the false alarm ratio while enabling the outage instances, especially, two or more outages to be detected correctly as much as possible. Therefore, the metric that is utilized to build and evaluate the classifier would be the outage detection rate. This metric, which is also known as recall value, is defined as $\frac{tp}{tp+fn}$. For each class of outage, the tp is the number of true positives (i.e., outages detected correctly) and fn is the number of false negatives (i.e., misclassified outages).

As explained earlier, there is a trade-off between the outage detection rate score for different classes. In other words, if one intends to predict all of the two or more outages correctly (i.e., maximizes the outage detection rate for that class), one might get less accurate results on one outage class and one needs to issue a great number of false alarms (i.e., outage detection rate for two other classes increase). In order to deal with this trade-off problem, the suggestion is setting different threshold values for the outage detection rate of different classes. For example, it could be assumed that the classifier should be able to deliver an outage detection rate of greater than 0.7 for two or more outages and a detection rate of greater than 0.5 for one outage. These values could be customized by the user.

One challenge with regards to the classification problem defined here is the presence of imbalanced classes. The majority class of outages is zero outages. The occurrence of one outage class is significantly smaller and the occurrence of two or more outages is rather infrequent. While the occurrence of one or two or more outages is considerably small, they are of interest to the utility company, and therefore an appropriate model should be able to identify them correctly as much as possible. Such a difference between the occurrence of classes is known as an imbalanced problem. Imbalanced class distribution of a data set is problematic as it can result in biased predictions and misleading accuracy for most classification learners [38]. A quantitative result is provided in Section 3.8 to demonstrate the impact of the imbalanced problem.

In order to address the imbalanced problem, a variety of methods has been proposed. These methods could be categorized under three well-established approaches of data-level, algorithm-level, and cost-sensitive learning. In the data-level approach, different mechanisms such as over-sampling the instances in the minority class, under-sampling the observations in the majority class, and generating synthetic data are employed to re-balance the data. In the algorithm-level approach, a bias is introduced in the objective function of classifiers to give different weights to the majority and minority classes. The rationale behind cost-sensitive learning methods is to evaluate the cost associated with misclassifying the observations. The objective is to take a decision to minimize the expected cost [38].

In this study, the algorithm-level approach is employed to tackle the imbalanced problem. In fact, the level of imbalance is very significant; as a result, data-level approach, especially generating synthetic data, won't be practical. Moreover, the desire is to design a classifier that has the ability to be customized by the user. In other words, it is desired that the user to have the ability to give customized importance to different classes with a flexible degree and desired outage detection rate. This is perfectly possible in the algorithm-level approach.

In order to carry out the classification task, the logistic regression is used as the baseline model. Logistic regression is among the most well-established classifier algorithms. While there could be several other choices; considering the size of the data set, the number of features, and the type of features, logistic regression would suffice for this problem. Especially, the loss function of the logistic regression could be easily modified to tackle the imbalanced problem. It is worth mentioning that some other models such as neural networks have the same ability and therefore the analyst should examine which algorithm suits the problem the best.

In logistic regression, the assumption is that all classes (i.e., primarily two classes) are equally important and hence have the same weight (i.e., importance) and the objective is to minimize the loss function as in (2) [12]:

$$logLoss = -\sum_{i=0}^{n-1}[y_i \cdot log(f(x_i)) + (1 - y_i) \cdot log(1 - f(x_i))] \tag{3.2}$$

where, y_i is the actual class, $f(x_i)$ is the predicted class, and n is the number of observations. However, in a weighted logistic regression, the importance of the classes is weighted so that different classes have different weights associated with them. considering the weight, w, (2) can be re-written as (3):

$$logLoss = -\sum_{i=0}^{n-1}[w \cdot y_i \cdot log(f(x_i)) + (1 - w) \cdot (1 - y_i) \cdot log(1 - f(x_i))] \tag{3.3}$$

It should be noted that logistic regression is a binary classifier, meaning it cannot handle target vectors with more than two classes. To make the multi-class classification possible, the logistic regression should be used in a procedure known as one-vs-all. In this procedure, a separate model is trained for each class to predict whether an observation belongs to that class or not versus to the other classes combined (thus making it a binary classification problem), and making the final decision at the end by looking at the results of all models [39]. Moreover, in order to avoid the over-fitting problem, the $L2$ regularization terms, with regularization rate of λ, [11] are added to the aforementioned loss function.

The weighted logistic regression would be the cornerstone of the classifier. However, to build a robust model, general steps such as data pre-processing, creating training and testing sets, tuning hyper-parameters through cross-validation, etc. have to be performed as well. The complete procedure for building the classifier is demonstrated in the flowchart shown in Fig. 3.9.

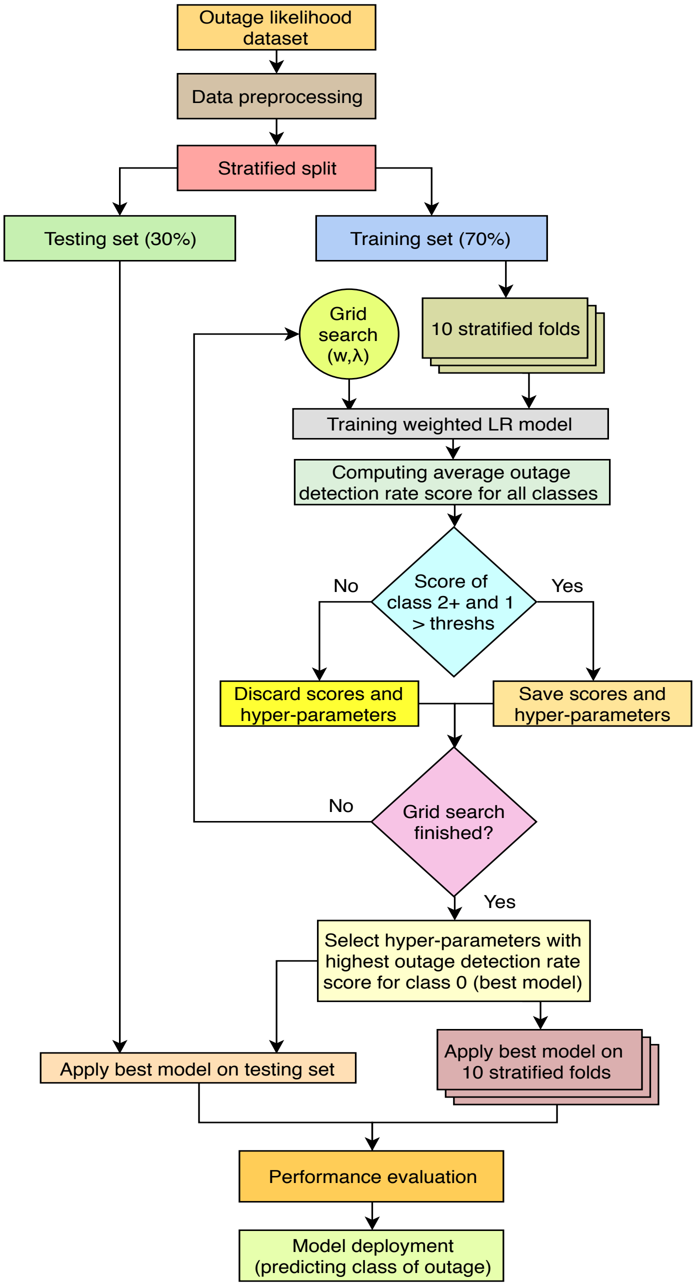

Outage likelihood dataset
Data preprocessing
Stratified split
Testing set (30%)
Training set (70%)
Grid search (w,λ)
10 stratified folds
Training weighted LR model
Computing average outage detection rate score for all classes
Score of class 2+ and 1 > threshs
No
Yes
Discard scores and hyper-parameters
Save scores and hyper-parameters
Grid search finished?
No
Yes
Select hyper-parameters with highest outage detection rate score for class 0 (best model)
Apply best model on testing set
Apply best model on 10 stratified folds
Performance evaluation
Model deployment (predicting class of outage)

Figure 3.9: Procedural flowchart of the proposed classifier

Several points should be made regarding the procedure as follows:

1. Data pre-processing includes handling missing data and outliers and normalizing the data.

2. The data is split into training and testing sets in a way that the proportion of classes in both sets is similar (A.K.A., stratified splitting)

3. The hyper-parameters of the logistic regression model include three weights (w) for classes as well as regularization rate (λ), and are tuned through cross-validation.

4. The threshold value for the outage detection rate for different classes could be customized by the user to satisfy the desired outage detection rates. If the outage detection rate for the critical class (i.e., two or more outages or one outage) is greater than a threshold, the hyper-parameters associated with that outage detection rate value would be stored.

5. The ability of the model to generalize is evaluated using k-fold cross-validation as well as its performance on the testing set.

The proposed algorithm allows predicting outages with a desired detection rate for different classes. The effectiveness of the proposed approach is demonstrated through a case study (third case study) in the next section.

3.8 Case Studies

In order to quantitatively show the practical issues that were discussed with regards to predicting lightning-induced outages and to demonstrate the superiority and effectiveness of the proposed approach, three case studies are provided as follows.

3.8.1 Case study 1 (benchmark results)

As explained earlier, clustering outages into a few manageable groups (three groups in this study) seems necessary to obtain the best possible predictive performance. In fact, it was shown that attempts to predict the exact number of outages could lead to a low degree of performance especially when the number of outages is large. It was also argued that this problem exacerbates because of the imbalanced problem.

In order to show the impact of the aforementioned issues, a case study will be carried out. In case study 1, the outages are not clustered; additionally, the proposed probability model for calculating the likelihood of outages is skipped. Three well-known machine learning classifiers: Random Forest (RF), Naive Bayes (NB), and Logistic Regression (LR) (not weighted) are used and are fed two main inputs of area number and number of thunderstorm events on a daily basis. Then, necessary hyper-parameters are tuned through 10-fold cross-validation. The objective is to predict the exact number of outages (i.e., zero to five). In order to explore the performance of the model, again a 10-fold cross-validation procedure is utilized and the average outage detection rate for each class (i.e., zero to five in this case study) is obtained, in which the results are provided in Table 3.4.

Table 3.4: RESULTS OF CASE STUDY 1

Model / Outage	**0**	**1**	**2**	**3**	**4**	**5**
RF	0.99	0.03	0.2	0.25	0.2	0
NB	0.97	0.2	0.26	0.25	0	0
LR	0.94	0.39	0.27	0.25	0	0

The results clearly highlight the impact of the aforementioned issues. In fact, as seen, the outage detection rate for large values of outages is very low, even in some cases zero. Moreover, due to the imbalanced problem, the models are biased toward the majority class (i.e., zero outage) and therefore deliver very low outage detection rate for minority classes (i.e., outage instances).

3.8.2 Case study 2 (imbalanced problem)

In this case study, outages are clustered to three groups (i.e., 0, 1, 2+) and the likelihood values for each group of outages are calculated. Then, the likelihood values are fed to the three aforementioned classifiers to predict the class of outages. However, in order to show the impact of the imbalanced problem, no action to deal with that problem is taken. Tuning the hyper-parameters and assessing the performance of the model are carried out through the 10-fold cross-validation again. The results (i.e., outage detection rates for three classes) are provided in Table 3.5.

Table 3.5: RESULTS OF CASE STUDY 2

Model / Outage	**0**	**1**	**2+**
RF	0.99	0.15	0.36
NB	0.96	0.31	0.40
LR	0.96	0.17	0.53

As seen in the table, clustering the outages improves the outage detection rates compared to the first case study. However, still, the models are biased toward the majority class (i.e., zero outage) and therefore deliver poor results for one outage and two or more outage classes.

3.8.3 Case study 3 (proposed approach)

In this case study, the proposed approach is implemented and its success in addressing the aforementioned issues is demonstrated. The outages are again clustered into three groups (i.e., 0, 1, 2+) and the likelihood values for each group of outages is calculated. Then, the likelihood values (i.e., outage likelihood dataset) are fed to the proposed classifier illustrated in Fig. 2.9. The threshold values that are considered for the most important class (i.e., two or more outages) is 0.8 and for the second important class (i.e., one outage) is 0.55. This means that the weights are tuned such that it is made sure to obtain those outage detection rate values on the cross-validation. The performance of the model (outage detection rates) is also evaluated on 10-fold cross-validation as well as on the testing set (30% of the whole data), where the results are provided in Table 3.6.

Table 3.6: RESULTS OF CASE STUDY 3

Set / Outage	**0**	**1**	**2+**
CV	0.82	0.57	0.84
Test	0.81	0.55	0.86

As seen in the table, by optimally tuning the weight values (w) for different classes, one is able to obtain outage detection rates that satisfy defined threshold values. As mentioned, there is a trade-off between outage detection rate values of different classes. In this case study, the highest importance is placed to two or more outage class (i.e., outage detection rate of 0.84) and lower importance to one outage class. As a result, some of the one outage instances are misclassified in favor of two or more outages, as the model is intentionally biased towards two or more outages, which represent the severest of outages.

One important observation that demonstrates the remarkable performance of the proposed approach is the outage detection rate obtained for zero outage instances. In fact, even though the classifier is intentionally biased to outage instances, it is able to detect zero outage observations with a high score of 0.82. This means that the number of false alarms issued by the model is significantly small. Another observation that proves the superior performance of the proposed approach is high outage detection rates that are obtained on the testing set (unseen data while developing the model). The values obtained on the testing set are significant and very similar to those obtained on cross-validation, demonstrating that the model is tuned properly. This indicates that the model is not over-fitting or under-fitting and is able to generalize very well.

3.9 Conclusions

A data-driven approach was proposed for predicting lightning-induced outages in power distribution systems on a daily basis. Based on this study, the following conclusions can be drawn.

1. In order to develop a practical approach, various types of information including historical records of outages and climatological variables should be obtained and processed.

2. A key step in building a realistic approach is to adequately define the extent

of the predictions' target area. Aggregating substations and creating broader geographical areas by using clustering algorithms seems a workable solution for this purpose. This helps to reduce the randomness of the number of outages and to obtain more accurate weather information.

3. In order to obtain the best possible predictive performance, lightning-induced outages should be categorized into a few manageable groups. For this study, the groups were zero outage, one outage, and two or more outages. These groups exhibit distinguishable characteristics with regards to the number of thunderstorm events.

4. The binomial probability model is an adequate model to find the likelihood of groups of outages (i.e., 0, 1, 2+) given a certain number of thunderstorm events and a specific area in the system.

5. An important issue that should be addressed to build a successful predictive model is the imbalanced problem. The weighted logistic regression model can handle this problem and can deliver an appropriate classification of different groups of outages.

Although many different pieces of information pertaining to lightning-induced outages were examined in this study, all possible factors were not accounted for due to lack of access to related data. Therefore, the performance of the proposed approach may be improved by the inclusion of additional climatological and geographical information (e.g., satellite data for more accurate identification on thunderstorm events). In fact, all the advantages of the proposed approach are built upon generic outage data collected by utilities, and typical daily weather forecast data, which is publicly available. This fact makes the implementation of the approach easily attainable within a great level of performance. However, the approach provides the flexibility to be improved by utilizing various other sources of data.

The results of this study could be very informative to utility companies in gaining insight about their lightning-induced outage issues, and to build better models for predicting those. Especially, by providing a preliminary but accurate prediction, the proposed approach enables operators to utilize satellite imagery or sophisticated lightning detection systems to find the exact locations in the system that could have high risk of a lightning-induced outage. Moreover, in this study, workable solutions have been proposed to some existing problems and data-driven insights have been produced that can be used by researchers to improve their outage predictive models.

CHAPTER 4: Statistical Analysis of Animal-Related Outages in Power Distribution Systems

4.1 Introduction

Interference from animals is one of the leading causes of outages that occur in power distribution systems. Different species of animals including birds, raccoons, snakes, and especially, squirrels, can come into contact with overhead distribution lines, and cause outages that pose serious challenges to utility companies and customers.

In order to help prevent the adverse effects of animal-related outages, electric utility companies usually take preventive measures that reduce the interaction between animals and overhead distribution lines. This includes designing a resilient distribution system and installing various protective guard devices, to name a few [40]. Despite such measures, animal-related outages occur on a very regular basis and remain a considerable concern for power utilities.

A viable strategy for utility companies to mitigate the impact of animal-related outages is to increase their awareness of patterns of these outages, habits of animals, and ultimately to build predictive models to estimate the frequency of these outages. Acquiring an advance knowledge of these outages and being able to predict their occurrence enable effective and timely decision-making actions, improving the reliability and operational integrity of the system [41].

In spite of the importance of predictive models for animal-related outages and substantial benefits that can be derived from them, developing such models has not been explored to the fullest extent in the past. The main reason for the perceived lack of progress on this critical problem was the limited access to sufficient amount of information pertaining to these outages. However, over the recent decade, with the explosion of data gathering efforts within the smart grid framework and other data acquisition systems, the necessary data for developing these predictive models has become available, resulting in some aspects of this problem being studied [41],[42], [43], [44], [45].

This chapter focuses on developing a data-driven approach for analyzing and predicting animal-related outages on a weekly basis. Although this problem has been considered in some studies before [41], [42], certain assumptions made in those studies, as well as some inputs used to train the proposed models, are unrealistic and can lead to obtaining biased results. Moreover, a fundamental analysis to show the hidden structure of these outages, the relationship between them and the animal population level, time, and weather-related factors has not yet been carried out. With this backdrop, this chapter aims at performing such rigorous analysis and providing a workable predictive model through a case study.

4.2 Problem Description

4.2.1 Objective

The main objective of this chapter is to propose an approach for analyzing animal-related outages and predicting them on a weekly horizon (preferably one week ahead). This is demonstrated through a case study. The majority of animal-related outages that occur in the area under study are caused by squirrels. As a result, the focus is only on this species and the data that pertains to them is utilized. If the necessary information for other animal species becomes available, this study could be expanded to build a predictive model for those animals as well.

One important factor for conducting the prediction task is the prediction horizon. As a matter of fact, performing the prediction for a very short-term horizon (i.e., hourly, daily) or very long-term horizon (i.e., yearly) may not be realistic nor practical. Moreover, defining the extent of the predictions' area (i.e., single substation, city, county, etc.) is a critical problem, which has been comprehensively discussed in [46].

In this study, the area under study is broad and includes regions in the states of North and South Carolina. However, if a considerable amount of data is available, the approach can be applied to carry out the prediction for smaller areas as well.

4.2.2 Approach

In order to build a predictive model for animal-related outages, it is essential to first analyze the structure of these outages. In particular, it is necessary to explore their trend, seasonal patterns, time dependency, and demonstrate that their occurrence is not random (albeit some degree of randomness may exist). This analysis is vital as it suggests that utilizing merely simple probabilistic approaches (i.e., fitting a specific distribution) is not enough to explain these outages and therefore more sophisticated models are required.

Once the aforementioned analysis is carried out, the factors that have influence on the occurrence of these outages ought to be identified and explored. In particular, the relationships between these outages and animal population size, season, and several weather-related factors is examined. The importance of these factors is shown and these factors are utilized to create the inputs for the predictive model.

Finally, a dynamic regression model is developed and it is demonstrated how the model can be utilized to predict the number of animal-related outages. Various statistical tests are carried out to validate the assumptions of the dynamic model. Also, the performance of the model is evaluated and is compared with two conventional machine learning models. Details about dynamic regression are provided in Section 4.5.

4.3 Data Description

The input data for the proposed approach is obtained from three main sources: 1) historically recorded outages, 2) wildlife resources, and 3) weather-related data. The power systems outage data is collected by Duke Energy - a major investor-owned utility company in the US. The data includes information on the exact time of sustained animal-related outages that occurred in the states of North Carolina and South Carolina between the years 2010 and 2014. The wildlife resource data is obtained from the North Carolina Wildlife Resources Commission and includes the hunter harvest survey estimates of various animal species for the aforementioned time span. The weather data is gathered from the NASA Langley Research Center's (LaRC) POWER Project funded through the NASA Earth Science/Applied Science Program. The weather data contains information on daily averaged data for the aforementioned date ranges and three parameters of temperature, wind speed, and precipitation for 100 geographical points within the area under the study. More discussions about the data will be provided in the upcoming sections.

4.4 Investigation on Animal-elated Outages and Influential Factors

4.4.1 Analyzing the structure of the outages

Fig. 4.1, top plot, demonstrates the number of animal-related outages that occurred in the area under study on a weekly basis. As seen, the average number of outages decreases over time. A simple linear regression model is fitted to show the negative trend ($Slope = -0.0526, t = -4.736, P = 0.00$). Moreover, analyzing the data reveals that the variation within the occurrence of these outages is not constant over time (Levene's test: $W = 2.99, df = 3, P = 0.03$). These observations suggest that some factors might have influenced the occurrence of the outages, which requires further investigation.

The next question that should be answered is whether or not the outage data shows a random behavior. In order to answer this question, the structure of the lag plot provided in Fig. 1, bottom left is investigated. The plot demonstrates that the data points tend to cluster (although noisily) along the diagonal. Such clustering suggests a moderate auto-correlation, and consequently a non-random behavior. To quantitatively check this assumption, a runs test is performed, wherein the result

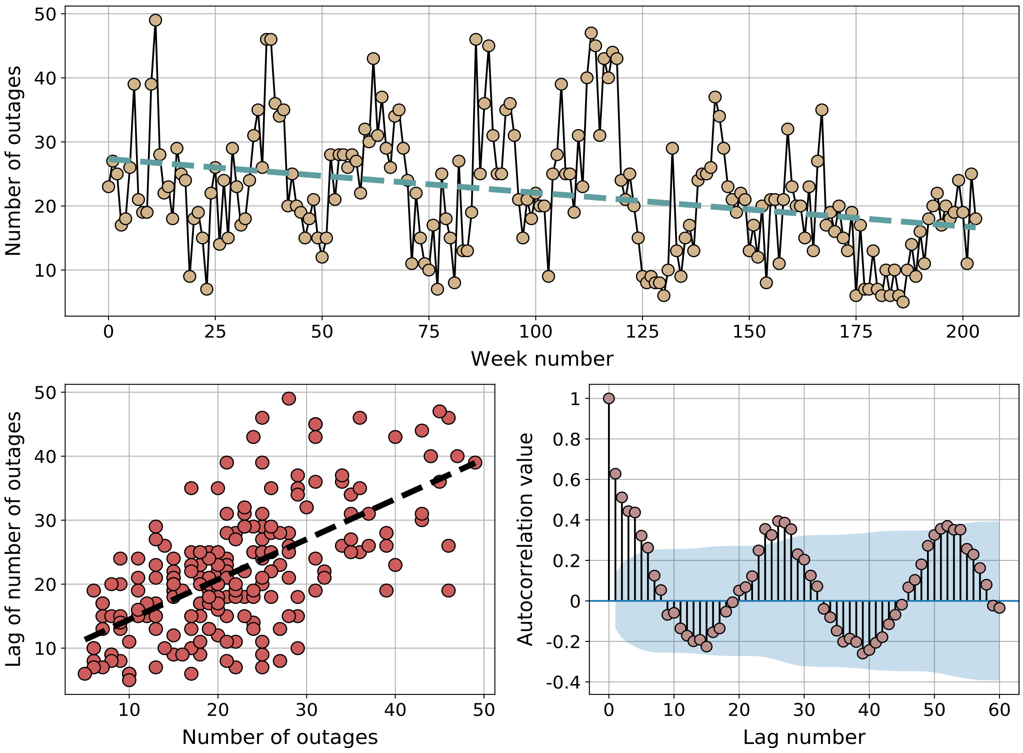


Figure 4.1: Exploratory data analysis of outages

$(Z = -6.42)$ confirms a non-random behavior. This means that these outages do not occur in a totally random manner, proving that a time-dependency exists in their occurrence.

In order to further examine the time-dependency within the data, the auto-correlations at varying time lags and demonstrate those in Fig. 1, bottom right plot is calculated. Based on the plot, the data has moderate auto-correlation, again confirming the non-random behavior in the data. However, the main observation is that the data exhibits a seasonal pattern (i.e., significant values at lags 26, 52, etc.). This seasonal behavior is important and should be explored further.

4.4.2 Exploring influential factors

4.4.2.1 Animal population

A critical factor that influences the number of animal-related outages is the population size of the animals (squirrels in this study) over time in the area under study.

Finding a very accurate data that shows the population level of squirrels is an extremely challenging task and would be beyond the scope of this study. Hence, this factor should be estimated. In order to estimate the squirrel population level, the hunter harvest surveys obtained from the North Carolina Wildlife Resources Commission are used. The survey shows the average number of squirrels harvested by hunters on an annual basis for the state of North Carolina. The state of South Carolina shares almost the same geographical characteristics and therefore it is assumed that the estimates for that state would be the same. The survey values are provided in Table 4.1. Such numbers would be a reasonable indicator of the population level of squirrels, wherein the years with higher harvest per hunter value had higher squirrel population level.

Table 4.1: Harvest per hunter for each year

Year	2010	2011	2012	2013	2014
Harvest / Hunter	8.33	8.59	8.56	7.81	7.52

The number of outages on an annual basis and the squirrels' population for those years are strongly correlated (Pearson $r = 0.98, n = 5, P = 0.003$). In order to further demonstrate the relationship between population and the number of outages, first look at the distribution of the number of outages for different population levels, shown in Fig. 4.2 right plot. As seen in this plot, as the population level increases,

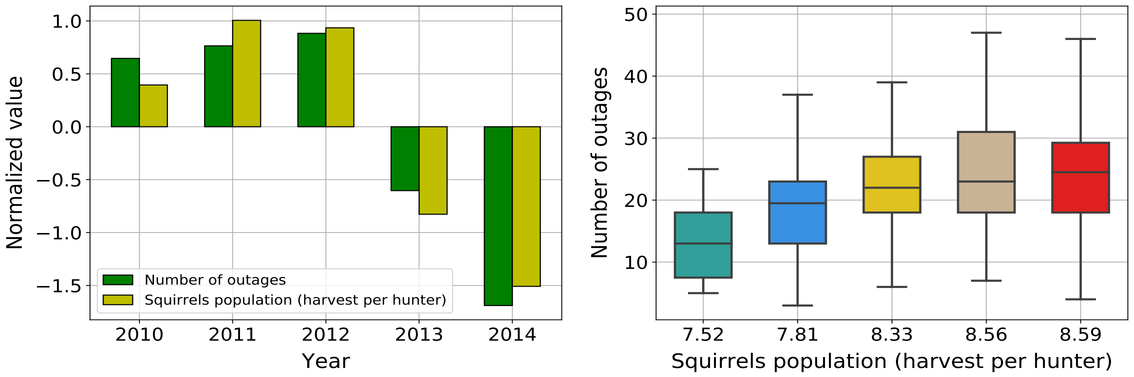


Figure 4.2: Relationships between squirrel population level and number of outages

the range and the median number of outages increases. Obvious differences between these box plots suggest that the population has a significant impact on the number of outages. Second, the normalized values (z-score) for the number of outages that occurred at each year and the corresponding population are compared in Fig. 2, left plot. According to the plot, the change in the population is highly correlated with the change in the number of outages. In fact, during the years 2013 and 2014, where a decrease in population is observed, a decrease in the number of outages could be seen as well. The negative change in the population over the last two years may explain the negative trend and smaller variation in the number of outages that was discussed earlier.

4.4.2.2 Season

As discussed earlier, animal-related outages are the result of the destructive interaction between animals and power distributions lines. Hence, during the time periods in which the animal activity increases, it is expected the utility company would experience a greater number of outages. Squirrels are typically active in all seasons [47]; however, their activity increases significantly during fall and spring months due to the search for food and breeding activities. On the other hand, during winter and summer months their activity decreases because of the temperature and gestation procedure.

In order to understand the impact of this seasonal behavior on the number of outages, some fluctuations of the top plot in Fig. 4.1 are removed by calculating the moving average values ($2 \times 8 - MA$). The resultant plot is illustrated in Fig. 4.3. As seen in the figure, each year, the number of outages experiences two peaks during the spring and fall, and two valleys during the winter and summer seasons. This observation perfectly matches the activity level of the squirrels during different seasons. Also, such observation explains the seasonal pattern captured in Fig 4.1, bottom right plot. Therefore, the season is a critical factor while building a predictive

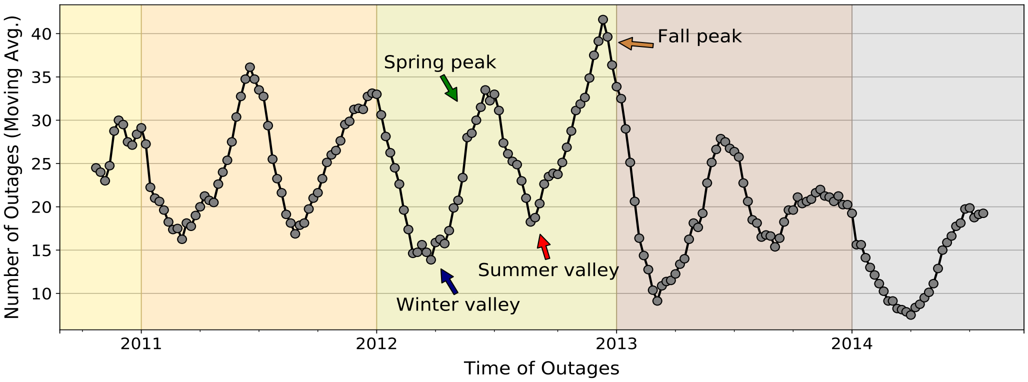


Figure 4.3: Smoothed number of outages over time

model for animal-related outages, as the difference in the number of outages between different seasons is extremely significant (Kruskal-Wallis $H = 53.65, df = 3, P = 0.00$).

4.4.2.3 Weather

In addition to the population level and season, the activity of squirrels and subsequently their interactions with power systems could be influenced by weather-related factors. The impact of various weather factors on squirrels has been investigated in great details [48]. In particular, according to many sources, it has been shown that high temperatures, strong winds, and precipitations can affect the behavior of squirrels and make them seek shelter and consequently become less active.

In order to explore how weather-related factors influence the activity of squirrels and the number of outages, different weather condition scenarios are created. It is worth mentioning that based on several analyses, it has been concluded that small changes in weather do not have much impact on activity of squirrels and the number of outages, and hence the extreme conditions should be considered. In order to define the scenarios, first, each day is categorized according to three factors: temperature, wind speed, and precipitation amount as follows:

1. Hot Day: 1 if temperature > 80 else 0

2. Windy Day: 1 if wind speed > 3.3 (m/s) else 0
3. Rainy Day: 1 if precipitation > 2.5 (mm/hr) else 0

Regarding this categorization, it should be noted that the weather factors are the average daily value. Also, the margins are defined based on values reported in several weather resources. It is worth mentioning that the selection of these values differs from region to region, and hence a careful study should be carried out by the analyst to find the proper values.

In order to determine whether the average number of outages for each condition statistically differs from each other, two-sample t-tests are performed. The results (Hot Day: $T = 2.09, P = 0.036$, Windy Day: $T = 4.11, P = 0.00$, Rainy Day: $T = 3.50, P = 0.00$) confirms the aforementioned assumption. This means that the average number of outages that occurred for example on windy days significantly differs from that on non-windy days and so on.

After this categorization is done, weather condition scenarios for each day are created. In fact, each day could take a position from 8 ($2 \times 2 \times 2$) possible positions as shown in Table 4.2. The average number of outages for each weather condition is calculated and provided in Table 4.2 as well. As seen in the Table, when the weather condition is fine (i.e., all values are 0), the activity of the squirrels is at the highest and, subsequently the average number of outages takes its highest value. On the

Table 4.2: Average number of outages for different weather conditions

Hot Day	**Rainy Day**	**Windy Day**	**Average Outage**
0	0	0	3.54
0	0	1	2.63
0	1	0	3.10
0	1	1	2.35
1	0	0	2.86
1	0	1	2.20
1	1	0	2.63
1	1	1	2.0

other hand, on harsh weather condition (i.e., all values are 1), the activity of squirrels is less and therefore the average number of outages takes its minimum value. These observations demonstrate how harsh weather conditions could affect the number of animal-related outages.

4.5 Prediction with Dynamic Regression Model

As shown in the previous section, a time-dependency and seasonal behavior could be observed in the occurrence of animal-related outages. Moreover, their frequency is influenced by wildlife resources as well as weather conditions. Therefore, in order to effectively predict their occurrence, a model is required that can capture the time-related patterns, and is able to explain the variations that are caused by wildlife and weather factors. Considering such requirements and the fact that the size of the dataset in this study is small, a dynamic regression model is the most appropriate approach.

A dynamic regression [49] is an extended ARIMA model with the inclusion of external factors. In order to explain the formulation of the dynamic regression model, first, a linear regression model as in (2) is considered.

$$y_t = \beta_0 + \beta_1 x_{1,t} + ... + \beta_k x_{k,t} + \epsilon_t \tag{4.1}$$

where y_t is a linear function of k explanatory variables ($x_{i,t}$) and ϵ_t is the error term, which usually is a white noise. Contrary to this formulation, in a dynamic regression model, the error term from the regression can contain the auto-correlation and follow an ARIMA model. Therefore, if the regression error term is defined as η_t and is assumed to follow an ARIMA(1,1,1) model, the dynamic regression formulation would become [11]:

$$y_t = \beta_0 + \beta_1 x_{1,t} + ... + \beta_k x_{k,t} + \eta_t \tag{4.2}$$

where:

$$(1-\Phi_1 B)(1-B)\eta_t = (1+\Theta_1 B)\epsilon_t \tag{4.3}$$

where Φ_1, B, Θ_1 are ARIMA parameters and ϵ_t is a white noise. Considering this formulation, the objective is to estimate the ARIMA model parameters and regression coefficients (β_i) such that the sum of squared ϵ_t values is minimized [50].

As explained, the prediction task is carried out on a weekly basis; hence, the y_t is the total number of outages that occurred in each week. The explanatory variables that will be used as inputs are squirrel population level, seasonal impact, and weather conditions.

The population is obtained on an annual basis; hence, each week in a year will take the same value. For making the prediction for future, the population size could be obtained from wildlife resources or, if not available, could be forecast. Seasons and weather conditions are categorical variables. In order to include them in the model, the data is examined on a daily basis, and it is determined on which combinations of season and weather conditions that day would fall, the expanding average of outages for that combination is calculated, and finally, the summative value of all days in a week is computed.

Fig. 4.4 is provided to clarify the aforementioned procedure. Consider a day in

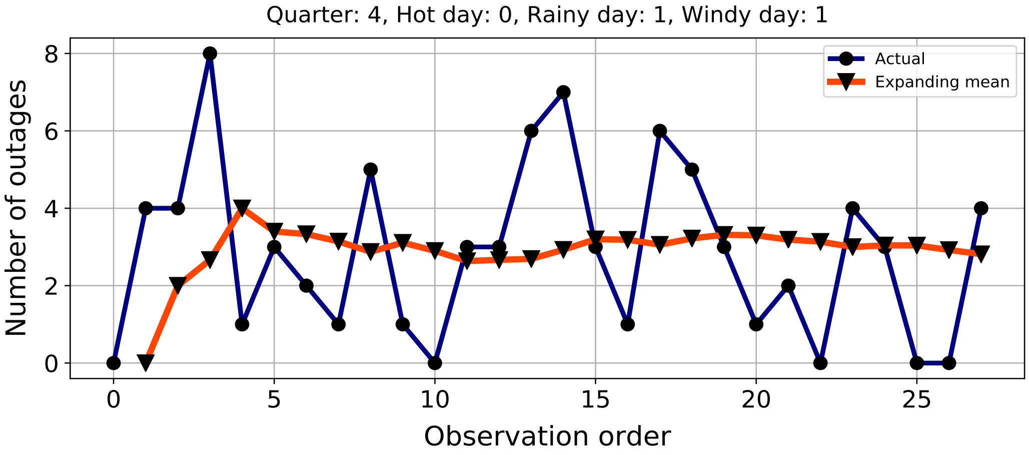


Figure 4.4: Feature engineering for season and weather conditions

the fall season that has the weather condition as Hot Day = 0, Windy Day = 1, and Rainy Day = 1. previous days that have exactly the same conditions are found, and the average value for them is calculated, and the resultant value is assigned for the day under the study. Actual and expanding average values for all days that have the aforementioned combination are shown in the figure. Then, the procedure is repeated for the other days in the week, and are added to find the weekly value. This results in a continuous input that has a considerable correlation with y_t (Pearson $r = 0.6, n = 204, P = 0.00$). This feature engineering is necessary for this study as the number of samples is small and hence inclusion of all categorical variables might lead to the over-fitting problem.

After the explanatory inputs are defined, the parameters of the model (i.e., ARIMA parameters including the seasonal terms and also regression coefficients) are estimated. In order to conduct this task, rolling based training and testing sets are created. At each step, different sets of parameters are considered, a grid search is performed, the sum of squared ϵ_t on the validation set is calculated, the combination of parameters that deliver the lowest error is selected and used to make the prediction on the test set. This procedure is started from the 53^{th} observation and is moved in steps of 1. At the end, the combination of parameters that results in the lowest error value on average in validation sets is selected as the optimal parameters. Considering the small size of the dataset, this procedure, which is known as k-fold cross validation, prevents over-fitting problem and helps demonstrate how the model results could be generalized.

The prediction results for the best model (Regression with $ARIMA(2,0,1)(1,0,0)_{52}$ errors) is plotted against the actual number of outages in Fig. 4.5, top plot. As seen in the figure, the model is able to capture the seasonal patterns and trends very well. Also, the model is able to capture some variations that might have been caused by weather factors. However, some unexpected variations in the data is not captured

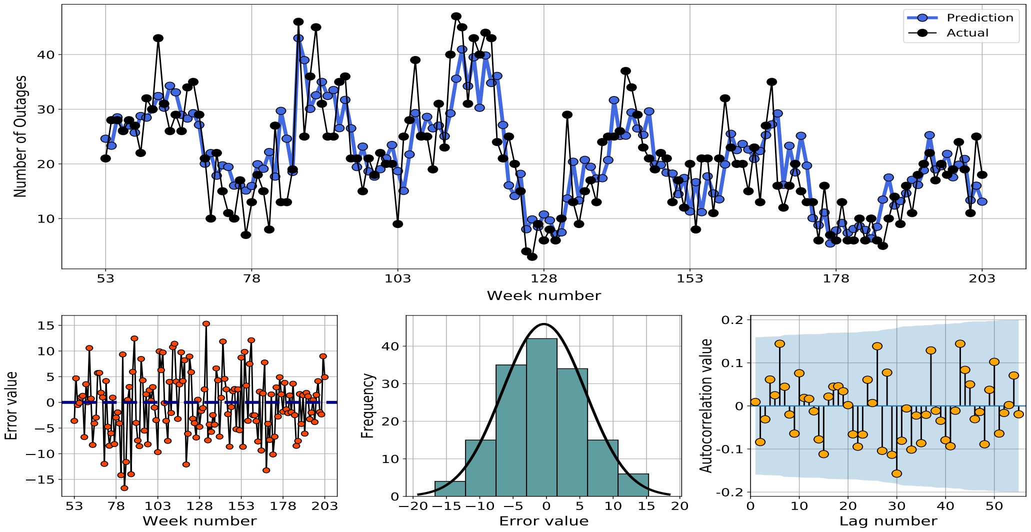


Figure 4.5: Demonstration of forecast and error terms

by the model (these may be random variations). As mentioned, the ϵ_t error terms of the dynamic regression model should be random and uncorrelated (i.e. white noise). Checking this assumption is necessary to validate that the model is appropriate. As a result, an error analysis is carried out. The results are presented in Fig. 4.5, bottom plots. As seen in the bottom-left plot, the average error value is close to zero and the variations stay much the same over time. The histogram plot shows that the error terms follow a normal distribution. Moreover, the auto-correlation plot is provided to demonstrate that there is no auto-correlation and subsequently non-random pattern in the error, as all values are close to zero. Therefore, these observations confirm that the error of the model satisfies the requirements and the model is an appropriate fit.

The RMSE and MAPE values for the prediction are 6.22 and 29%, respectively. Considering the range of the number of outages, which is around 45, and the fact that the randomness in the behavior of animals and occurrence of outages can be considerable, the model performance is remarkable.

In order to show the effectiveness of the proposed model, its performance with other conventional methods is compared. As a result, two popular models are selected, namely ridge regression and K-Nearest Neighbor algorithms, and the aforementioned explanatory variables as well as various time-related factors are used as inputs to those models and their hyper-parameters are tuned. Afterward, four different testing periods are selected from the data to compare the results of those two models with the dynamic regression model. The RMSE values for all methods are provided in Fig. 4.6.

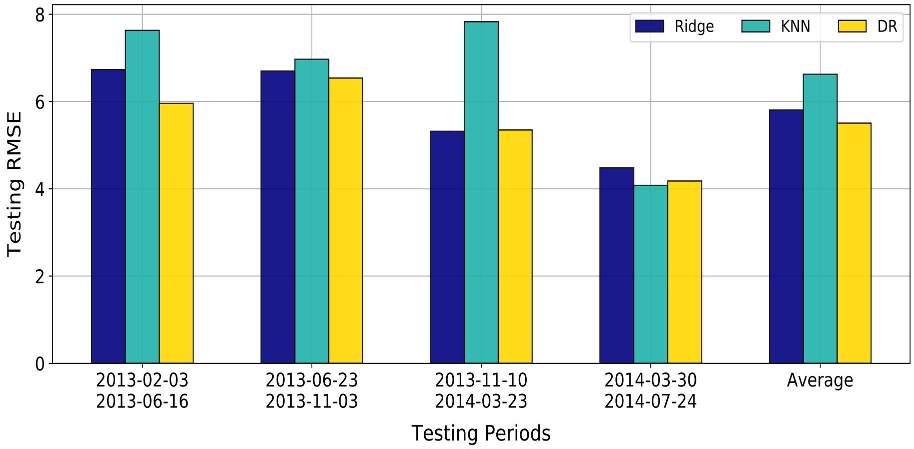


Figure 4.6: RMSE values for different models and testing periods

It is worth mentioning that for each testing set, the training set is the data that falls prior to them. As seen in the figure, generally, the dynamic regression model (DR) either outperforms the other models or delivers a very similar performance. On average, the dynamic regression model delivers the lowest RMSE value, confirming that the proposed approach is suitable for this dataset and effective to predict the number of animal-related outages.

4.6 Conclusions

A data-driven approach was presented for exploring animal-related outages and building a proper model to predict their occurrence on a weekly basis. Based on this study, the following conclusions can be drawn.

1. The occurrence of these outages (average, variation) changes over time due to influences of various factors.

2. These outages exhibit timely and seasonal behavior.
3. Population size of animals that cause outage has a significant relationship with the number of outages.
4. The average number of outages varies significantly during different seasons.
5. Extreme weather conditions can influence the occurrence of these outages.
6. The dynamic regression model is an effective predictive model that can take the aforementioned factors into account and provide a reasonable prediction for animal-related outages on a weekly basis.

CHAPTER 5: Power Distribution System Outage Root Cause Analysis by Using Association Rule Mining

5.1 Introduction

5.1.1 Motivation

Outages in power distribution systems can seriously endanger the system operation in different ways. As a matter of fact, distribution outages negatively influence the system reliability since they are responsible for a considerable number of major interruptions that customers experience [51]. Furthermore, outages exert damaging impacts on system safety and security and result in heavy costs for distribution utilities [52]. Therefore, utilities either seek to find practical solutions aimed at preventing specific outages or attempt to take effective measures to properly and quickly restore the system after outages occur. For attaining either, it is essential to acquire a deeper understanding of the primary causes of outages and to identify significant variables related to those causes.

Power distribution utilities are usually interested in preventing avoidable outages. One main course of action to fulfill this objective is to introduce necessary modifications to system outage management based on the knowledge acquired from outages that have occurred in the past. For instance, if it turns out that a distribution system had experienced most of its vegetation-related outages during the months of June and July, then the utility will recognize the need to carry out essential preventive maintenance for those two months. In addition to modifying outage management practices, such knowledge is highly beneficial for improving the design of existing distribution systems to reduce the number of future outages [53]. In order to gain this knowledge, it is required to carry out an in-depth root cause analysis for different outages.

Nevertheless, most often, distribution outages have proved to be unavoidable. For example, even with the implementation of different preventive measures such as installing animal guards or designing large clearances between phase and ground wires, substantial numbers of animal-related outages occur in the system. In these cases, distribution utilities attempt to take an appropriate response to the outages either by predicting it or identifying the causes immediately after the outage. The task of predicting or identifying outages has been considered extremely challenging due to the random nature of outages and the numerous contributing factors. However, over the recent years, with the explosion in data gathering within the smart grid framework, applications of advanced data analytics techniques combined with the traditional rigorous mathematical modeling have facilitated the tasks of outage prediction and identification [54]. In fact, these applications have been identified as viable technologies with the capability to provide necessary actionable information support for utilities to predict or identify different outage with satisfactory accuracy [55], [56], [57]. For the purpose of developing these models, however, it is crucial to identify and utilize the factors that are related to each outage cause. For example, the authors in [56] consider six features of: weather condition, season, time of day, faulty phases, protective device activated, and the location where the outage happened as the most influential factors for vegetation and animal-related outages and use them as inputs to build a power distribution outage cause identifier.

Consequently, characterizing outages according to their underlying causes and identifying significant variables that strongly impact the outage frequency are extremely valuable as they allow utilities to find solutions to restrict specific causes and to give an appropriate response to unavoidable outages.

5.1.2 Literature Review

By this time, several studies have been conducted to analyze the characteristics of various outages, which have been caused primarily by animals or vegetation-related issues. These studies mainly take advantage of statistical techniques and data analytics algorithms.

For instance, the authors in [53] propose a statistical data mining approach to perform a root cause analysis of animal-related outages. In their work, the impact of six factors of weather condition, season, the day of the week, time of day, outagey phases, and protection device activated is considered to be significant on the frequency of such outages. In another attempt, by utilizing four statistical measures of actual, normalized, relative, and likelihood, the authors in [58] investigate tree-related outages with respect to six factors: weather condition, season, time of day, the number of outagey phases, location and the clearing device. Furthermore, the same authors in [59] use logistic regression to explore the influence of the aforementioned factors on tree outages by evaluating significance levels resulting from regression. In their study, the weather condition is observed to be the most influential factor. In [60], the authors review two statistical methods, namely hypothesis test, and stepwise regression and introduce two new methods to select proper features for identifying root causes of different outages. They employ these methods to examine six factors explained in [56].

5.1.3 Contributions

Although basic statistical analysis provides a general understanding of the primary causes of outages, it falls short of describing nuanced conditions that lead to a outage. In fact, gaining a deeper understanding of the contributing factors for each outage type by using conventional statistical methods could become extremely time consuming as it requires performing various analyses. On the other hand, applying sophisticated methods such as stepwise and logistic regression or artificial neural networks can produce deeper insight into the underlying causes; however, it would be computationally burdensome and might require a tremendous amount of running time [60]. The reason for the extra computation time requirement is that a substantial number of inputs, which are based on the many attributes that are involved in the study, have to be generated and fed into these methods.

In order to overcome the aforementioned problems, this study proposes an approach for outage cause analysis based on association rule mining. Association rules, introduced by [61], belong to a category of uncomplicated but remarkably powerful regularities in binary data. They initially were employed to mine large collections of basket data type transactions for association rules between sets of items. However, due to their proven capabilities for finding interesting associations, or correlations among data items, they have received increasing attention in different fields for practical applications. The main advantage of association rule mining over the aforementioned methods is that it can easily analyze the co-occurrence of different attributes to find any association or correlation. Also, since association rule algorithms were originally developed to be applied to extremely large transaction data sets, they are very effective with regard to the computation time requirement. In fact, as will be demonstrated in the following sections of this study, the association rule mining is capable of extracting comprehensive patterns from outage data sets within a short amount of time. To implement it, however, in-depth data preparation is required. Moreover, there are several practical issues associated with outage data sets that ought to be addressed. Therefore, this study provides a step-by-step procedure that fully deals with necessary data preparation, practical issues associated with outage data sets, and implementation of association rule mining. Furthermore, this procedure is applied to investigate a real-world outage data set. As a result of the case study, causal structures and frequent patterns for vegetation, animal, equipment failure, public accident, and lightning-related outages, which have drawn less attention previously in the literature are explored.

5.1.4 Study Organization

This chapter is organized as follows. Association rule mining is discussed in Section 5.2. In Section 5.3, the problem of data insufficiency is explained, and a practical solution to deal with it is provided. The proposed methodology is elaborated in Section 5.4. Section 5.5 provides a case study to illustrate the implementation of the proposed methodology. Finally, Section 5.6 draws conclusions on the effectiveness of association rule mining in outage cause analysis.

5.2 Association Rule Mining

As mentioned earlier, association rules were initially introduced to mine large collections of basket data type transactions to investigate how items or objects are related to each other. However, nowadays, because of their effectiveness, they are widely employed in different domains to mine for causal structures and to identify frequent patterns in various data sets.

An association rule is a causality, where a rule is defined as an implication of the form $X \Rightarrow Y$, with two conditions of $X, Y \subseteq I \ \ and \ \ X \cap Y =$, where I is a set of n binary attributes called items [61]. The itemsets X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule, respectively.

Given the set of transactions, numerous rules can be generated. However, the rules that are of actual interest to data analysts, which provide useful information, have to fulfill certain constraints. The major constraints related to the "support" and "confidence" of a rule [61]. Support determines how often a rule is applicable to a given data set, while confidence determines how frequently items in Y appear in transactions that contain X. Support and confidence can be mathematically expressed as (1) and (2), respectively.

$$Support(X \Rightarrow Y) = \frac{\sigma(X,Y)}{N} \tag{5.1}$$

$$Confidence(X \Rightarrow Y) = \frac{\sigma(X,Y)}{\sigma(X)} \tag{5.2}$$

where, σ is summation notation, and N represents the total number of all transactions.

In this formulation, the association rule problem is usually decomposed into two main problems as follows [11].

- Generating frequent itemsets: The first problem is to discover the itemsets whose occurrences surpass a minimum support. These itemsets are called frequent or large itemsets;

- Generating rules: The second problem is to generate association rules from those large itemsets with the constraint of minimal confidence.

While handling the second problem is straightforward, addressing the first problem can be challenging. Actually, finding all frequent itemsets in a database is difficult since it requires searching over all possible itemset combinations [62]. Therefore, different techniques have been proposed to effectively deal with this issue. One of the most well-established techniques with this regard is the Apriori algorithm. Apriori is the first association rule mining algorithm that utilizes the support-based pruning for the purpose of controlling the exponential growth of candidate itemsets systematically [62]. This algorithm is employed in this study to find frequent itemsets and to generate rules based on confidence.

In order to rank the resultant rules from the Apriori algorithm, there are different measures. As mentioned, support and confidence are the primary metrics to evaluate the quality of the rules generated by the algorithm as they indicate the statistical significance and the strength of a rule, respectively. In addition, there are other metrics, one of which, is "lift" value of a rule. The lift value can be stated as (3)

$$Lift(X \rightarrow Y) = \frac{Support(X \Rightarrow Y)}{Support(Y) \times Support(X)} \tag{5.3}$$

Lift value is an indicator of the strength of a rule over the random co-occurrence of the antecedent and the consequent [63]. A lift ratio larger than 1.0 implies that the relationship between the antecedent and the consequent is more significant than when the two sets are independent. The larger the lift ratio, the more significant the rule.

5.3 Generating Synthetic Data

Over recent years, with the deployment of an enormous number of intelligent electronic devices and various sensors, rich data for studying different outages have become available [64]. However, having access to sufficient amounts of data for pattern recognition is not always possible. Furthermore, a significant challenge with real-world outage data is the lack of adequate information relating to specific types of outages as their occurrence is not frequent. For example, only 3% of outages occurred in and around Charlotte, North Carolina region, between the years of 2009 and 2014 were due to lightning. Hence, if one wants to find patterns associated with lightning outages, one might face the insufficient data problem. In fact, such inadequacy in data can pose serious problems to the efficient operation of association rule mining algorithms as it leads the algorithm to fail to generate requisite large itemsets.

In order to address the aforementioned issue, one practical solution is to generate synthetic data based on the available real data and add them to the dataset [65]. However, this task should be carried out such that hidden complex patterns within the data set are preserved as much as possible. In order to comply with this requirement, synthetic data can be generated based on the Synthetic Minority Over-sampling Technique (SMOTE). This method, introduced in [65], is of particular interest to analysts due to its simplicity and effectiveness. SMOTE has been tested on different datasets, with varying degrees of imbalance and varying amounts of data in the training set, successfully providing a diverse testbed [66]. For the purpose of creating new artificial samples, this method uses the k-nearest neighbor algorithm and bootstrapping [67]. The process for generating artificial data based on this method is illustrated in Fig. 5.1. As seen in the figure, the procedure includes four steps as follows [65].

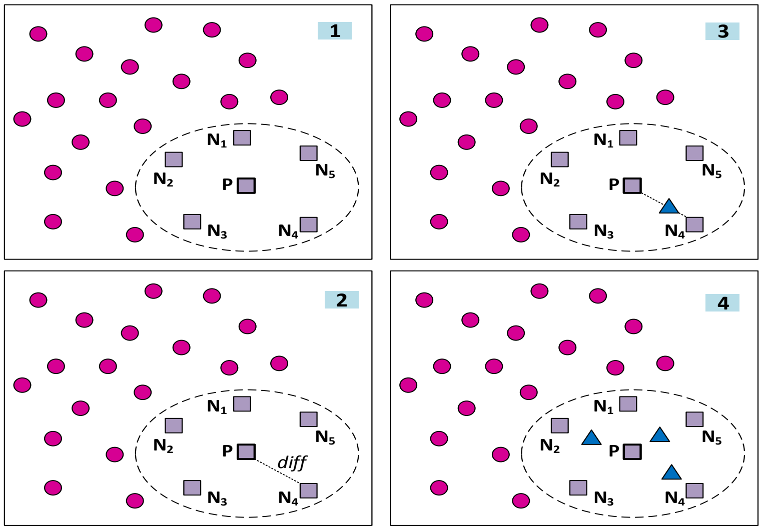


Figure 5.1: SMOTE algorithm

- Step 1: For each minority sample P, its k neighbors (shown with squares) are found (5 neighbors in this example), and an example out of k neighbors (N_4) is randomly chosen;

- Step 2: The difference between the sample under consideration and its selected neighbor is taken;

- Step 3: The difference is multiplied by a random number between 0 and 1 to generate synthetic data (shown with triangles);

- Step 4: The generated synthetic data are added to the dataset.

Synthetic data that are generated by this method mimic the original observed data and preserve the relationships between variables; consequently, they are considered as one feasible solution to the problem of data insufficiency.

5.4 Proposed Approach

The structure of the proposed approach is illustrated in the flow chart provided in Fig. 5.2. In what follows, each block of the flow chart is described in detail.

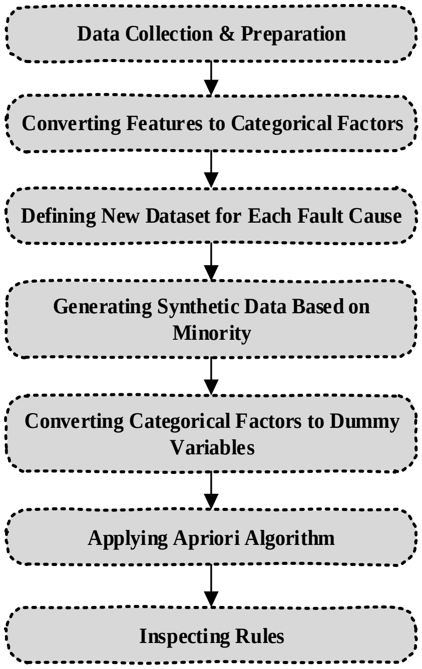

Data Collection & Preparation
Converting Features to Categorical Factors
Defining New Dataset for Each Fault Cause
Generating Synthetic Data Based on Minority
Converting Categorical Factors to Dummy Variables
Applying Apriori Algorithm
Inspecting Rules

Figure 5.2: Technical flow chart of the proposed approach

1. *Data Collection and Preparation*: An essential step in performing outage analysis for pattern recognition is to collect and prepare a considerable amount of outage data. In a typical outage data, one column is assigned to the outage cause (i.e. vegetation, animal, etc.) and the rest of the columns are specified for the features. Such data can be gathered from various sources such as a utility companyâs electronic devices, sensors, weather stations, and etc. However, as

mentioned earlier, occasionally having access to a sufficient amount of data is not possible; therefore, the lack of data ought to be artificially compensated.

In order to effectively gather and prepare data, this task is split into different categories as follows.

1.1. *Obtaining data for various features*: The most vital part of this task is to obtain data for as many features as possible. In order to analyze a specific outage, a large amount of information pertaining to that outage ought to be available. In fact, without having necessary features that make significant contribution to a specific outage, understanding the characteristics of that outage would be practically impossible.

1.2. *Data cleansing*: Real-world outage data sets usually contain a significant amount of input errors, duplicate data, outliers, extreme and unexpected values. These errors can perturb the operation of data mining algorithms. Therefore, the raw data should be explored to identify and remove these errors.

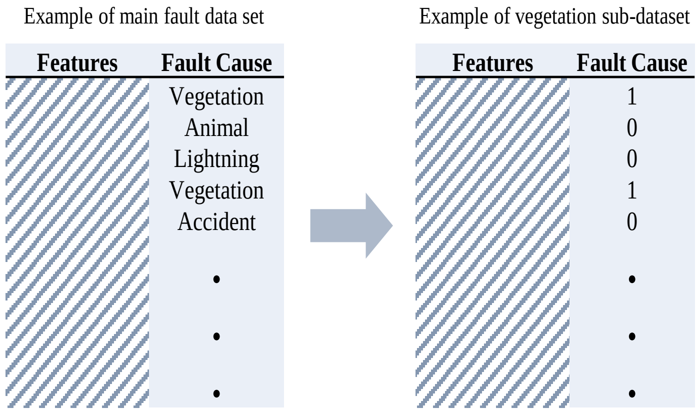
1.3. *Dealing with missing values*: outage datasets typically beset with a substantial number of missing values. The missing data, which can be because of errors in measurements and sensor malfunction, pose serious problems for data analysis. As a result, it is necessary to deal with them. One simplistic approach to address this issue is to remove the rows and columns that contain missing values; nevertheless, this leads to losing valuable information. Therefore, alternative approaches that deal with them properly must be employed.

2. *Converting continuous features to categorical factors*: outage data sets are often described with various features that have different formats. Some of them can be continuous, while the others are categorical. However, in order to develop the

proposed methodology, it is necessary to convert all continues features to categorical ones. To fulfill this task, those features can be categorized into different groups and then represented by group number. For instance, vegetation-related outages that occurred in Charlotte, North Carolina region, between the years of 2009 and 2014 were associated with different ranges of wind speed. This range continuously varies from 1 m/s to 19 m/s. To make the wind speed feature categorical, it can be categorized into 7 groups as [1,3.5), [3.5,6), [6,8.5), [8.5,11), [11,13.5), [13.5,16), and [16,19] and then represented by categorical factors of 1 to 7, respectively. The selection of the number of categories depends on the desired resolution.

3. *Defining new dataset for each outage cause*: The next step is to create a new dataset for each outage cause. In order to perform this, first, the main outage data set is replicated by the number of all outage causes or by the number of outage causes that are the targets for the study. For example, if the purpose is to find patterns for vegetation and animal-related outages, then original data set will be replicated two times. The new data sets are called sub-dataset, which their title is the outage cause. Afterward, the column that shows the outage cause in each sub-dataset is transformed such that it only contains binary values of 1 and 0. The value would be 1 if the outage cause is the same as the title of the sub-dataset; otherwise, it would be 0. An example of this transformation for vegetation-related outages is illustrated in Fig. 5.3. It is worth mentioning that the features for all sub-datasets are the same. Moreover, all sub-datasets have the same number of samples. The only difference is in the column that shows the outage cause.

4. *Generating synthetic data for each sub-dataset*: In order to address the problem of data insufficiency, synthetic data are to be generated for each sub-dataset.



Example of main fault data set

Features	Fault Cause
▨	Vegetation
▨	Animal
▨	Lightning
▨	Vegetation
▨	Accident
▨	•
▨	•
▨	•

Example of vegetation sub-dataset

Features	Fault Cause
▨	1
▨	0
▨	0
▨	1
▨	0
▨	•
▨	•
▨	•

Figure 5.3: Illustration of defining sub-dataset for vegetation-related outages

As illustrated in Fig. 1, artificial data are generated based on the minority group. The minority group in each sub-dataset includes those samples in which their outage cause is represented by 1. Therefore, by applying the SMOTE algorithm, the number of these samples is increased such that finally each sub-dataset contain almost the same number of samples that are represented by 0 and 1.

5. *Converting categorical factors to dummy variables*: Next, for all sub-datasets, all features are to be converted into dummy variables. In order to conduct this, the name of the feature is added to the category number to create a new string, then the former value is replaced with this string. For instance, as mentioned, the wind speed can be converted to categorical factors represented by 1 to 7. If for a specific outage sample the wind speed is represented by 5, it should be converted into WIND5. Likewise, the same procedure ought to be followed for the column that demonstrates the outage cause. For example, for vegetation sub-dataset, if the value for outage cause is 1, it should be converted to VEGETATION1. Taking this step is necessary for interpreting the rules generated by association mining.

6. *Mining association rules for each sub-dataset*: By following the procedure explained thus far, different sub-datasets are generated that resemble basket data type transactions; therefore, they are qualified to be explored by association rule mining. In order to perform association rule analysis, the Apriori algorithm is utilized for each sub-dataset. For this reason, different constraints are to be first determined as follows.

    6.1. *Minimum support*: In order to find frequent itemsets, the minimum support value should be defined. In general, selection of this value depends on the dataset characteristics such as the number of samples, as well as the number of rules that the analyst intends to generate. For example, the authors in [62] consider minimum support values ranging from 2% to 0.5% for a dataset that contains 100,000 samples. In fact, increasing the value for minimum support will boost the statistical significance of the rules; however, it will lead to fewer rules to be generated.

    6.2. *Minimum confidence*: For the purpose of determining the strength of rules, minimum value for confidence ought to be defined. Similar to minimum support value, increasing minimum confidence value results in finding stronger but fewer rules. Again, the selection of this value is at analyst's discretion.

    6.3. *Setting RHS of rules*: By applying the Apriori algorithm, numerous rules might be found; however, since the purpose of the study is to find the patterns for outages' causes, those rules are desired whose RHS is the target outage cause. Therefore, for each sub-dataset, the RHS of the rules is to be set to dummy variable that is created for the outage cause. For instance, the RHS of the rules for analyzing vegetation sub-dataset is to be set to VEGETATION1.

7. *Inspecting rules*: As mentioned, applying the Apriori algorithm results in generating a variety of rules. In order to identify best rules, which show the patterns for outage cause, they are required to be sorted by the lift value. Afterward, those with the highest lift value are selected and interpreted by the analyst.

5.5 Case Study

In order to demonstrate how the proposed approach is to be properly implemented in practice, a case study is investigated in this section. The main goal is to characterize vegetation, animal, equipment failure, lightning and, public accident-related outages that commonly occur in and around Charlotte region according to their underlying causes and to identify significant variables that strongly impact the frequency of these outages. To carry out this task, the proposed procedure is followed step-by-step, and results are achieved and discussed. It is worth mentioning that various factors could affect the frequency of different outages, and based on the results, it is clear that those factors could have high dependency on the geographical location of the distribution system, the environmental characteristics of the region, etc. Therefore, it would not always be possible to generalize the rules that are achieved in this study. However, the proposed approach may be applied on any outage dataset to find the patterns of interest.

5.5.1 Implementation

The proposed algorithm is implemented as follows.

1. *Data collection and preparation*: In this study, outage data collected by Duke Energy is utilized. The data includes information on outages that occurred in Charlotte, North Carolina region, between the years 2009 and 2014. The data is comprised of almost 55,000 samples and almost 20 features. The outages occurred due to various causes. Fig. 5.4 demonstrates the distribution of different outages based on their causes.

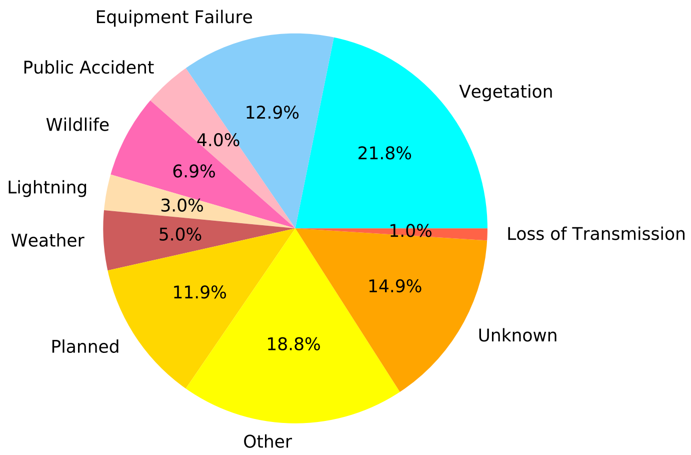


Figure 5.4: Distribution of different outages for Charlotte between the years 2009 and 2014

1.1. *Obtaining data for various features*:

Every time a outage occurs in distribution systems within the domain of Duke Energy, several pieces of information related to that distribution outage are recorded into a dataset as one record entry. Each outage record is described with almost 20 features. Nevertheless, among the features, some are irrelevant to the purpose of this study; hence, they are excluded from this analysis. Those features that are selected to be included are: weather condition, time, voltage level, circuit number, outagey phases, and activated clearing device. The time is split by the authors into three groups of the month, time of day, and weekday. Furthermore, the authors have added four additional useful features of temperature, humidity percentage, dew, and wind speed to the available data from external sources.

1.2. *Data cleansing*: The data set contains a few number of input errors and duplicate data. Therefore, by exploring data set in the R programming language, all errors are identified and removed.

1.3. *Dealing with missing values*: There is a large number of missing values in the dataset. In order to effectively deal with them, the "randomForest-

SRC" algorithm developed by [68] in the *R* programming language is implemented in this study. In this method, the missing data is imputed by randomly drawing values from non-missing values. For this purpose, a random forest is grown and is used to impute missing data.

2. *Converting continuous features to categorical factors*: In what follows, all features are fully described, and it is explained how they are converted to categorical factors. It is worth mentioning that the exact name of the feature, which is utilized in different steps of the algorithm is provided in the parenthesis.

   - *Weather condition (WEATHER)*: Each outage is associated with a weather condition. Weather conditions are categorized into 10 groups: calm, extreme cold, rain, wind, the combination of wind, rain, and lightning, lightning, snow, ice, extreme heat, and the combination of wind and rain. These are represented by factors 0 to 9, respectively.

   - *Month (MONTH)*: Month can affect the frequency of outages with different causes. Months are represented by numbers 1 to 12.

   - *Weekday (WEEKDAY)*: Weekday can make impact on some specific outages such as public accident-related outages as the car traffic differs between days. In addition, the authors demonstrated in [64] that weekday takes high importance with regard to identifying equipment failures. Therefore, this feature is included in the analysis, which the days are represented by numbers 1 to 7, where Monday is represented by 1.

   - *Time of day (TIME)*: Time of day is highly useful for outage root cause analysis since several environmental parameters are associated with it. In order to add this feature to each outage sample, first, different time intervals are created. 6am-12pm, 12pm-5pm, 5pm-8pm, 8pm-12am, 12am-6am are defined as morning, afternoon, evening, night and midnight, respectively. Then, the aforesaid groups are represented by 1 to 5, respectively.

- *Clearing device activated (CD)*: When a outage occurs, the information about the device that was activated to clear it is recorded as a feature. This feature is extremely valuable as it helps in finding the patterns for post-outage situations. There are 12 clearing devices in the dataset: substation device, feeder breaker, line recloser, line fuse, transformer fuse, transformer CSP, HPP/MHO breaker, service/secondary device, disconnect device, jumper, sectionlizer, and transmission device. These devices are represented by 0 to 11, respectively.

- *Voltage level (VOLTAGE)*: Distribution systems are comprised of different voltage levels, which all level are exposed to various outages. There are five different voltage levels for existing distribution circuits in Charlotte: 2kV, 4kV, 7kV, 12kV, and 24kV. In the dataset, these groups are represented by numbers 1 to 5, respectively.

- *outagey phases (PHASE)*: Different line phases can be affected by a outage. Duke Energy captures the outagey phases and records them in the dataset. Each possible combinations of outagey phases are represented by a number. All combinations are included in the analysis; however, two of them, namely three-phase and one-phase are observed to be necessary for interpreting the results of this study. These two combinations are represented by number 5 and 10, respectively.

- *Circuit (CIRCUIT)*: There are more than 160 distribution circuits associated with outage samples available in Duke Energy dataset. These circuits differ in terms of age, length, location and voltage level. Each circuit is represented by a specific number.

Different environmental features including wind speed, humidity, dew point value, and temperature can exert significant effect on outage frequency. Therefore, they are included in the analyses, and are described as follows.

- *Wind speed (WIND)*: outages occur in a broad range of wind speed. In this dataset, the range covers wind speeds of 1m/s to 19m/s. The wind speed is categorized into 7 groups of [1,3.5), [3.5,6), [6,8.5), [8.5,11), [11,13.5), [13.5,16), and [16,19] and then represented by categorical factors of 1 to 7, respectively.

- *Humidity (HUMIDITY)*: The range for humidity varies from 20% to 96%. Similar to wind, humidity values are categorized in 7 groups. The categories are [20,36), [36,46), [46,56), [56,66), [66,76), [76,86), and [86,96]. These groups are represented by categorical factors of 1 to 7, respectively.

- *Dew (DEW)*: Dew values associated with the outages in the dataset range between 0°F and 74°F, which are categorized into 7 groups of [0,13), [13,23), [23,33), [33,43), [43,53), [53,63), and [63,73] and represented by 1 to 7, respectively.

- *Temperature (TEMPERATURE)*: Similar to the other environmental variables discussed, 7 categories are created for temperature. Temperature values range through 16°F and 90°F. The created categories are [16,30), [30,40), [40,50), [50,60), [60,70), [70,80), and [80,90] and represented by 1 to 7, respectively.

3. *Defining new dataset for each outage cause*: The next step is to create a new dataset for each target outage cause. As mentioned earlier, in this case study, vegetation, animal (wildlife), equipment failure, lightning and public accident-related outages will be explored; consequently, five sub-datasets will be created.

In order to carry this out, the procedure illustrated in Fig. 3 is followed five times.

4. *Generating synthetic data for each sub-dataset*: The next step is to generate synthetic data for each sub-dataset. In order to conduct this task, SMOTE algorithm which was explained earlier is utilized. To perform this algorithm, Random Over-Sampling Examples (ROSE) package developed by [69] in the *R* programming language is implemented five times, one time for each sub-dataset. After the method is applied, the percentage of instances in each sub-dataset whose outage cause value is represented by 1 increases. In general, the decision on the amount of synthetic data to be generated depends on the data size, types of data, features, as well as other factors. Therefore, different amounts of synthetic data should be generated and tested to understand how it affects the performance of the data analytics algorithm. For example, the authors in [65] consider different amounts of synthetic data ranging between 100% and 500%. In this study, the number of synthetic data varied between 200% and 700% based on the outage type. By creating such synthetic data, each sub-dataset would contain sufficient amounts of information relating to each outage cause.

5. *Converting categorical factors to dummy variables*: Then, by following the procedure explained in Section 5.4, part 4, all categorical factors are converted to dummy variables for all sub-datasets. For example, as mentioned earlier, one feature is month, which takes values of 1 to 12. These values are converted to MONTH1,..., MONTH12. This should be performed for all available features. Similarly, the same procedure ought to be followed for the column that demonstrates the outage cause.

6. *Mining association rule*: Generated sub-datasets, which look like basket data type transactions, are ready to be investigated by association rule mining. In or-

der to mine these sub-datastes, "arules" package [70], which is based on Apriori algorithm is implemented in R programming language. Necessary constraints are specified as follows.

6.1. *Minimum support*: The minimum support to generate large itemsets in this study is selected to be 0.0018, which is equivalent to 100 occasions. It is assumed that if an itemset appears more than 100 times, it can be assumed to be frequent. In fact, it is decided not to select a high value for the minimum support as it is preferred that the algorithm generates more rules. These rules could be further investigated and interpreted to select the most important ones. It is worth mentioning that selecting a low value is always a better option compared to selecting higher minimum support because the latter option could result in missing possible important rules.

6.2. *Minimum confidence*: The minimum confidence is selected to be 70%. As mentioned earlier, the selection of this value depends on analyst's decision.

6.3. *Setting RHS of rules*: For each sub-dataset, the RHS of the rules is set to dummy variable that is created for that outage cause. For instance, the RHS of the rules for analyzing vegetation sub-dataset is set to VEGETATION1.

7. *Inspecting rules*: In order to find the top-ranked rules, which better show the patterns for outage causes, all resultant rules are inspect by applying the inspect function defined in the Apriori package. The procedure is to sort all rules and identify the high-quality rules whose lift values is significant. In this study, those rules whose lift is greater than 1.4 are selected and presented.

5.5.2 Results and Discussion

In what follows, the results of association rule mining for vegetation, animal, equipment failure, public accident, and lightning-related outages are presented and discussed. It is worth mentioning that the LHS, RHS, support, confidence and lift values of rules are provided. Also, the rules are sorted based on their lift value.

5.5.2.1 Vegetation-related outages

The strongest rules achieved for vegetation-related outages are provided in Table 5.1.

Table 5.1: RESULTS OF ASSOCIATION RULE MINING FOR VEGETATION-RELATED outageS

LHS		**RHS**	**Support**	**Confidence**	**Lift**
{WEATHER8, WIND4}	⇒	{VEGETATION1}	0.003337871	0.9728261	1.944528
{WEATHER8, CD3, DEW4}	⇒	{VEGETATION1}	0.002480094	0.9500000	1.898902
{WEATHER8, HUMIDITY5}	⇒	{VEGETATION1}	0.003636228	0.9466019	1.892110
{MONTH2, CD2, TEMPERATURE2}	⇒	{VEGETATION1}	0.001957969	0.9459459	1.890799
{WEATHER17, TEMPERATURE6, WIND4}	⇒	{VEGETATION1}	0.002237679	0.9302326	1.859390
{WEATHER17, HUMIDITY6, WIND4}	⇒	{VEGETATION1}	0.002424152	0.9154930	1.829928
{TIME5, HUMIDITY7, DEW7}	⇒	{VEGETATION1}	0.002405505	0.9084507	1.815852
{WEATHER7, TEMPERATURE2, HUMIDITY6}	⇒	{VEGETATION1}	0.002890335	0.9011628	1.801284
{WEATHER17, WIND7}	⇒	{VEGETATION1}	0.002013911	0.8503937	1.699805
{MONTH2, CD2}	⇒	{VEGETATION1}	0.002386857	0.8951049	1.789176
{CD2, HUMIDITY7}	⇒	{VEGETATION1}	0.002591978	0.8424242	1.683875
{CD10}	⇒	{VEGETATION1}	0.003804054	0.7669173	1.532948
{WEATHER7}	⇒	{VEGETATION1}	0.009006657	0.7606299	1.520381
{WIND7}	⇒	{VEGETATION1}	0.009808492	0.7377279	1.474603

It is worth mentioning that there are several rules that include location (i.e. circuit number) in their LHS; however, due to the security reasons they are not provided in the table. According to the results, it can be understood that extreme heat, ice, snow, and combinations of wind and rain are the most frequent weather conditions for vegetation-related outages. Moreover, very low temperature, very high humidity, high wind, and location play crucial roles in such outages. Also, line reclosers and sectionlizers are the primary devices that clear these outages.

5.5.2.2 Animal-related outages

Table 5.2 provides the considerable rules extracted for animal-related outages.

Similar to vegetation-related rules, the animal-related rules that contain location in their LHS are not provided in the table. However, based on the analysis, it is noticed that the location make a great impact on these outages. In summary, based on Table II, it can be concluded that animal-related outages are highly dependent on month. In fact, most of such outages in Charlotte, North Carolina region, occur during the

Table 5.2: RESULTS OF ASSOCIATION RULE MINING FOR ANIMAL-RELATED outageS

LHS		RHS	Support	Confidence	Lift
{MONTH6, CD4, DEW5}	⇒	{ANIMAL1}	0.002647920	0.9594595	1.917810
{MONTH9, CD4, DEW5}	⇒	{ANIMAL1}	0.005183956	0.9553265	1.909549
{MONTH12, CD4, HUMIDITY5}	⇒	{ANIMAL1}	0.008316706	0.9550321	1.908961
{MONTH10, TIME3, CD4}	⇒	{ANIMAL1}	0.005370429	0.9442623	1.887434
{MONTH12, CD4, DEW5}	⇒	{ANIMAL1}	0.004568594	0.9423077	1.883527
{MONTH12, CD4, TEMPERATURE5}	⇒	{ANIMAL1}	0.002088500	0.9411765	1.881265
{MONTH11, PHASE10, CD4}	⇒	{ANIMAL1}	0.011244336	0.9363354	1.871589
{MONTH10, CD4, DEW4}	⇒	{ANIMAL1}	0.007272456	0.9285714	1.856070
{MONTH11, CD4, WIND2}	⇒	{ANIMAL1}	0.013239599	0.9281046	1.855137
{MONTH11, CD4}	⇒	{ANIMAL1}	0.034385664	0.8995122	1.797985
{MONTH12, CD4}	⇒	{ANIMAL1}	0.029164413	0.8901537	1.779279
{MONTH10, CD4}	⇒	{ANIMAL1}	0.025062002	0.8604353	1.719876
{MONTH11}	⇒	{ANIMAL1}	0.065452104	0.7164727	1.432118

last three months of the year. Furthermore, it can be understood that dew points that are above average make a significant contribution to such outages. Finally, it can be observed that most of the animal-related outages are cleared by transformer fuse.

5.5.2.3 Equipment failure-related outages

The rules achieved for equipment failures are provided in Table 5.3.

Table 5.3: RESULTS OF ASSOCIATION RULE MINING FOR EQUIPMENT FAILURE-RELATED outageS

LHS		RHS	Support	Confidence	Lift
{WEATHER1, WEEKDAY2, WIND2}	⇒	{EQUIPMENT1}	0.004307532	0.9130435	1.825032
{WEATHER1, MONTH1, WEEKDAY2}	⇒	{EQUIPMENT1}	0.004177000	0.9105691	1.820086
{VOLTAGE24, WEEKDAY7, CD7}	⇒	{EQUIPMENT1}	0.005780670	0.9037901	1.806536
{TEMPERATURE 7, HUMIDITY3, WIND2}	⇒	{EQUIPMENT1}	0.002797098	0.8928571	1.784683
{WEATHER1, PHASE10, HUMIDITY2}	⇒	{EQUIPMENT1}	0.001920674	0.8879310	1.774836
{WEATHER1, PHASE10, DEW1}	⇒	{EQUIPMENT1}	0.002517389	0.8766234	1.752234
{WEATHER1, TEMPERATURE1, DEW1}	⇒	{EQUIPMENT1}	0.007011394	0.8764569	1.751901
{WEATHER1, TEMPERATURE1}	⇒	{EQUIPMENT1}	0.007011394	0.8744186	1.747827
{WEATHER1}	⇒	{EQUIPMENT1}	0.007570813	0.8565401	1.712090
{TEMPERATURE1}	⇒	{EQUIPMENT1}	0.016018051	0.7182274	1.435625
{DEW1}	⇒	{EQUIPMENT1}	0.023141328	0.7095483	1.418277

Based on this table, it is noticed that extreme cold and extreme hot weather conditions have a huge impact on equipment failures. In fact, a majority of these outages occur when the temperature, humidity, and dew point values are either very low or very high. Furthermore, it is found that circuits with the highest voltage (i.e. 24kV) are main locations for equipment failures. Also, weekday makes a contribution to these outages.

5.5.2.4 Public accident-related outages

Table 5.4 summarizes the rules with the highest strength for outages with regards to public accidents.

Table 5.4: RESULTS OF ASSOCIATION RULE MINING FOR PUBLIC ACCIDENT-RELATED outageS

LHS		**RHS**	**Support**	**Confidence**	**Lift**
{VOLTAGE24, CD9}	⇒	{ACCIDENT1}	0.002219031	0.9370079	1.872933
{MONTH10, CD1}	⇒	{ACCIDENT1}	0.003132750	0.9230769	1.845087
{MONTH11, CD1}	⇒	{ACCIDENT1}	0.002983572	0.9142857	1.827515
{MONTH10, CD2}	⇒	{ACCIDENT1}	0.003170045	0.8947368	1.788440
{MONTH2, WIND7}	⇒	{ACCIDENT1}	0.003244634	0.8923077	1.783584
{HUMIDITY3, WIND7}	⇒	{ACCIDENT1}	0.003244634	0.8923077	1.783584
{WEEKDAY5, WIND7}	⇒	{ACCIDENT1}	0.003244634	0.8923077	1.783584
{WEATHER2, DEW2}	⇒	{ACCIDENT1}	0.002107148	0.8897638	1.778499
{PHASE5, CD9}	⇒	{ACCIDENT1}	0.009025304	0.8897059	1.778384
{CD2, TEMPERATURE3}	⇒	{ACCIDENT1}	0.005911201	0.8781163	1.755218
{CD1, DEW5}	⇒	{ACCIDENT1}	0.006489268	0.8721805	1.743353
{MONTH1, CD1}	⇒	{ACCIDENT1}	0.003710817	0.8689956	1.736987
{TIME5, CD9}	⇒	{ACCIDENT1}	0.003020866	0.8617021	1.722409
{CD1}	⇒	{ACCIDENT1}	0.032502284	0.7678414	1.534796
{CD9}	⇒	{ACCIDENT1}	0.012363175	0.7542662	1.507661

According to the analysis, these outages are highly affected by location (circuit number); however, as mentioned, due to security reasons, the rules that have circuit number are not presented in the table. Based on the rest of results, it is understood that rain weather condition, high winds, low visibility (i.e. TIME5), and month have great impacts on these type of outages. Also, feeder breaker and jumpers can be identified as main devices that clear accident-related outages.

5.5.2.5 Lightning-related outages

Table 5.5 demonstrates the important rules for lightning-related outages.

Table 5.5: RESULTS OF ASSOCIATION RULE MINING FOR LIGHTNING-RELATED outageS

LHS		**RHS**	**Support**	**Confidence**	**Lift**
{WEATHER5, MONTH3}	⇒	{LIGHTNING1}	0.002871688	0.9685535	1.935988
{WEATHER5, CD4}	⇒	{LIGHTNING1}	0.018442203	0.9509615	1.900824
{WEATHER5, HUMIDITY4}	⇒	{LIGHTNING1}	0.010852742	0.9494290	1.897761
{WEATHER5, TIME2}	⇒	{LIGHTNING1}	0.012288586	0.9374111	1.873739
{WEATHER5, DEW6}	⇒	{LIGHTNING1}	0.005445018	0.9240506	1.847034
{WEATHER5, DEW7}	⇒	{LIGHTNING1}	0.010144144	0.9189189	1.836776
{WEATHER5, HUMIDITY6}	⇒	{LIGHTNING1}	0.018927033	0.9168925	1.832726
{WEATHER5, CD3}	⇒	{LIGHTNING1}	0.003263282	0.9114583	1.821864
{WEATHER5, VOLTAGE24}	⇒	{LIGHTNING1}	0.029444123	0.9090386	1.817027
{WEATHER5, MONTH7}	⇒	{LIGHTNING1}	0.158539542	0.7100977	1.419375

It is obvious that lightning situation is a necessity for these type of outages. However, based on the table, it is understood that month, high voltage, high humidity and high dew points also can affect lightning outages. Moreover, location plays a key role for these outages; nevertheless, due to security reasons, circuit numbers are not presented in the table. With respect to clearing device, it can be mentioned that the majority of outages due to the lightning are cleared by line fuse and line recloser.

5.6 Conclusions

In this study, a novel approach for outage root cause analysis by utilizing association rule mining was proposed. For this purpose, a detailed procedure, which includes several steps was provided to effectively deal with required data preparation, practical problems associated with real-world outage data set, and implementation of association rule mining. In order to show the application of the proposed approach in practice, a case study was selected and investigated. By employing the proposed approach various outages including vegetation, animal, equipment failure, lightning, and public accident-related outages were characterized according to their underlying causes, and variables that strongly impact their frequency were identified. Moreover, the co-occurrence of these variables was investigated. The results demonstrate that applying this approach successfully yields revealing insights into the causal structures and frequent patterns of different outages that commonly occur in power distribution systems.

CHAPTER 6: Power Distribution System Equipment Failure Identification Using Machine Learning Algorithms

6.1 Introduction

In power distribution systems, providing customers with the most reliable supply of electricity in the form of an uninterrupted service is an outright necessity. Reliability plays a crucial role in reducing the cost of electricity and bringing customers' satisfaction. While distribution utilities essentially aim at maintaining the reliability at its highest level, outages caused by various outages pose serious challenges to attaining this goal [71]. Therefore, the problem of producing an appropriate response to the outage comes to the fore [72].

In order to adequately address the issues brought about by outages, identification of the cause is necessary. As a matter of fact, acquiring knowledge of the outage's cause immediately after the outage is extremely beneficial in terms of reducing the outage's duration as 1) it gives the operation and maintenance crew the chance to find the evidence of the outage more quickly, and 2) it enables utilities to dispatch the right crew with the specific equipment which is necessary to repair or replace the damaged components [72].

Several studies have been carried out to identify major sources of outages including animals and vegetation in distribution systems. The authors in [72] have developed a power distribution outage cause classifier in order to address the problem of cause identification. In their work, two classification methods, logistic regression, and artificial neural network are investigated by using six features of weather condition, season, time of day, the number of phases affected, protective device activated and the circuit where the outage happened as inputs. Moreover, the same authors in [73], employ an extended version of the fuzzy classification algorithm to identify three outage causes: vegetation, wildlife, and lightning. In their work, it is demonstrated that the proposed algorithm could deal with imbalanced data problem more effectively and delivers better performance compared to the neural network.

The authors in [74] have investigated tree outages in distribution systems. Four measures of actual values, normalized values, relative, and likelihood values are employed to examine the characters of tree outages. Furthermore, the significance of several factors that make contributions to tree outages is evaluated by using logistic regression analysis.

One of the main causes for the outage in distribution systems is equipment failure, which includes transformer failure, and broken insulator, to mention a few [75], [76]. Unlike the animal and vegetation-related outages, equipment failure identification has drawn less attention. However, over recent years, with the deployment of an enormous number of intelligent electronic devices and various sensors, necessary data for studying equipment failure have become available. Analyzing such data by employing analytical methods could shed light on understanding the characteristics of equipment failures and their causes. For instance, the authors in [76], by using chronological failure data, propose an approach based on Bayes methodology and applications of population Monte Carlo to predict the number of equipment failures and to formulate replacement strategies.

This chapter focuses on developing an approach for identifying equipment failure outages. This task is considered as a binary classification problem in which outages are categorized into two classes of equipment failure and non-equipment failure ones. The target class is identified by using machine learning classification methods. To carry out this study, actual outage data collected by Duke Energy are utilized.

In this chapter, several features that have impacts on equipment failures are described and their relationships are statistically examined. Moreover, this chapter discusses two main practical issues which are usually faced in outage cause identification problems, namely presence of imbalanced classes, and feature selection and provides two solutions to address these problems. Fig. 6.12 provides the technical flowchart of the approach. Each block of the flowchart would be fully explained in following sections.

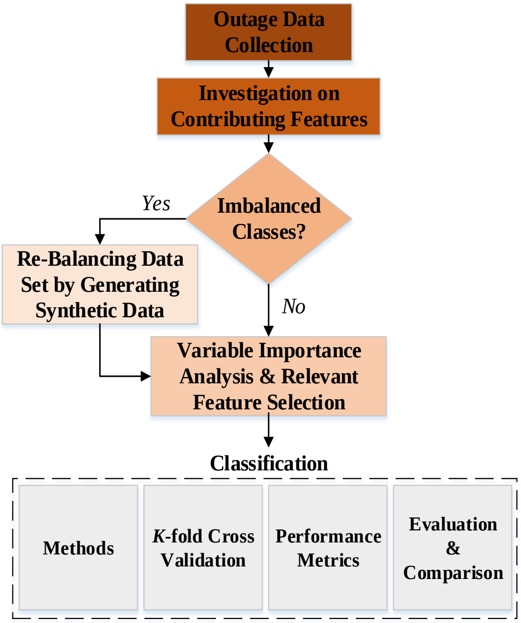

Outage Data Collection
Investigation on Contributing Features
Imbalanced Classes?
Yes
Re-Balancing Data Set by Generating Synthetic Data
No
Variable Importance Analysis & Relevant Feature Selection
Classification
Methods
K-fold Cross Validation
Performance Metrics
Evaluation & Comparison

Figure 6.1: Technical flow chart of the approach

The organization of this chapter is as follows. First, the data set is described in Section 6.2. In Section 6.3, several features that have impacts on equipment failure are discussed and their relations are graphically demonstrated. In Section 6.4, the problem of imbalanced classes is discussed and addressed. Determining the significance of all features and selecting the relevant features are discussed in Section 6.5. The features are fed into the classification methods and classification's performances are compared and evaluated in Section 6.6. Finally, Section 6.7 closes the study by drawing conclusions.

6.2 Data Description

In this chapter, outage data collected by Duke Energy is utilized. The data includes information on outages that happened in the Charlotte, North Carolina region, between the years 2009 and 2014. The data is comprised of almost 55,000 samples and 20 features.

Among the features, some are relevant to the purpose of this study including time, weather condition, voltage level, circuit number, outagey phases, activated clearing device and outage cause. The time is split by the authors into three groups of the month, time of day, and weekday. Furthermore, the authors added four additional useful features of temperature, humidity percentage, dew, and wind speed to the available data from external sources. It is worth mentioning that continuous variables, such as temperature, are categorized into different groups, and represented by group numbers.

As was mentioned earlier, this study deals with a binary classification problem for cause identification. Therefore, the feature that demonstrates the outage cause is composed of only 0 and 1. If the outage outage is equipment failure, the outage cause value would be 1; otherwise, it would be 0. It should be mentioned that almost 11% of outages are due to equipment failures.

6.3 Investigation on Contributing Features

In this study, 12 features that make contributions to equipment failure are statistically examined: weather conditions, month, weekday, time of day, clearing device activated, voltage level, outage phases, circuit number, wind speed, humidity level, dew, and temperature. Figs. 6.2 to 6.13 illustrate the results of the analysis for above-mentioned features, respectively. For example, according to Fig. 2, it can be observed that weather conditions are categorized into 10 groups: calm, extreme cold, rain, wind, the combination of wind, rain and lightning, lightning, snow, ice, extreme heat, and the combination of wind and rain. Also, it is understood that most of the equipment failures (73%) occur during calm weather conditions. Rain condition and the combination of wind, rain and lightning are the second and third most frequent weather conditions for the equipment failures. To perform this analysis, a few assumptions are made as follows.

- In order to carry out the "Time of day" analysis, different time intervals are created. 6AM-12AM, 12AM-5PM, 5PM-8PM, 8PM-12AM, 12AM-6AM are defined as Morning, Afternoon, Evening, Night and Midnight, respectively.

- There are more than 160 distribution circuits recorded in Duke Energy data set. These circuits differ in terms of length, location and voltage level. Each circuit has its own number of failure outages.

- Continuous variables including wind speed, humidity level, dew, and temperature are categorized into different groups and are represented by group number. For instance, equipment failures occur for different ranges of wind speed. In this study, the range varies from 1 m/s to 19 m/s. The wind speed is categorized into 7 groups.

Figure 6.2: Weather analysis

6.4 Imbalanced Classification Problem

One main challenge with real-world outage data is the presence of imbalanced classes. As equipment failure is only one of the various causes for the outage, non-equipment failure outages considerably outnumber the equipment failure class in the data set. Imbalanced class distribution of a data set is problematic as it can result in biased predictions and misleading accuracy for most classification learners [77].

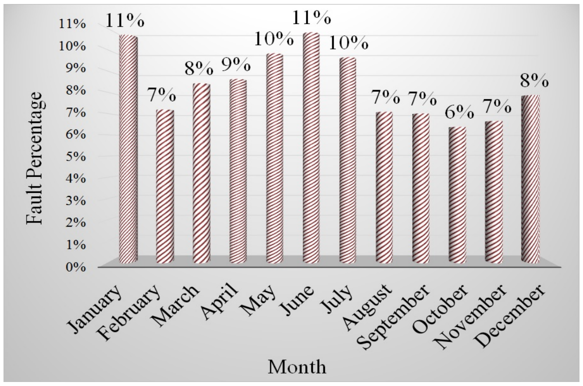


Figure 6.3: Month analysis

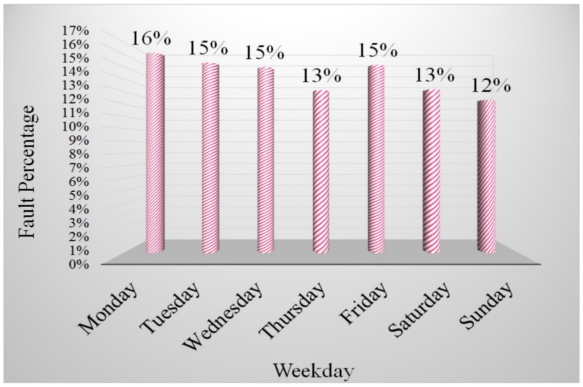


Figure 6.4: Weekday analysis

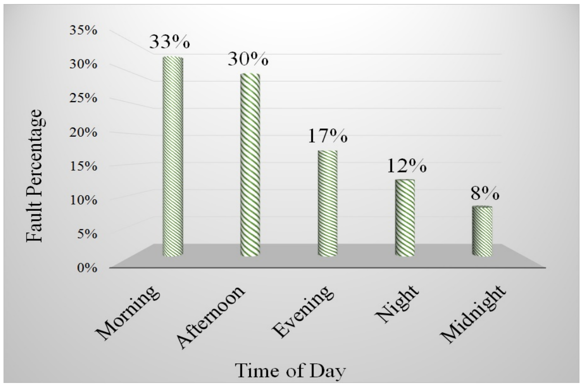


Figure 6.5: Time of day analysis

In order to address this problem, a variety of methods has been proposed. These methods could be categorized under two well-established approaches of cost-sensitive and re-balancing the data set [77]. The rationale behind cost-sensitive based methods is to evaluate the cost associated with misclassifying the observations. The objective

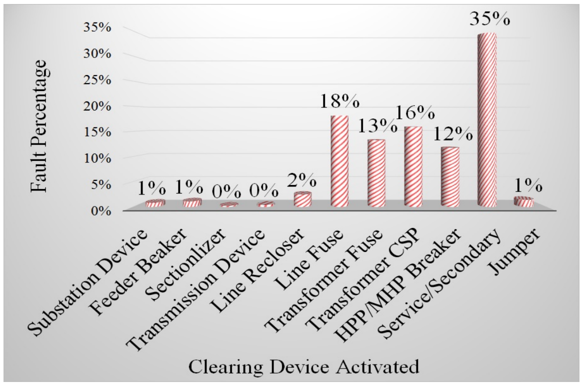


Figure 6.6: Clearing device analysis

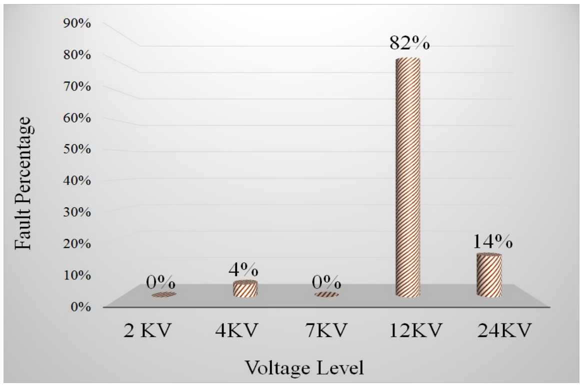


Figure 6.7: Voltage level analysis

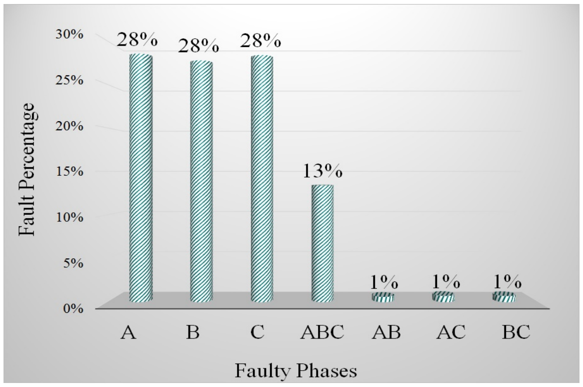


Figure 6.8: outagey phases analysis

is to take a decision to minimize the expected cost. In the re-balancing approach, different mechanisms such as over-sampling the instances in the minority class (i.e. duplicating some of the minority class), under-sampling the observations in the majority class (i.e removing some of the majority class), and generating synthetic data

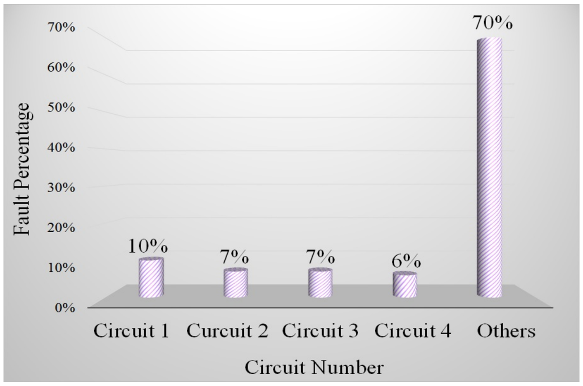


Figure 6.9: Circuit number analysis

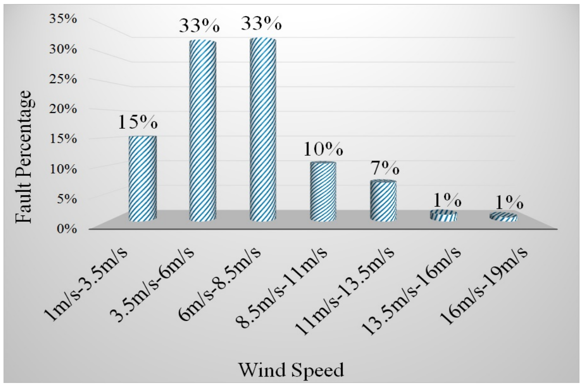


Figure 6.10: Wind speed analysis

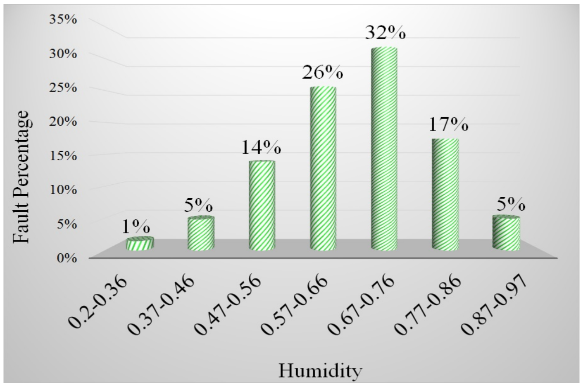


Figure 6.11: Humidity level analysis

are employed. Over and under-sampling mechanisms have captured significant attention in the past; however, they are identified with various shortcomings. As a matter of fact, over-sampling could lead to the over-fitting problem and under-sampling may cause losing important information [78]. The over-fitting problem is elaborated in the

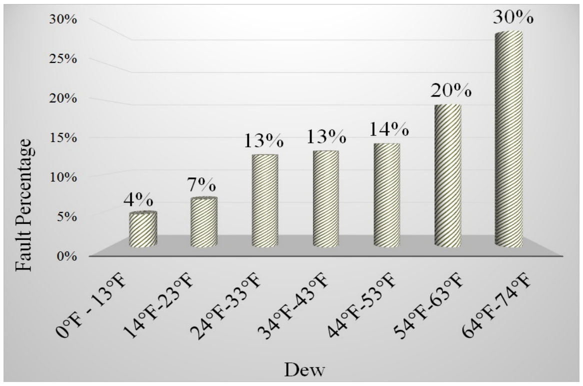


Figure 6.12: Dew analysis

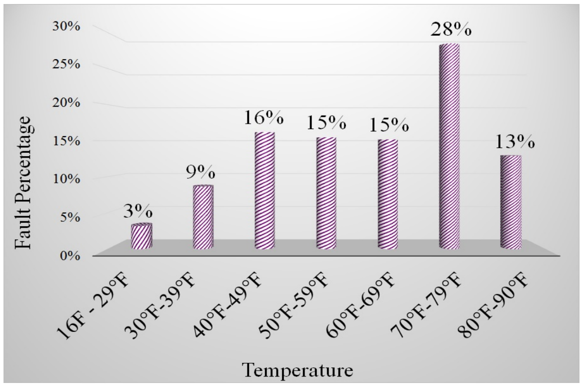


Figure 6.13: Temperature level analysis

next section.

In order to overcome the aforementioned issues, generating synthetic data is introduced, which is of particular interest to analysts due to its simplicity and effectiveness. This method, which is known as Synthetic Minority Over-sampling Technique (SMOTE) [79], is a combination of oversampling and undersampling; however, the oversampling is performed by creating artificial data according to feature space similarities from minority samples rather than merely replicating the minority class. In order to create the random samples, k-nearest neighbor algorithm and bootstrapping are used in this method. The procedure is as follows [79].

- Step 1: The difference between the feature vector (sample) under consideration and its nearest neighbors is taken;

- Step 2: The difference is multiplied by a random number between 0 and 1 and added to the feature vector under consideration.

Since the samples created by this method are not exact copies of the instances in the minority class, the classifier does not over-fit.

As mentioned earlier, in outage data set used in this study, only 11% of outages are due to equipment failure; as a result, non-equipment outages represented by 0 considerably outnumbers the equipment failures represented by 1. To tackle this issue, by using Random Over-Sampling Examples (ROSE) package developed by [80] which is based on SMOTE, synthetic data are generated according to equipment failures and are added to the data set. The algorithm is implemented in the *R* programming language. After applying the method, the percentage of equipment failure instances are increased to 50%, which makes the data set balanced.

6.5 Variable Importance Analysis and Feature Selection

Outage data sets are usually described with various features, while some of them might be irrelevant to the classification. In fact, learning classifiers tend to over-fit with the presence of features that make no contribution to the response. Over-fitting occurs when the model is too complicated for the data and is caused by adding too many independent variables into the model. When over-fitting happens, the model becomes tailored to the sample data, while it shows a poor performance for the new data. In addition to the over-fitting problem, the inclusion of unnecessary features into the model substantially increases the running time, computation cost, and decreases the accuracy; hence, it is necessary to study the importance of features in order to identify and remove the redundant features. This procedure, which is known as feature selection, leads to a better learning accuracy, lower running time and could prevent unwanted problems [81].

In order to perform the feature selection analysis, there are two general approaches: 1) the minimal-optimal feature selection which identifies a small subset of variables that deliver the best possible classification performance, and 2) the all-relevant feature selection which identifies all features that are in some circumstances relevant for the classification [82].

All-relevant feature selection, which is fairly new in the field of feature selection, is found to be highly useful. As a matter of fact, discovering all relevant features instead of only the non-redundant ones gives a profound understanding about the problem and the mechanism behind the subject of the interest. Moreover, it provides a quantitative measure of the importance of the features [83].

In this study, Boruta algorithm developed by [82] is used to perform the all-relevant feature selection. To find the relevant features and their importance, this algorithm carries out a top-down search which compares original features' importance with importance achievable at random, estimated using their permuted copies, and progressively eliminates irrelevant features to stabilize that test. One advantage of this algorithm is that it provides a quantitative measure for the importance of the features [82].

By implementing the Boruta algorithm in the *R* programming language, all 12 features selected in this study are confirmed to be important; consequently, they will all be used as inputs to classifiers discussed in the next section. Moreover, the importance of the features is achieved and is demonstrated in Fig. 6.14.

According to this figure, it can be understood that the clearing device activated, circuit number, and weather conditions take on the highest importance with regard to identifying equipment failures, while dew value is the least important feature.

6.6 Classification

6.6.1 Classification Methods

In this study, three well-established classification methods of decision tree, logistic regression, and naive Bayesian classifier are employed. These methods are briefly described as follows [84], [85].

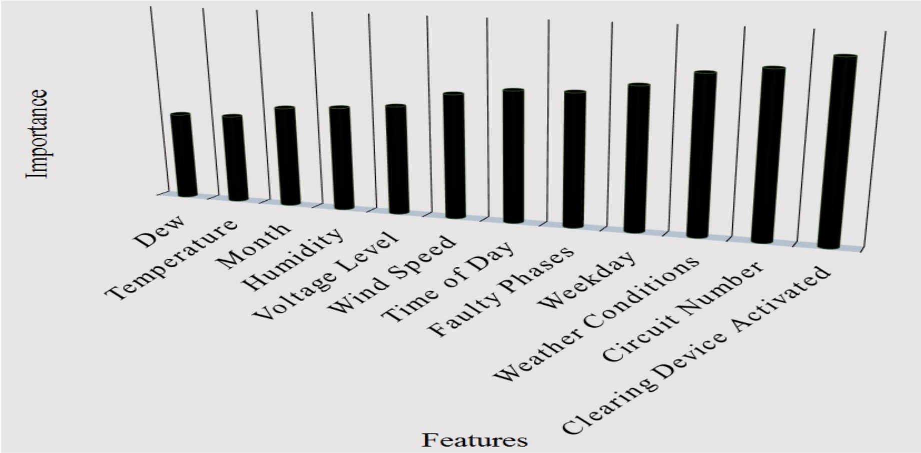


Figure 6.14: Feature importance

6.6.1.1 Decision Tree

A decision tree is a tree-structured plan that classifies the target value by recursively partitioning the data. The internal nodes of a decision tree indicate different features, the branches between the nodes demonstrate the possible values that features can take, and terminal nodes represent the class of the outcome.

6.6.1.2 Logistic Regression

Logistic regression, which is especially popular for binary classification problems, is a probabilistic method that predicts the probability of an outcome by fitting data to a logistic function. The resultant probability values can be converted into classes by using different functions.

6.6.1.3 Naive Bayesian Classifier

The naive Bayesian classifier is a probabilistic method which is based on the Bayes' theorem of conditional probability. The naive Bayesian classifier uses all the attributes contained in the data individually and assumes that they are equally significant and independent of each other.

6.6.2 K-fold Cross Validation

In order to evaluate different settings for the aforementioned methods including pruning the decision tree and regularization for logistic regression, k-fold cross-validation technique is employed where k is selected to be 10 in this study. In k-fold cross-validation, the data set is randomly split into k smaller sets, where k-1 folds are utilized for training the model, while the remaining subset is used as validation data to compute the performance of the model. This procedure then is repeated k times, with each of the k folds being used exactly once as the validation data. The results are then averaged to produce the final value. In fact, this procedure is necessary to avoid problems such as over-fitting. Moreover, without having validation data, the evaluation metrics no longer report on generalization performance.

By using k-fold cross validation technique, three discussed classification methods are trained and tested.

6.6.3 Performance Metrics

Machine learning classifiers are typically evaluated by using confusion matrix. Table 6.1 demonstrates a general confusion matrix. This matrix compares the number of instances that are classified as 0 or 1 (negative, and positive) versus the real observation. In the confusion matrix, TP, FP, FN and TN stand for True Positive, False Positive, False Negative and True Negative, respectively. For example, FP is the number of negative examples incorrectly classified as positive. By utilizing confusion matrix, three metrics of accuracy, True Positive Rate (TPR), and False Positive Rate (FPR) are defined as in (1) to (3), respectively [78].

Table 6.1: GENERAL CONFUSION MATRIX

	Predicted Value		
Observed Value		Negative	Positive
	Negative	TN	FP
	Positive	FN	TP

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \tag{6.1}$$

$$TPR = (TP)/(TP + FN) \tag{6.2}$$

$$FPR = (FP)/(FP + TN) \tag{6.3}$$

In addition, another metric, Receiver Operating Characteristic (ROC) curve is used in this study. TPR (Y-axis) could be plotted against FPR (X-axis). The resulting graph is called ROC curve. The ROC curve provides the ability to measure the performance of the classifier over the entire operating range. The ideal point on the ROC curve would be (0,100), that is all positive examples are classified correctly and no negative examples are misclassified as positive. The closer the curve follows the left-hand border and then the top border of the ROC curve, the more accurate the model.

6.6.4 Evaluation and Comparison

The aforementioned metrics are calculated for the three methods and the results are provided in Table 6.2. According to this table, it can be understood that the decision tree delivers better performance compared to the logistic regression and the naive Bayesian classifier as it provides higher accuracy, predicts more positive instances correctly (i.e. higher TPR value), and classifies fewer negative examples incorrectly (i.e. lower FPR value).

Table 6.2: METRIC VALUES FOR CLASSIFIERS

Method	**Accuracy**	**TPR**	**FPR**
Decision Tree	0.87	0.89	0.14
Logistic Regression	0.66	0.68	0.35
Naive Bayesian Classifier	0.62	0.63	0.38

Moreover, the ROC curves are demonstrated in Fig. 6.15. As can be observed from this figure, the decision tree is closer to an ideal classifier since its curve goes

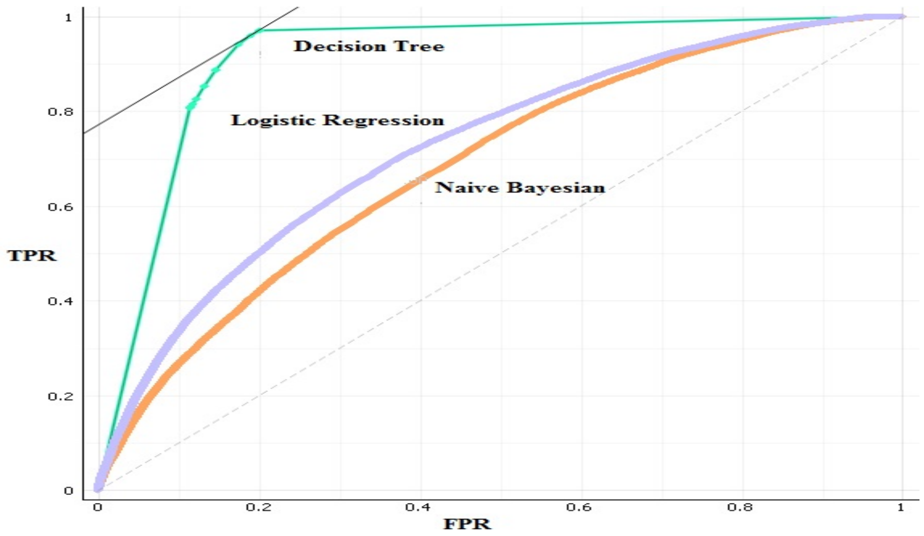


Figure 6.15: ROC curve for classification methods

straight up the Y axis, and then along the X axis. However, the logistic regression and the naive Bayesian classifier both have fewer characteristics in common with an ideal classifier.

Although the logistic regression and the naive Bayesian classifier have their own advantages such as being less prone to over-fitting and could be computationally faster, the decision tree has distinguishing characteristics, which makes it a more effective method to deal with the problem of identifying equipment failure outages. As a matter of fact, decision trees have been proven to be particularly suitable for classification problems in which decision boundaries (i.e. the lines that are drawn to separate different classes) are highly non-linear. Since the number of features used in this study is relatively small and they are not sparse, the decision boundary is very likely to be non-linear. Consequently, decision tree suits better to this problem compared to logistic regression, which is a better fit for problems with linear boundary decisions. Moreover, decision trees are able to easily handle feature interactions, while naive Bayesian classifiers make the assumption of independent features. In real-world outage data sets, due to the nature of features, it is almost impossible that this assumption will completely hold, which places serious limitations on the performance of the naive Bayesian classifier. In addition, as mentioned earlier, synthetic data is generated in this study for equipment failures by using the SMOTE algorithm. The authors in [79] explained that in circumstances in which synthetic data is generated by this algorithm, decision trees tend to generalize better.

As a result, based on these comparisons, it can be argued that applying the proposed approach by using a decision tree delivers an excellent performance by identifying 89% (TPR) of the equipment failure outages correctly.

6.7 Conclusion

The main objective of this chapter was to develop an approach for identifying equipment failures in distribution systems. For this purpose, the problem was defined as a binary classification in which outages were categorized into two classes: equipment failure and non-equipment failures. The classification task was performed by employing three well-established classifiers: decision tree, logistic regression, and naive Bayesian classifier. The chapter also addressed two practical issues with outage data, namely the presence of imbalanced classes and relevant feature selection. The SMOTE method and Boruta algorithm were discussed and employed as effective solutions to address the aforementioned issues, respectively. The results demonstrated that by applying the approach, 89% of the equipment failures could be identified correctly.

CHAPTER 7: CONCLUSIONS AND RECOMMENDATIONS

7.1 Conclusions

With the explosion in data gathering within the smart grid framework, data analytics has emerged as a desirable feature in maintaining power system operating security, as it enables effective and timely decision-making actions by operators as well as planners. Advanced statistical and machine learning techniques, combined with the traditional rigorous mathematical model methodologies, can provide the necessary actionable information support for improving the outage management systems. This work envisioned an integrated framework that will demonstrate the use of innovative data analytics technologies for early warning of degrading operating conditions that may imperil system security. With the high cost of improving reliability and increased calls for faster restoration of power in the face of severe storms, the need for advanced outage management has grown. The integration of contextual information into comprehensive models for predicting, identifying and analyzing outages for the purpose of making decisions or taking actions is increasingly becoming urgent.

In this dissertation, various approaches for predicting, identifying, and analyzing outages in power systems were successfully developed. Various practical challenges that are faced in outage-related problems were also discussed and workable solutions to address those problems were provided. The proposed approaches were categorized into three main groups of 1) predicting unavoidable outages 2) preventing avoidable outages by obtaining insights from the data, and 3) managing the situation during the outages. In what follows, the main conclusions drawn for each category are provided.

7.1.1 Predicting unavoidable outages

Three different predictive approaches for analyzing and forecasting the number of vegetation, lightning, and animal-related outages on different horizons were proposed. It was demonstrated that in order to develop practical approaches, various types of information including historical records of outages, climatological, and geographical variables should be obtained and processed. Moreover, successful approaches require a great deal of attention to the data pre-processing task. In particular, conducting a comprehensive missing value imputation and outlier detection and handling process is essential.

On majority occasions, a key step in building a realistic approach is to adequately define the extent of the predictionsâ target area. Aggregating substations and creating broader geographical areas by using clustering algorithms seem like a workable solution for this purpose. This helps to harness the randomness of the number of outages and to obtain more accurate weather information. Throughout this dissertation it was shown that due to the complex nature of outages and several factors that influence their occurrence, utilizing simplistic approaches such as calculating the average number of outages based on historical data does not lead to obtaining an accurate predictive model; therefore, more sophisticated models are required. Moreover, the occurrence of these outage depends on various factors which are subject to change during the time; hence, models that take these factors into considerations are required.

With regards to vegetation-related outages, it was demonstrated that they occur due to various reasons; however, they could be categorized into two main groups of growth-related and weather-related outages. Each category of the vegetation-related outage reveals a different pattern, and therefore requires a different treatment. Such categorization is necessary to build an accurate model. Time series models can successfully explain the patterns existing within growth-related vegetation outages and are able to produce convincing predictions. Machine learning regression models that can handle the non-linear interaction in the data are effective tools for predicting weather-related vegetation outages.

With regards to lightning-induced outages, it was demonstrated that to obtain the best possible predictive performance, those outages should be categorized into a few manageable groups. These groups exhibit distinguishable characteristics with regards to the number of thunderstorm events. The binomial probability model is an adequate model to find the likelihood of groups of outages given a certain number of thunderstorm event sand a specific area in the system. An important issue that should be addressed to build a successful predictive model is the imbalanced problem. The weighted logistic regression model can handle this problem and can deliver an appropriate classification of different groups of outages.

For animal-related outages, it was shown that the occurrence of these outages changes over time due to the influence of various factors. These outages exhibit timely and seasonal behavior. The population size of animals that cause outage has a significant relationship with the number of outages. The average number of outages varies significantly during different seasons. Extreme weather conditions can influence the occurrence of these outages. The dynamic regression model is an effective predictive model that can take the aforementioned factors into account and provide a reasonable prediction for those outages on a weekly basis.

7.1.2 Preventing avoidable outages by obtaining insights from the data

With the increasing requirements on power distribution utilities to ensure system reliability, utilities seek to find practical solutions that enable them to restrict specific outages. For achieving this, it is crucial to acquire a profound understanding of different outages by exploring their underlying causes and identifying key variables related to those causes. For this purpose, a novel approach for outage root cause analysis by using association rule mining was developed.

It was demonstrated that the association rule mining is capable of extracting comprehensive patterns from outage data sets within a short amount of time. To implement it, however, in-depth data preparation is required. Moreover, there are several practical issues associated with outage data sets that ought to be addressed. Therefore, a step-by-step procedure that fully deals with necessary data preparation, practical issues associated with outage data sets, and implementation of association rule mining was provided. Furthermore, this procedure was applied to investigate a real-world outage data set. As a result of the case study, causal structures and frequent patterns for vegetation, animal, equipment failure, public accident, and lightning-related outages were explored.

7.1.3 Managing the situation during the outages

As was mentioned, after an outage occurs, utility companies are to produce an appropriate response to make sure that the restoration process is fast and smooth and also the customers are well-informed about the situation. Identifying the cause of the outage is an important task for reaching the aforementioned objectives. As a result, a model to identify equipment failure outages was developed. For this purpose, the problem was defined as a binary classification in which outages were categorized into two classes: equipment failure and non-equipment failures. The classification task was performed by employing three well-established classifiers. The research also addressed two practical issues with outage data, namely the presence of imbalanced classes and relevant feature selection. The SMOTE method and Boruta algorithm were discussed and employed as effective solutions to address the aforementioned issues, respectively. The results demonstrated that by applying the approach, 89% of the equipment failures could be identified correctly.

All the advantages of the proposed approaches were built upon generic outage data collected by utilities, and typical daily weather forecast data, which is publicly available. This fact makes the implementation of the approach easily attainable within a reasonable level of accuracy. However, the approach provides the flexibility to be improved by utilizing various other sources of data. Results of this dissertation could be very informative to utility companies in gaining insight about their outage problems, and to build better models for predicting them.

7.2 Recommendations

7.2.1 Obtaining additional data

Although many different pieces of information pertaining to different outages were considered in this dissertation, all possible factors were not accounted for due to lack of access to related data. As a result, the performance of the proposed approach may be improved by the inclusion of additional climatological and geographical information (e.g., satellite images, LiDar data, etc.). Moreover, obtaining a considerably larger scale of outage data may open up unique opportunities for utilizing the most advanced predictive models including deep learning algorithms for predicting different outages.

7.2.2 Predicting outage duration

In recent years, with the increasing requirements imposed on power distribution companies to ensure system reliability and customer and regulator satisfaction, analytical tools that enable utilities to more effectively manage their responses to the power outages have become a vital necessity [86]. These tools, which are commonly known as Outage Management System (OMS), are computer systems that assist utilities in detecting, tracking, and analyzing customer service interruptions, ranging from individual outages to system-wide disturbances.

The OMS can support various functions including collecting data, producing reports, analyzing customer calls, managing crews, just to mention a few. More recently, as the grid is becoming smarter, the OMS may be given the potential capabilities to offer more sophisticated services, one of the most significant of which, would be providing an estimate for outage duration (a.k.a. restoration time).

Having the ability to estimate the outage duration is of great importance for utilities as it 1) increases the key stakeholders' (i.e. customers, regulators, media, etc.) awareness of the situation, which eventually improves their satisfaction, 2) facilitates the communication and planning between the field crew and operators, which ensures that crews are deployed most effectively, and 3) supplies ancillary services with necessary information, which helps to run the restoration effort more smoothly [87].

Historically, utilities were relying upon educated guesses made based on managerial judgment and previous experience to give an estimate for outage duration [87], [88]. Nevertheless, over recent years, with the explosion in data gathering within the smart grid framework and major advances in big data analytics, the OMS has become able to support utility companies in enhancing the estimation of outage duration [88], [89]. In order to develop systems that are capable of conducting this task, however, it is crucial to acquire a deeper understanding of the outage duration by exploring underlying causes and identifying key variables. Furthermore, analyzing the outage duration and examining the variables that make major contributions provide the utilities with a wealth of information that can be used to improve the design of existing distribution systems and to carry out essential preventive maintenance for the purpose of reducing the duration of future outages [89].

In order to develop sophisticated models for estimating the outage duration, it is essential to investigate the factors that make contributions to the duration. However, having limited access to sufficient amounts of real-world data and lack of adequate information relating to outage duration have resulted in serious challenges for data analysts to build critical features that could be used as inputs for duration estimation algorithms. Therefore, in order to ease this burden, a comprehensive statistical analysis of the outage duration in distribution systems was carried out [90] with two primary goals:

- Analyzing and interpreting the characteristics of the outage duration with respect to various contributing features.

- Identifying significant variables that strongly impact the duration.

By analyzing different variables and implementing the random forest algorithm in *R* programming language, it can be demonstrated that various factors play an important role in predicting the duration of outages. The importance of different variables is shown in Fig. 7.1.

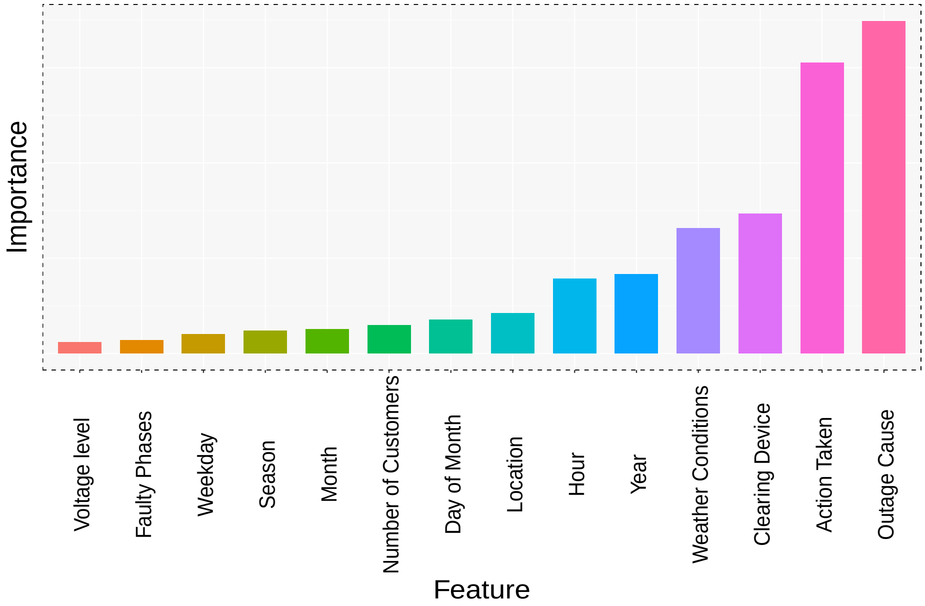


Figure 7.1: Feature importance

According to this figure, it can be understood that outage cause, necessary actions taken by the crew to restore the system, protection device activated, and weather conditions take the highest importance. Nevertheless, the voltage level and outagey phases are found to be the least significant features. The results achieved in this analysis could provide the necessary background for building sophisticated outage duration estimation models.

There are a few approaches in the literature that attempt to build a predictive model for outage duration; however, a comprehensive model that can deliver a remarkable accuracy has not been developed. As a result, one interesting area for outage management is to build a predictive model for estimating the duration of outages, by the inclusion of variables discussed in the aforementioned study.

7.2.3 Some potentially interesting topics in outage management

Based on different analyses, the following areas could be some potentially important and interesting topics in outage management systems:

1. analyzing and predicting the impact of outage (i.e., number of customers affected);
2. handling the imbalanced classification problem with novel approaches;
3. analyzing transformer failure and identifying the transformers with high risk of failure;
4. analyzing and predicting the occurrence and impact of weather-related outages;
5. building probabilistic models for predicting accident (i.e., vehicle accident)-related outages;
6. investigating the applications of convolutional neural networks on images taken from trees for detecting hazardous areas.

REFERENCES

[1] P. A. Kuntz, R. D. Christie, and S. S. Venkata, "Optimal vegetation maintenance scheduling of overhead electric power distribution systems," *IEEE Trans. on Power Delivery*, vol. 17, Oct 2002.

[2] P. C. Chen and M. Kezunovic, "Fuzzy logic approach to predictive risk analysis in distribution outage management," *IEEE Trans. on Smart Grid*, vol. 7, Nov 2016.

[3] D. M. Ward, "The effect of weather on grid systems and the reliability of electricity supply," *Climatic Change*, vol. 121, Nov 2013.

[4] M. Panteli and P. Mancarella, "Influence of extreme weather and climate change on the resilience of power systems: Impacts and possible mitigation strategies," *Electric Power Systems Research*, vol. 127, 2015.

[5] A. Bahmanyar, S. Jamali, A. Estebsari, and E. Bompard, "A comparison framework for distribution system outage and fault location methods," *Electric Power Systems Research*, vol. 145, 2017.

[6] L. Xu and M.-Y. Chow, "A classification approach for power distribution systems fault cause identification," *IEEE Trans. on Power Systems*, vol. 21, Feb 2006.

[7] L. Xu, M. Y. Chow, and L. S. Taylor, "Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification e-algorithm," *IEEE Trans. on Power Systems*, vol. 22, Feb 2007.

[8] L. Xu, M. Y. Chow, J. Timmis, and L. S. Taylor, "Power distribution outage cause identification with imbalanced data using artificial immune recognition system (airs) algorithm," *IEEE Trans. on Power Systems*, vol. 22, Feb 2007.

[9] Z. S. Hosseini, M. Mahoor, and A. Khodaei, "Ami-enabled distribution network line outage identification via multi-label svm," *IEEE Trans. on Smart Grid*, vol. 9, Sept 2018.

[10] D. W. Wanik, E. N. Anagnostou, B. M. Hartman, M. E. B. Frediani, and M. Astitha, "Storm outage modeling for an electric distribution network in northeastern usa," *Natural Hazards*, vol. 79, Nov 2015.

[11] D. T. Radmer, P. A. Kuntz, R. D. Christie, S. S. Venkata, and R. H. Fletcher, "Predicting vegetation-related failure rates for overhead distribution feeders," *IEEE Trans. on Power Delivery*, vol. 17, Oct 2002.

[12] S. D. Guikema, R. A. Davidson, and H. Liu, "Statistical models of the effects of tree trimming on power system outages," *IEEE Trans. on Power Delivery*, vol. 21, July 2006.

[13] D. Wanik, J. Parent, E. Anagnostou, and B. Hartman, "Using vegetation management and lidar-derived tree height data to improve outage predictions for electric utilities," *Electric Power Systems Research*, vol. 146, 2017.

[14] M. Doostan and B. H. Chowdhury, "Power distribution system fault cause analysis by using association rule mining," *Electric Power Systems Research*, vol. 152, 2017.

[15] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, Oct 2004.

[16] B. Iglewicz and . Hoaglin, David C., *How to detect and handle outliers*. Milwaukee, Wis. : ASQC Quality Press, 1993.

[17] A. Smola and S. Vishwanathan, *Introduction to Machine Learning*. Cambridge University Press, 2008.

[18] E. Alpaydin, *Introduction to Machine Learning, Second Edition*. The MIT Press, 2009.

[19] P. Chen, T. Dokic, N. Stokes, D. W. Goldberg, and M. Kezunovic, "Predicting weather-associated impacts in outage management utilizing the gis framework," in *2015 IEEE PES Innovative Smart Grid Technologies Latin America*, Oct 2015.

[20] L. Breiman and A. Cutler, *Random Forests-Classification Description*. Department of Statistics Berkeley, 2007.

[21] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[22] R. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. 2013.

[23] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*. Springer International Publishing, 2016.

[24] T. Miyazaki and S. Okabe, "Experimental investigation to calculate the lightning outage rate of a distribution system," *IEEE Transactions on Power Delivery*, vol. 25, pp. 2913–2922, Oct 2010.

[25] P. Chen and M. Kezunovic, "Fuzzy logic approach to predictive risk analysis in distribution outage management," *IEEE Transactions on Smart Grid*, vol. 7, pp. 2827–2836, Nov 2016.

[26] D. M. Ward, "The effect of weather on grid systems and the reliability of electricity supply," *Climatic Change*, vol. 121, pp. 103–113, Nov 2013.

[27] M. Panteli and P. Mancarella, "Influence of extreme weather and climate change on the resilience of power systems: Impacts and possible mitigation strategies," *Electric Power Systems Research*, vol. 127, pp. 259 – 270, 2015.

[28] A. Piantini and J. M. Janiszewski, "Lightning-induced voltages on overhead linesâapplication of the extended rusck model," *IEEE Transactions on Electromagnetic Compatibility*, vol. 51, pp. 548–558, Aug 2009.

[29] L. Xu, M. Chow, and L. S. Taylor, "Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification e-algorithm," *IEEE Transactions on Power Systems*, vol. 22, pp. 164–171, Feb 2007.

[30] L. Xu, M. Chow, J. Timmis, and L. S. Taylor, "Power distribution outage cause identification with imbalanced data using artificial immune recognition system (airs) algorithm," *IEEE Transactions on Power Systems*, vol. 22, pp. 198–204, Feb 2007.

[31] P. Kankanala, S. Das, and A. Pahwa, "Adaboost$^+$: An ensemble learning approach for estimating weather-related outages in distribution systems," *IEEE Transactions on Power Systems*, vol. 29, pp. 359–367, Jan 2014.

[32] D. Zhu, D. Cheng, R. P. Broadwater, and C. Scirbona, "Storm modeling for prediction of power distribution system outages," *Electric Power Systems Research*, vol. 77, no. 8, pp. 973 – 979, 2007.

[33] N. Balijepalli, S. S. Venkata, C. W. Richter, R. D. Christie, and V. J. Longo, "Distribution system reliability assessment due to lightning storms," *IEEE Transactions on Power Delivery*, vol. 20, pp. 2153–2159, July 2005.

[34] A. Smola and S. Vishwanathan, *Introduction to Machine Learning*. Cambridge University Press, 2008.

[35] E. Alpaydin, *Introduction to Machine Learning, Second Edition*. The MIT Press, 2009.

[36] B. Shahbaba, *Biostatistics with R: An Introduction to Statistics Through Biological Data*, pp. 221–234. New York, NY: Springer New York, 2012.

[37] N. Balakrishnan, V. Voinov, and M. Nikulin, "Chi-squared goodness of fit tests with applications," in *Chi-Squared Goodness of Fit Tests with Applications* (N. Balakrishnan, V. Voinov, and M. Nikulin, eds.), pp. 197 – 213, Boston: Academic Press, 2013.

[38] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358 – 3378, 2007.

[39] R. Rifkin and A. Klautau, "In defense of one-vs-all classification," *J. Mach. Learn. Res.*, vol. 5, pp. 101–141, Dec. 2004.

[40] J. Burnham, R. Carlton, E. A. Cherney, G. Couret, K. T. Eldridge, M. Farzaneh, S. D. Frazier, R. S. Gorur, R. Harness, D. Shaffner, S. Siegel, and J. Varner, "Preventive measures to reduce bird-related power outages-part i: electrocution and collision," *IEEE Transactions on Power Delivery*, vol. 19, pp. 1843–1847, Oct 2004.

[41] 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, pp. 1765–1771, Nov 2009.

[42] M. Gui, A. Pahwa, and S. Das, "Bayesian network model with monte carlo simulations for analysis of animal-related outages in overhead distribution systems," *IEEE Transactions on Power Systems*, vol. 26, pp. 1618–1624, Aug 2011.

[43] L. X. and, "A classification approach for power distribution systems fault cause identification," *IEEE Transactions on Power Systems*, vol. 21, pp. 53–60, Feb 2006.

[44] L. Xu, M. Chow, and L. S. Taylor, "Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification *e*-algorithm," *IEEE Transactions on Power Systems*, vol. 22, pp. 164–171, Feb 2007.

[45] M. Doostan and B. H. Chowdhury, "Power distribution system fault cause analysis by using association rule mining," *Electric Power Systems Research*, vol. 152, pp. 140 – 147, 2017.

[46] M. Doostan, R. Sohrabi, and B. H. Chowdhury, "A data-driven approach for predicting vegetation-related outages in power distribution systems," *arXiv preprint arXiv:1807.06180*, 2019.

[47] "North carolina wildlife resources commission, available online at https://www.ncwildlife.org/,"

[48] O. Vaczi, B. Koosz, and V. Altbacker, "Modified Ambient Temperature Perception Affects Daily Activity Patterns in the European Ground Squirrel (Spermophilus citellus)," *Journal of Mammalogy*, vol. 87, pp. 54–59, 02 2006.

[49] A. E. Pankratz, *Forecasting with Dynamic Regression Models*. John Wiley and Sons, 1991.

[50] R. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. Online at http://otexts.org/fpp/, 2012.

[51] S. Kazemi, M. Lehtonen, and M. Fotuhi-Firuzabad, "Impacts of fault diagnosis schemes on distribution system reliability," *IEEE Transactions on Smart Grid*, vol. 3, pp. 720–727, June 2012.

[52] F. E. Landegren, J. Johansson, and O. Samuelsson, "A method for assessing margin and sensitivity of electricity networks with respect to repair system resources," *IEEE Transactions on Smart Grid*, vol. 7, pp. 2880–2889, Nov 2016.

[53] and L. S. Taylor, "Analysis and prevention of animal-caused faults in power distribution systems," *IEEE Transactions on Power Delivery*, vol. 10, pp. 995–1001, April 1995.

[54] O. Vaczi, B. Koosz, and V. Altbacker, "Modified Ambient Temperature Perception Affects Daily Activity Patterns in the European Ground Squirrel (Spermophilus citellus)," *Journal of Mammalogy*, vol. 87, pp. 54–59, 02 2006.

[55] D. T. Radmer, P. A. Kuntz, R. D. Christie, S. S. Venkata, and R. H. Fletcher, "Predicting vegetation-related failure rates for overhead distribution feeders," *IEEE Transactions on Power Delivery*, vol. 17, pp. 1170–1175, Oct 2002.

[56] L. X. and, "A classification approach for power distribution systems fault cause identification," *IEEE Transactions on Power Systems*, vol. 21, pp. 53–60, Feb 2006.

[57] L. Xu, M. Chow, and L. S. Taylor, "Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification *e*-algorithm," *IEEE Transactions on Power Systems*, vol. 22, pp. 164–171, Feb 2007.

[58] L. Xu, M. Chow, and L. S. Taylor, "Data mining and analysis of tree-caused faults in power distribution systems," in *2006 IEEE PES Power Systems Conference and Exposition*, pp. 1221–1227, Oct 2006.

[59] L. Xu, M. Chow, and L. S. Taylor, "Analysis of tree-caused faults in power distribution systems," in *2003 North American Power Symp.*, pp. 1221–1227, Oct 2003.

[60] Y. Cai, M. Chow, W. Lu, and L. Li, "Statistical feature selection from massive data in distribution fault diagnosis," *IEEE Transactions on Power Systems*, vol. 25, pp. 642–648, May 2010.

[61] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *SIGMOD Rec.*, vol. 22, pp. 207–216, June 1993.

[62] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, (San Francisco, CA, USA), pp. 487–499, Morgan Kaufmann Publishers Inc., 1994.

[63] M. Forghani and F. Karimipour, "Extracting human behavioral patterns by mining geo-social networks," in *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2014.

[64] M. Doostan and B. H. Chowdhury, “Power distribution system equipment failure identification using machine learning algorithms,” in *2017 IEEE Power Energy Society General Meeting*, pp. 1–5, July 2017.

[65] N. V. Chawla and *et. al*, “Smote: Synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[66] O. Maimon and L. Rokach, *Data mining and knowledge discovery handbook.* Springer, 2005.

[67] O. Maimon and L. Rokach, *Data mining and knowledge discovery handbook.* Springer, 2010.

[68] H. Ishwaran and U. B. Kogalur, “Package randomforestsrc,” *[Online] Available FTP: cran. r-project. org/web/packages/randomForestSRC/randomForestSRC. pdf*, pp. 321–357, 2015.

[69] N. Lunardon, G. Menardi, and N. Torelli, “Rose: a package for binary imbalanced learning,” 2014.

[70] M. Hahsler, B. Grun, , and K. Hornik, “The arules package: Mining association rules and frequent itemsets,” 2008.

[71] S. Kazemi, M. Lehtonen, and M. Fotuhi-Firuzabad, “Impacts of fault diagnosis schemes on distribution system reliability,” *IEEE Transactions on Smart Grid*, vol. 3, pp. 720–727, June 2012.

[72] L. X. and, “A classification approach for power distribution systems fault cause identification,” *IEEE Transactions on Power Systems*, vol. 21, pp. 53–60, Feb 2006.

[73] L. Xu, M. Chow, and L. S. Taylor, “Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification *e*-algorithm,” *IEEE Transactions on Power Systems*, vol. 22, pp. 164–171, Feb 2007.

[74] L. Xu, M. Chow, and L. S. Taylor, “Data mining and analysis of tree-caused faults in power distribution systems,” in *2006 IEEE PES Power Systems Conference and Exposition*, pp. 1221–1227, Oct 2006.

[75] D. G. Kreiss, “Fault and equipment failure analysis in distribution systems using intelligent techniques,” in *2007 IEEE Power Engineering Society General Meeting*, pp. 1–2, June 2007.

[76] P. M. Djuric, M. M. Begovic, and J. Ferkel, “Prediction of power equipment failures based on chronological failure records,” in *2006 IEEE International Symposium on Circuits and Systems*, pp. 4 pp.–1210, May 2006.

[77] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358 – 3378, 2007.

[78] O. Maimon and L. Rokach, *Data mining and knowledge discovery handbook*. Springer, 2010.

[79] N. V. Chawla and *et. al*, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[80] N. Lunardon, G. Menardi, and N. Torelli, "Rose: a package for binary imbalanced learning," 2014.

[81] C. C. Aggarwal, *Data Classification: Algorithms and Applications*. Chapman & Hall/CRC, 1st ed., 2014.

[82] M. B. Kursa, W. R. Rudnicki, *et al.*, "Feature selection with the boruta package," *J Stat Softw*, vol. 36, no. 11, pp. 1–13, 2010.

[83] R. Nilsson, J. M. Peña, J. Björkegren, and J. Tegnér, "Consistent feature selection for pattern recognition in polynomial time," *Journal of Machine Learning Research*, vol. 8, no. Mar, pp. 589–612, 2007.

[84] E. Alpaydin, *Introduction to machine learning*. MIT press, 2009.

[85] U. Stańczyk and L. C. Jain, *Feature selection for data and pattern recognition*. Springer, 2015.

[86] P.-C. Chen and M. Kezunovic, "Fuzzy logic approach to predictive risk analysis in distribution outage management," *IEEE Transactions on Smart Grid*, vol. 7, no. 6, pp. 2827–2836, 2016.

[87] M.-Y. Chow, L. S. Taylor, and M.-S. Chow, "Time of outage restoration analysis in distribution systems," *IEEE Transactions on Power Delivery*, vol. 11, no. 3, pp. 1652–1658, 1996.

[88] Y. Jiang, C.-C. Liu, M. Diedesch, E. Lee, and A. K. Srivastava, "Outage management of distribution systems incorporating information from smart meters," *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 4144–4154, 2016.

[89] K. Sridharan and N. N. Schulz, "Outage management through amr systems using an intelligent data filter," *IEEE Transactions on Power Delivery*, vol. 16, no. 4, pp. 669–675, 2001.

[90] M. Doostan and B. H. Chowdhury, "A data-driven analysis of outage duration in power distribution systems," in *2017 North American Power Symposium (NAPS)*, pp. 1–6, IEEE, 2017.