Estimation of Resource Allocation Based on Disturbance Prediction Data with Use of Statistics, Machine Learning and Data Analysis

Richard Luong, Lund University
Rebecka Domeij, Lund University

Spring 2017

Abstract

The electrical power grid is one of modern society's most important infrastructures and both power distributors and the Swedish government are investing large amount of resources to ensure continuous delivery of power. By predicting future outages with an automatic prediction system, the distributors could prevent long restoration times and economic loss. In this work, the authors evaluate the possibility of predicting outages based on statistical and machine learning methods and the relative importance of the di erent factors. The study uses open weather data and data on the power grid gathered from Swedish Meteorological Institute and E.ON, respectively.

The result shows that while maintaining the same false positive rate as currently used manual prediction methods, the automatic prediction is able to increase the true positive rate from 20% to 33%. The authors conclude that wind gust is the most important factor in predicting weather related outages and that the models are better to predict outages during stronger winds. However, further data analysis is warranted before the automatic prediction can be implemented in a real world context.

Keywords: outages, prediction, negative binomial regression, logistic regression, decision tree

Acknowledgements

We would like to thank everyone who has contributed to our work and made this thesis possible.

First, we would like to thank everyone at E.ON who has shared your valuable time, expertise, and experience. A special thanks to Mikael Hakansson for being our supervisor and contact person at E.ON for your availability and interest in our work. Without the data provided by Torbjörn Stenström and your data expertise, this master thesis would never have been accomplished.

We are grateful to our university supervisor Kalle Åström for your insights of our work and the trip to Paris, your positive and inspiring attitude and your directions when we stumbled in the dark.

Furthermore, we would like to give an honourable mention to our master thesis buddies at E.ON, Anna Rydberg and Jesper Wendell, for interesting discussions, fun lunches and supporting spirits. Thank you, Malin and David Domeij for your critical eye and proofreading of the thesis before publication. Finally, a warm thank you to all our friends and family who have supported us during this time.

Lund, May 2017

Rebecka Domeij & Richard Luong

Abbreviations

AIC	Akaike information criterion [1]
AUC	Area under the curve [2]
AUC 0.1	Area under the curve for FPR <0.1
CART	Classi cation and Regression Tree [3]
DT	Decision Tree [3]
FN	False negative
FP	False positive
FPR	False positive rate [2]
LOG	Logistic Regression [4]
NB	Negative Binomial Regression [4]
PDF	Probability distribution function
ROC	Receiver operating characteristic [2]
RQ	Research question
SMHI	Swedish Meteorological and Hydrological Institute [5]
SVK	Svenska kraftnat [6]
TN	True negative
TP	True positive
TPR	True positive rate [2]
VHI	Vakthavande ingenjor

Contents

Chapter 1

Introduction

In 2005, a storm made 730,000 people in southern Sweden powerless for up to 45 days, see [7]. The Swedish power distribution companies were not organized to handle such an extensive power outage. Hence, few customers were compensated and the society incurred severe economic loss. After this, the Swedish government [8] changed the energy legislation to strengthen the customer's rights with requirements of a maximum of 24 hours of coherent power failure and of compensation rights after 12 hours of coherent power failure. In order to follow the new legislation, power distribution companies increased their e orts to minimise the impact of weather related disturbances. Data collection and data analysis are crucial elements in this process.

By increasingly relying on data analysis in its decision process, the energy industry is following a global trend. According to Gartner [9], the leading information technology research and advisory company, one of top trends in technology in 2017 is machine learning and arti cial intelligence. These methods are used to help companies utilize data more e ciently. Gartner [10] also estimates that by 2018 more than half of the world's largest organizations will be using advanced analytics and algorithms.

This thesis is written in cooperation with the operations department at E.ON. The operations department is already using data driven decision tools to predict mechanical and operational failures in the power grid [11], but not in the case of weather related disturbances. Failing to predict weather related outages is costly. Overestimating the disturbances causes unnecessary restoration costs while underestimating them causes delays in repair and later on compensation costs to customers. Increasing the prediction accuracy is therefore highly important and this thesis investigates if this can be achieved by introducing machine learning and data driven decision tools. The operations department furthermore makes the current prediction using data on wind strength only. However, machine learning facilitates incorporating many more factors in the decision process. This thesis will therefore

also investigate the importance of including other factors in prediction of weather related outages. Specifically, the effects of historical weather data, grid structure, and ground conditions will be investigated.

1.1 Purpose and research questions

The purpose of this master thesis is first to examine and identify causalities between weather conditions and power grid disturbances and second to propose an approach for resource allocation based on disturbance prediction data. The aim of the thesis can be summarised as the following research questions (RQ):

RQ1: Can we design an automatic prediction program that increases the prediction as compared with current manual prediction methods?

RQ2: Which factors have a significant impact on disturbances in the electrical power grid?

1.2 Scope

The thesis focuses on using three different classification and regression methods. Regarding data accessibility, due to the nature of the provided historical outage data the thesis investigates only outages from 2015. Furthermore, only weather stations which have been updated between 2015-01-01 and 2016-09-30 are considered.

Regarding the number and the type of factors considered, the thesis uses existing research as basis together with human experience from the operations department. An important aspect according to the operations department is to find the variables that explain most of the disturbances. In this thesis, weather related disturbances relate to disturbances caused by falling trees and strong winds. Other weather related disturbances such as ice storms or lightning strikes are less frequent or costly to the company.

We decided to focus on developing general models, which can describe differences between regions as well as be implemented for the total grid. For this reason, only the most important variables describing regions are included. If the analysis is deemed useful, it can easily be extended to include factors related to the total power grid. The thesis only focuses on the local grid, as the regional grid is almost completely protected from falling trees. The analysis focuses on ten substation areas in the southern part of Sweden. These were chosen based on the length of overhead lines in the forest and that during the last three years the area has been affected by strong winds and disturbances.

1.3 Related articles

Zhou et al. [12] aim to help decision makers decide maintenance need for overhead lines based on weather related disturbances, speci cally daily wind gust speed and the logarithm of lightning stroke current. Zhou et al. use two methods, a Poisson regression and Bayesian network model, and conclude that historical data most likely follow a Poisson distribution due to rare events of large numbers and a variance that increases with the mean. However, the Bayesian model is easier to implement and update and it captures the distribution of failure events. Moreover, Zhou et al. nd evidence that overhead lines are the most vulnerable parts, and that weather related failures are random and di cult to completely prevent.

Billinton et al. [13] and Alvehag and Soder [14] discuss the use of two-state and three-state models based on high wind conditions. Billington et al. use a Markov approach and nd that distribution lines are highly a ected by weather conditions and that this is best captured through a three-state model with normal, adverse and major adverse weather condition states. The two-state model with only normal and adverse weather conditions underestimates the failure rate of extreme weather conditions. Alvehag and Soder use a time-varying reliability model and conclude that two- and three-state models simplify important factors. Both the failure rate and restoration time are assumed to be constant in the state-models while in the real world these factors are time-varying.

Radmer et al. [15] investigate the connection between tree-caused failures and annual average daily minimum and maximum temperature, annual daily precipitation and time since tree trimming. Using linear, multivariate and exponential regression models and an arti cial neural network model, they nd that a model which directly uses growth vegetation variables to capture failure rates is better than integrating a vegetation model with a failure rate model. The multivariate and neural network performed best based on lowest root-weighted mean-square error. The neural network captured the historical data better than the multivariate, but due to a limited data set, it had the largest prediction error. The multivariate model was slightly better at predicting unknown failure rates.

Another approach to model power outages is to use a negative binomial regression model which Liu et al. [16] use when analysing hurricane related failures in the Gulf Coast of the United States region. They conclude that falling trees cause the most damage and that maximum wind gust is related to outages, but as a single parameter, it poorly predicts their occurrence. For this reason, seven types of data were used in the model: hurricane-related outage rates, electric power system inventories, wind speed, rainfall, land cover type, tree type and soil drainage level. Including company and hurricane indicators as explanatory variables improved the t, but a drawback is that the results are not applicable to other situations and the possibility

to make predictions is limited. Liu et al. also test two di erent area resolutions: grid of one square kilometer and zip code for the United States. The zip code resolution provided predictions that are more accurate. The analysis demonstrates that high resolution data such as land cover type, tree type and soil drainage level were not providing better predictions. They nd that larger resolution better captures the variability in outages and that for the zip code model, the Poisson variability was therefore less important.

Han et al. [17] revisit the issue in Liu et al., and attempt to nd company and hurricane speci c variables that can be used instead of the indicator variables. They focus on variables describing the electric power system structure, such as number of transformers, length and type of lines and number of customers and on variables characterising the hurricane, such as landfall of the hurricane, time since previous hurricane, radius of maximum wind and central pressure di erence. The data set had both collinearity and overdispersion, which was taken into account by using a negative binomial regression and a transformation of the variables by a principal components analysis. Due to the geographical di erences between the investigated areas, di erent variables had substantial impact on the predictions. For example, the impact of the wind variables were not consistent across the regions where some of the variables had a positive correlation with outages. They discuss that a possible explanation could be that some of the regions had experienced outages caused by more ooding or thunderstorms and less by strong winds. In general, they conclude that having more overhead components leads to higher number of outages during hurricanes.

In addition to previous related literature, this thesis makes several contributions. We investigate the relationship between outages and weather conditions in Sweden using a new data set. The relationship is analysed in several ways, such as by probability distribution functions and by modelling data based on di erent wind limits, number of outages and characteristics of the region. Drawing on previous research we develop prediction models where wind limits, number of outages and characteristics of the region are important input variables. In contrast to previous literature, we use decision trees as prediction method in the analysis and additional evaluation criteria (ROC curves and AUC values). Moreover, the model comparison metrics, such as the di erent objective functions, are of interest since they relate to implementation and integration with businesses. Business integration is something that has not been addressed in previous research and is especially of interest. With this, we hope to contribute with insights for future work.

1.4 Thesis outline

In Chapter 2, brief information about the Swedish energy market and E.ON is described. Thereafter relevant theory for the analysis is presented. In

Chapter 4, the data and some of the limitations are described. The methodology and the result are presented in the two following chapters. Finally, the results are discussed and evaluated in Chapter 7. Chapter 8 contains a conclusion and Chapter 9 outlines some suggestions for future work.

For a short condensed reading of the thesis, see 2.3.2, 5.3, 6.1-6.2 and 6.4-6.6. Chapter 2.3.2 describes the routines and current situation at E.ON. Chapter 5.3 discusses the di erent objective functions used in the analysis. Chapters 6.1-6.2 and 6.6 show results related to the relationship between wind gusts and outages while 6.4-6.6 show results related to the di erent models in question.

Chapter 2

E.ON and the Swedish energy market

In this chapter, a summary of the company and the Swedish energy market is presented. Information about the Swedish energy market and E.ON has been gathered from interviewing employees at E.ON and from internal information systems.

2.1 The Swedish energy market

The Swedish energy market is regulated by the governmental transmission company, *Svenska kraftnät* (SVK), which is responsible for ensuring that Sweden's transmission system for electricity is safe, environmentally sound and cost-effective [6]. SVK ensures that the supply-demand energy relationship is in balance and owns the national grid connected to the power plants. Most of the Swedish energy production is in the northern part while the majority of the energy consumption is in the southern part.

The power is transformed from 400 kV to lower distribution voltages by several substations, before it is consumed by households and companies. Figure 2.1 gives a more detailed description of the Swedish energy distribution system. The system is divided into three parts: the national, regional and local power grid. These are distinguished by the various substations and electrical currents.

Once the electricity has been transformed at the primary substation, private companies are responsible for the delivery and quality of the energy. Around 50% of the total regional and local grid is owned by the three largest power distribution companies: E.ON, Vattenfall and Elevio.

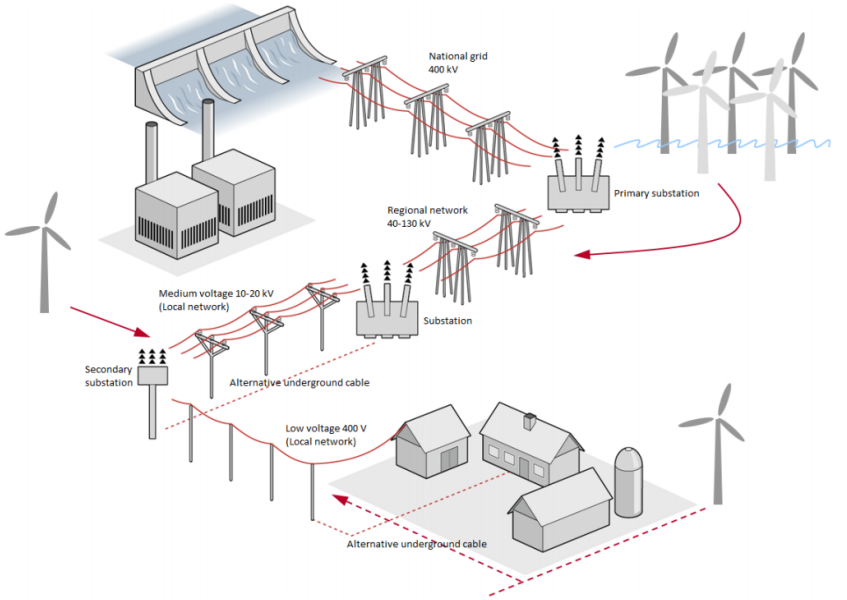

Figure 2.1: The value chain from production to consumption where the local grid is called Local network in the gure and the ten substations relate to Substation in the gure, as illustrated by E.ON [18].

2.2 Disturbances on the power grid

The higher the voltage, the more households are connected and a ected by disturbances on the power grid. Therefore, large investments are made to secure the regional grid from disturbances. Weather related disturbances, such as falling trees caused by strong winds, ice storms and lightning strikes, are unpredictable. The most harmful disturbance is falling trees, which require extensive preventive work. The lines are, for example, secured from falling trees by forest gates which ensure space between the grid and the forest. Another method is to cut down trees with dangerous height and width outside of the forest gate with the help of helicopters. However, this cost intensive e ort is mostly done on the regional grid due to the widespread area and length of the local power grid.

As described in [19], most of the local grid has forest gates but these are usually only 10 to 15 meters and thus not completely secured from falling trees. A preventive method is to have underground lines or isolated lines so that a fallen line is still functional. Both these methods are used on the local grid to some extent, but costs are too large for securing the complete grid. Hence, the reconstruction of the power grid is mostly focused on certain high risk regions to optimize investments and consequences from disturbances.

2.3 E.ON

E.ON is a global energy provider, with its headquarter located in Düsseldorf, Germany. Globally, E.ON has over 33 million customers whereof approximately one million are in Sweden. The Swedish office, located in Malmö, covers both the Swedish and Danish market. E.ON is one of the largest power grid owners in Sweden, with a power grid stretching over 130,000 km. The majority of the power grid is located in the southern areas of Sweden.

2.3.1 Operations Department

The operations department of E.ON is responsible for supervising and monitoring the power grid. Its two operation centers handle all disturbances and are responsible for communicating information to all affected parties within the company. E.ON does not maintain their power grid by themselves. Instead, the company has several framework agreements with subcontractors in different regions of Sweden. The subcontractors are responsible for the actual work and maintenance of the power grid as well as the repair in case of normal disturbance levels. In case of larger disturbances, which are often caused by strong weather conditions, E.ON can order extra work force from the subcontractors. This work force will be mobilised to restore the power grid to a functional state after the storm has passed the affected areas.

2.3.2 Routines during major disturbances

The person in charge of ordering extra work force is called *Vakthavande Ingenjör* (VHI). This person is also in charge of the organisation during major weather related disturbances. Today, E.ON has nine employees that have this role in addition to their normal duties. The VHIs have worked or are working in the operations department and have experience and knowledge about daily operations, disturbances in general and what affects the power grid. The main task for the VHI is to receive reports and forecasts from the Swedish Meteorological and Hydrological Institute (SMHI) and decide whether internal and external workforce need to be mobilised. The VHI is also the primary contact for the operations center in case of a larger disturbance and it is his or her responsibility to further initiate contact with affected parties within and outside the organisation.

The nine employees assume the role as VHI every ninth week. During this week, the designated employee receives both mail and text message alerts from SMHI if the weather forecasts indicate severe weather conditions within the next coming days. The main information source for the VHI is *Blast och Snö* [20], which is a subscription service from SMHI especially developed for energy providers. The portal contains detailed information about the coming weather with focus on wind speed, rain- and snowfall and

lightning. If needed, the VHI can also contact a meteorologist at SMHI for a more detailed forecast. With this information, the VHI decides if certain areas of Sweden need extra mobilised workforce.

In very severe weather conditions, e.g. storms a ecting large parts of Sweden, there can be a need of mobilising more than one subcontractor or even the complete organisation. In these cases, a central disturbance organisation will be formed within E.ON with the aim to coordinate the resources and to lead the reconstruction work. The operations department is in charge of this temporary organisation. Hence, the extra resource allocation needed is currently based on the wind and weather forecast from SMHI and previous knowledge and experience of the VHI about power grid disturbances.

Chapter 3

Theory

The aim of this chapter is to provide sufficient information to understand the theories and concepts used in the thesis. The different models in the thesis use both statistical and decision tree methods and are evaluated by methods such as goodness of fit, confusion matrix, ROC curves and hold out testing.

3.1 Negative binominal regression

Negative binomial regression [4] is often used when the response variable is countable and when there is overdispersion in the Poisson regression. Most regression models are derived from an underlying probability distribution function (PDF). The negative binomial regression is normally derived from a Poisson-gamma mixed distribution but can also be derived directly from the generalised linear distribution or from other mixed distributions. The negative binomial regression used in R is derived from a Poisson model with gamma heterogeneity where the gamma noise has a mean of one. R [21] is an open source programming language and software environment for statistical computing and graphics.

The variance for the mixed distribution, $V(Y) = \quad + \frac{2}{V}$, consist of the first term being the Poisson variance and the second term being the variance from the one-parameter gamma distribution. By transforming the gamma scale function, v to , the variance can be expressed in a direct relationship of and the amount of overdispersion as, $V(Y) = \quad + \quad ^2$.

The expression is called the negative binomial ancillary or heterogeneity function, which often is referred to as the overdispersion parameter. In this form, the variance is related to an otherwise normal Poisson model. The parameter is estimated using maximum likelihood and in the case of being statistically close to zero, overdispersion can be neglected and a normal Poisson model can be used. The negative binomial regression can be thought of as an extension to the Poisson model as the basic regression to model count

responses. Count responses are non-negative integers, heteroskedastic and right-skewed with an increasing variance per mean.

In some cases negative binomial models cannot fully explain the variance in the data set, i.e. the variance produced by the model is larger than the negative binomial distribution and the model is overdispersed. This is often due to the limitation that the negative binomial distribution de nes speci c expected values of count. In data sets that don't include zeros or have an excess of zeros, other forms of negative binomial and Poisson models are developed such as zero-in ated and zero-truncated models. These models are more advanced and therefore less used.

3.1.1 Overdispersion

Hilbe [4] de nes overdispersion as when the response variance is greater than the mean. The problem with overdispersion is that standard errors may be underestimated and hence predictors appear to be signi cant when this is not the case. Before dealing with overdispersion, the signi cant p-values cannot be truly accepted and need to be used with caution.

Overdispersion is tested by the Pearson-based dispersion statistic. It is de ned as the ratio of the Pearson statistic to the degrees of freedom = $\frac{^2}{n\ p}$. Here the degrees of freedom is calculated as the number of observations less predictors.

The Pearson chi-square statistic,

$$^2 = \sum_{i=1}^{n} \frac{(y_i \quad u_i)^2}{V(u_i)}; \tag{3.1}$$

is de ned as the sum of square of the raw residuals divided by the variance. The raw residual is the observation less the mean. There is no overdispersion in the data if = 1 while a value above one indicates overdispersion. For large data sets a value of 1.05 is large enough to consider the data set to be overdispersed.

Data with dispersion statistics over one are not always truly overdispersed. Apparent overdispersion can happen when the model omits important predictors, the data includes outliers, the model fails to include a su cient number of interaction terms, a predictor needs to be transformed to another scale or the link function is misspeci ed. In these cases, the overdispersion can be dealt with by for example removing the outliers from the data or transforming the explanatory variables to other scales. There are also several methods such as scaling of standard errors that deal with transforming the standard errors when they appear to be overdispersed.

3.2 Logistic regression

A logistic regression is often used when the response variable is binary, see [4]. A binary response variable has a special relationship between its mean and variance, since $\mathbb{V}[Y] = p(1 \quad p)$ is a function of the mean $\mathbb{E}[Y] = P(Y = 1) = p$.

The model can explain the often nonlinear relationship between probability p and the possible explanatory variables. The mean becomes a monotone function for each variable when all the other explanatory variables are xed where

$$p = f(x^{0};\ ) = \frac{exp(x^{0}\ )}{1 + exp(x^{0}\ )} \tag{3.2}$$

can only take a value between zero and one for all x and .

The inverse of the logistic function, logit, $log(\frac{p}{1\ p}) = x^{0}$ is linear in the parameter and can be interpreted in the same way as a linear model. Logit is therefore easier to understand than the previous Equation 3.2 where the right-hand side is a nonlinear function of the predictors. The logit can be interpreted as that $_i$ represents the e ect on the probability of increasing the i-th predictor by one unit while holding all other variables constant.

3.3 Goodness of t

Goodness of t-methods describe how well a statistical model ts a set of observations. The Pearson chi-squared test and the Akaike information criterion (AIC) are two of the methods mentioned by Claeskens et al [1]. The AIC is de ned as

$$AIC(M) = 2(loglikelihood_{max}(M) \quad dim(M)) \tag{3.3}$$

where the $dim(M)$ is length of Model M's parameter vector. A larger model will always have a higher maximum log-likelihood, since it increases with the number of parameters. However, a less complex model is preferable and by penalising the larger model with $dim(M)$ the AIC score can be used to nd the optimal model. The model with the highest AIC score is the preferred one. One drawback with AIC is that while it can be used to nd the optimal model, it does not test the quality of the models.

The optimal subset of parameters can be found using two di erent algorithms: stepwise forward selection and stepwise backward elimination. The stepwise forward method starts with an empty set of parameters and successively adds more, while the stepwise backward method starts with the total set of parameters and successively eliminates the ones with low AIC.

3.4 Decision Tree

Decision trees are often used as a classi cation method in data mining, as it serves as a white box model where every step is observable. Decision trees were rst described by Breiman et al. [3] in 1984.

The data set in a decision tree consists of objects with a set of attributes or properties. Each attribute, discrete or continuous, measures some important feature of an object. Each object belongs to one class. A simple example of a decision tree is shown in Figure 3.1. All nodes in the tree are connected through branches, with questions or tests for each possible outcome. In the example, the example questions is "*Is wind gust above 16.25 m/s?*".

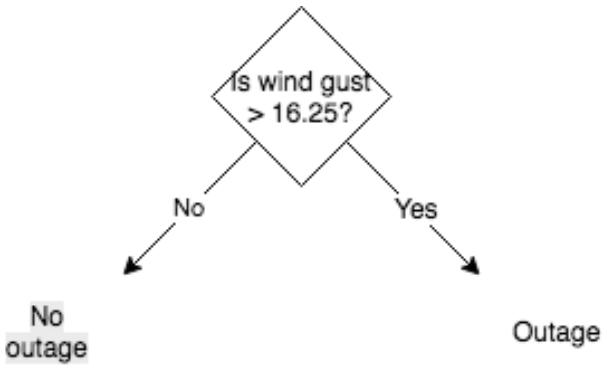

Figure 3.1: A decision tree for predicting outages. Note that this is an example and not part of the result.

The goal of the decision tree is to predict the class of an object. By starting at the root of the tree, evaluating each test and taking the appropriate branch, one can classify new objects. The process continues until a leaf is encountered and the object is asserted to belong to a class.

With a large amount of splits and increased complexity of a decision tree, the training set can be classi ed with a low amount of falsely classi ed objects. However, the essence of the decision tree is to move away from the training set and have the ability to classify future observations. If given a choice between two trees that correctly classify the training set, the simpler one is preferred as it is more likely to capture structure inherent in the problem.

There are several methodologies used for decision trees. Classi cation and Regression Tree (CART) is one of the most used algorithms for decision trees. Other common tree methods are ID3 and C4.5.

3.4.1 CART methodology

The CART methodology was developed by Breiman et al. [3] in 1984. It uses learning samples, a set of historical data with pre-assigned classes for each observation. By representing the decision tree by a set of questions, the learning samples can be split into smaller and smaller parts.

CART methodology consists of three parts: construction of maximum tree, choice of the right tree size and classi cation of new data using constructed tree. Constructing the tree is most time consuming. The tree is built following the splitting rule which splits the learning sample into smaller trees. Two splitting rules in CART are the Gini criterion and the Twoing rule. The Gini criterion works by attempting to separate classes by focusing on the largest or most important class in a node. The Twoing rule focuses on segmenting the classes into two groups that add up to 50 percent of the data. Example of these two rules are shown in Figure 3.2 and 3.3.

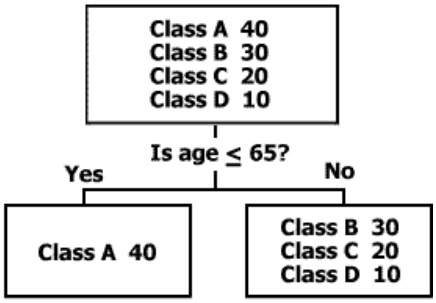

Figure 3.2: The Gini criterion classi es the largest or most important class in one node, as illustrated by Salford Systems [22].

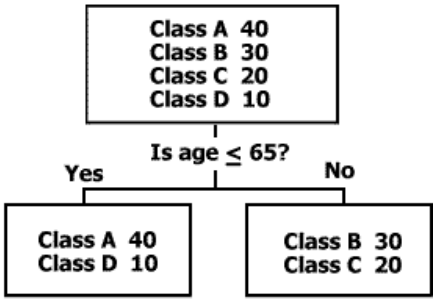

Figure 3.3: The Twoing rule separates into two equally large nodes, as illustrated by Salford Systems [23].

Advantages of CART are that the methodology is non-parametric and does not require the input variables to be speci ed in any speci c form and that it works well with both continuous and categorical variables. CART does not require the variables to be selected in advance, and will eliminate

the insigni cant ones from the nal decision tree. Another advantage of decision trees in general are that they are easy to interpret.

Disadvantages are that all splits produced by CART are perpendicular to axis and if the underlying data have a more complex structure, it may not be caught by the methodology. According to Timofeev [24] and [25], if the underlying data is nonlinear, the methodology might not be suitable and a more complex tree is needed to explain the data.

3.5 Confusion matrix

A confusion matrix, or contingency table as described by Fawcett [2], is a matrix commonly used within statistics and machine learning for visualizing the performance of an algorithm or a classi cation method. The columns represent the predictions of the algorithm classifying the case of an event occurring and not occurring, while the rows represent the actual outcome. The matrix comprises the number of true positives, false positives, false negatives and true negatives.

True positive (TP) if the instance is positive and classi ed as positive
False positive (FP) if the instance is negative but classi ed as positive
False negative (FN) if the instance is positive but classi ed as negative
True negative (TN) if the instance is negative and classi ed as negative

False positive and false negative are also called Type I error and Type II error, respectively. Based on these counts, several metrics can be calculated, such as the true positive rate (TPR).

$$TPR = \frac{true\ positive\ events}{total\ positive\ events} \tag{3.4}$$

is calculated as the true positive events divided by the total number of positive events. The total positive events consists of all TP and FN.

3.6 ROC curves

Receiver operating characteristic (ROC) curves [2] are a common way of investigating the performance of a binary classi er outputted score. The curve is acquired by plotting the true positive rate and false negative rate of each method. A tuning parameter is introduced and a ects the outcome by biasing the classi er and changing its decision boundary.

The theoretically best classi er would have its ROC curve close to the top left corner, where the true positive rate is one and the false positive rate (FPR) is zero. This point is in practice impossible to reach, since classi ers are usually never perfect. Closeness to the top left corner can be measured

by the area under the curve (AUC). This is a metric to transform the two-dimensional representation of the ROC curve into a scalar measure.

One should aim for as high true positive rate as possible to the lowest false positive rate. If a model has a higher true positive rate at a certain false positive within a specific range, this model should be the most preferable in this range. A hybrid model can be created whose combined AUC is larger than the two classifiers separately. Research by Flach and Wu [26] show that any concavities in a hybrid model can also be equalised.

3.7 Hold out test

When training and testing different models, it is common to divide the data set into two sets. Hold out testing is a popular way of testing the accuracy of a forecasting method. This is done by choosing two independent sets as training and test sets, respectively. The training set is used for training the model, and the other one for testing. By conducting a test with a data set that has not been used in training, one can measure the model's performance on unseen (future) data. According to Var [27] using hold out testing is a good way to validate the estimated function and make sure it performs well when applied on another sample. However, Tan et al. [28] point out that having a holdout sample reduces the estimation sample.

The size of the test set should correlate to the forecast horizon. For classification problems, one typically uses stratified sampling, so that the test set contains roughly the same proportions of class labels as the original sample. K-hold out test is a variant of the standard hold out test, where k% of the original data set is used as test set. Sampling is done without replacement. The hold out is usually repeated n times, yielding n random partitions of the original sample. The n results are averaged (or otherwise combined) to produce a single estimation.

Chapter 4

Data description

The data used in this thesis is provided by E.ON and SMHI. The SMHI data is public and available through their open API. [29]

The initial data set contains outage data from year 2015 to 2016 for ten chosen substations in Sweden. The weather data is from the same time period. The time period was chosen because the system E.ON uses today was implemented in 2014 and full data exists from this time period and onwards. Data before 2015 exists, but is not stored in the same way.

4.1 Weather data

The variable used is wind gust (m/s), which is de ned as the highest 2-second wind during an hour by SMHI [30]. The data is gathered from the closest weather station which is based on the shortest distance on the surface between the substation and the weather station. Each weather station is updated hourly. One limitation with the data set is that the location of the weather station and the power grid is most likely di erent and therefore an observed data point from the weather station might not describe the exact weather condition at the power grid. This problem could be overcome by for example interpolating the data between the weather stations by using a geographical information system program.

4.2 Disturbances

When there is a disturbance on the regional or local grid down to 10 kV the disturbance is registered in one of E.ON's databases. The outage is logged with an outage ID and other relevant information such as number of a ected customers, cause of disturbance, a ected components, feeder and the location of the outage. The exact location of the broken line is unknown and instead the outage is located to the closest protective device. The disturbance could be between 0.5 and 2 kilometers from the protective device

depending on the structure of the grid. The line could also be broken at several places that all have the same protective device. Since most of the grid structure is circular, the broken line can be switched o and the electricity can be transported through another line. By switching on and o lines combined with human investigation the exact location can be found. Disturbances on the local grid with lower voltage than 10 kV are not registered automatically in the system, but registered when customers report that they are powerless.

There are some limitations with the disturbance data set that can a ect the result. First, the exact location of the disturbance is unknown and one outage ID can actually capture several disturbances. Second, the cause of disturbance is manually adjusted in the system and categorised after the outage is resolved. Therefore, there is a risk that the categorisation of data used in this thesis, is not correct and that there are other categories which could explain falling trees on the lines.

4.3 Substation area information

Data about the grid structure combined with the terrain in the substation area is provided by E.ON. There exists information about the type and length of the line in a substation area. Lines are speci ed as belonging to three di erent types: isolated overhead line, not isolated overhead line and underground line. E.ON uses the terrain map from *Lantmateriet*, the Swedish National Land Survey [31], for risk analyses and grid planning. The data is speci ed by both the type of line and the type of land it passes through in a substation area. The information is speci ed as meter of length. A higher land resolution could a ect the result positively but due to implementation and integration reasons for E.ON, this type of data was not possible to obtain.

4.4 Data preprocessing

The data described in chapters 4.1-4.3 was downloaded into separate les and then preprocessed and merged into one single data le. Three di erent time periods were used (one, six and 24 hours) and thus three di erent data les were created. Each row in the data le represents one single time period for a speci c substation. The state of the other parameters for this speci c time and substation was then added to the row. Outages are logged both as countable and binary for each row in the data set. There could occur several outages during the di erent time periods since the outage is logged with the exact time. Rows with missing values are removed to reduce bias in the data set.

Chapter 5

Method

We have aimed to develop generalised models which can easily be applied to other geographical areas. This paper mainly focuses on a subset of power stations in Sweden, but we believe that the models can cover a larger area with similar results.

5.1 Choice of parameters

The parameters in question for model design are displayed in Table 5.1, where Variable is the name of the parameter, Variable description describes the parameter and Type describes the format of the parameter.

Table 5.1: Parameters in question for model design.

Variable	Variable description	Type
y_1	Outage	Binary
y_2	Number of outages	Integer
x_1	Wind gust	Real number
x_2	Length of overhead line, isolated	Real number
x_3	Length of overhead line, not isolated	Real number
x_4	Length of overhead line, in forest	Real number
x_5	Length of overhead line, not in forest	Real number
x_6	Length of overhead line, in forest, isolated	Real number
x_7	Length of overhead line, in forest, not isolated	Real number
x_8	Length of underground line, in forest	Real number
x_9	Length of underground line, not in forest	Real number
x_{10}	Substation ID	String

There are two responses, y_1 and y_2, which are used in different models. Negative binomial regression requires a countable response variable, while logistic regression requires a binary response variable. For evaluation reasons, the models that use y_2 as response variable are converted into y_1.

Of the eight explanatory variables, only five were used in the final models. The name of the substations, x_{10}, was not included. This was because of previous work, Liu et al. [16] and Han et al. [17], which discuss the issues with description variables and that the attributes that characterise a station can be explained by other variables. By using parameters as forest type distribution and length of each power line type, the substation ID variable could be omitted.

Other parameters that were not used are those related to underground line since these parameters are not affected by wind. Hence, x_8 and x_9 were not included. Variables x_6 and x_7 are subsets both of x_4 and of x_2 and of x_3, respectively. Several initial tests with different combinations of the variables were conducted where a combination of x_1, x_2, x_3, x_6 and x_7 provided the best results. To conclude the parameters which are considered in the final models are

$$y_1 | | | y_2; x_1; x_2; x_3; x_6; x_7; \tag{5.1}$$

where the response variable changes for the negative binomial regression.

5.2 Implementation of models

A set of different models was developed based on combinations of three different classification and regression methods, one wind gust limit and three different time periods. In total, these combinations resulted in 18 different models.

The different time periods were chosen to one, six and 24 hours. The reason is that the original data received from SMHI is per one hour, the prognoses the VHI receives from the service *Blast och Snö* is updated every six hours and finally, the VHI does a prognosis on a daily basis. The parameter x_1,which describes wind gust, was adjusted for each time period where the maximum value for the time period was used. All the other explanatory variables were not affected by the different time periods.

A hold out test was conducted where all models were trained on data for 2015, while data for 2016 was used for testing and evaluation. For each time period, two models were developed. One model which tried to explain all future data and another that only concerned future data with wind gusts over 15 m/s. The wind limit was set to 15 m/s because the first warning from *Blast och Snö* starts at 15 m/s and because the VHI is only concerned about wind gusts above this limit.

The regression models were evaluated using the AIC-criteria where the model resulting in lowest AIC was used as the final model. The stepwise-forward method was used to find the optimised model due to computational reasons and the small data set used for training the methods.

The decision tree models used the CART methodology and the Gini criterion as the splitting rule. The models used a maximum number of

splits of 10 and a cost function of 1:10 between the false positive and false negative error. Several di erent splits were investigated where the 10-splits provided the best result in relationship to complexity. The cost function is related to the objective function which is described in Chapter 5.4.1.

5.3 Final models

The table below shows the nal models used and their respective model number which will be used as reference throughout the thesis. The method of the models are decision tree (DT), negative binomial (NB) and logistic regression (LOG).

Table 5.2: The nal models with their respective model number.

Model	Method	Time	Wind
1	DT	1	0
2	DT	6	0
3	DT	24	0
4	NB	1	0
5	NB	6	0
6	NB	24	0
7	LOG	1	0
8	LOG	6	0
9	LOG	24	0
10	DT	1	15
11	DT	6	15
12	DT	24	15
13	NB	1	15
14	NB	6	15
15	NB	24	15
16	LOG	1	15
17	LOG	6	15
18	LOG	24	15

5.4 Model evaluation

In order to evaluate the di erent models, an objective function was developed. As a complement, ROC curves were plotted and AUC calculated. The estimation behind the objective function is described in the two following chapters.

5.4.1 Objective function

By minimising the objective function, the model with the best performance can be found. This objective function sets the proportion between the cost of a real outage not being classi ed as an outage and the cost of falsely predicting an outage when nothing happened in reality. Hence, the function weights the false positive and false negative errors for each model.

The cost associated with the false positive error is the cost of mobilising extra workforce,

$$c_{fp} = cost \quad number\ of\ workers = 2{;}500 \quad 4 = 10{;}000; \tag{5.2}$$

which is an estimation based on the following assumptions. The cost is regulated through framework agreements between E.ON and its subcontractors. Since these agreements vary between subcontractor areas, a mean cost of 2,500 SEK per worker is used. According to several VHIs, four workers per a ected area is usually ordered. Therefore, a mean number of four workers is used in the cost estimation. It should be noted, that the cost are approximations and the estimated cost may di er from real costs. However, the estimated cost makes it possible to compare di erent models in relation to each other.

Putting subcontractor workers in preparation is a way of minimising the risk of not being prepared in case of an outage. It is when an outage occurs and the subcontractors are not prepared that the large costs occur for E.ON. This is due to the change in energy legislation for power failures and other costs associated with outages, such as goodwill and damage costs.

When a customer is a ected by an outage for over 12 hours, E.ON will compensate the customer with a minimum of 900 SEK. For every additional day with coherent power failure, the compensation will increase with the same minimum amount per day according to public information from E.ON [32]. The other costs are di cult to estimate and therefore the cost associated with the false negative error is limited to the compensation cost. This is represented as

$$c_{fn} = \frac{compensation \quad number\ of\ customers}{number\ of\ outages} = 100{;}000 \tag{5.3}$$

whereof it is the minimum cost E.ON needs to pay per outage. The mean compensation per outage is calculated by the total compensation paid between 2015 and 2016 divided by the number of outages during the same period.

The cost constants are summarised into the following objective function

$$c = n_{fp} \quad c_{fp} + n_{fn} \quad c_{fn}; \tag{5.4}$$

where n_{fp} and n_{fn} are associated with the number of false positive and false negative errors while c_{fp} and c_{fn} are the costs associated with the two

types of errors. The two cost constants have a factor of 1:10 and hence, the relationship between the false positive and false negative errors is 1:10.

The objective function is adjusted to take into account the time period for each data set. A false prediction for an outage in the one hour data set is not believed to be equivalent to one in the 24-hour data set. In reality, a VHI needs to decide if extra workforce is required two days in advance to cover the non-normal working hours (evenings and weekends) and the person makes only one decision every 24 hours. Therefore, a model cannot be punished more than once every 24 hours for falsely predicting an outage or failing to predict an outage. The constants in the adjusted objective function are divided by the number of times each time period occurs per 24 hours. Thus, the objective function for the one-hour data set is divided by 24 and the six-hour data set is divided by four,

$$c_{Adjust} = n_{fp} \quad \frac{c_{fp}}{(1 \; or \; 4 \; or \; 24)} + n_{fn} \quad \frac{c_{fn}}{(1 \; or \; 4 \; or \; 24)} \tag{5.5}$$

and these adjusted constants are summarised in Table 5.3. Notice that the constants for the objective function are the same as for the 24 hour adjusted objective function.

Table 5.3: Constants used in the adjusted objective function.

Constant	Value
$c_{fp;24h}$	10000
$c_{fn:24h}$	100000
$c_{fp;6h}$	2500
$c_{fn;6h}$	25000
$c_{fp;1h}$	417
$c_{fn;1h}$	4167

5.4.2 VHI objective function

By estimating the current objective function for E.ON, a reference point can be made to all the other models. The VHI objective function is calculated with the same cost estimations as the objective function. The number of occurrences, i.e. the number of false positive errors and false negative errors, are estimated by cross-checking all the times a VHI had ordered extra workforce within the dates of outages. Since the VHI only makes a decision once, as described above, the VHI objective function is calculated based on the 24-hour data set. Also, the VHI bases his or her decision on *Blast och Snö*'s first warning at wind gust forecasts above 15 m/s and hence only data above 15 m/s is relevant.

Only data for 2016 was available and the VHI objective function is therefore calculated for 2016-01-01 to 2016-09-30. During this time period, extra

workforce was ordered two times in the investigated subcontractor area of which one was a correctly predicted outage. The wrongly predicted outage happened during the high season for strike lightning and will therefore not be considered as a false positive error. Hence, in the VHI objective function there are no false positive errors. During the same period, ve outages occurred in total of which one was predicted. Thus, the false negative error is four. The number of times where there was no outage and the VHI did not order extra workforce were 142 in the data set. This is displayed in the confusion matrix in Table 5.4. It is important to notice that the confusion matrix for the VHI consists of no false positive errors.

Table 5.4: Confusion matrix of the VHI during 2016-01-01 and 2016-09-30

	Not order workforce	Order workforce
No outage	142	0
Outage	4	1

5.5 Relative importance of variables

Four tests are conducted to investigate the relative importance of the parameters and their characteristics. Three of the tests analyse if higher wind gusts, overhead lines in forest and if not isolated overhead lines are better predictors than their opposite. The fourth test analyses if there are similarities or di erences between substations based on number of outages. All tests are based on the total data set and hence, a hold out test is not conducted.

The wind gust test is done by splitting the data set by the wind gust limit of 15 m/s. The models are based on data with wind gusts below 15 m/s and data with wind gusts above 15 m/s, respectively. For the substation test, the data set is divided based on the number of outages. The models are based on data with ve substations with the highest number of outages and on data with ve substations with the lowest number of outages.

The two other tests are conducted di erently. Instead of splitting the data set, di erent parameters are included in the models. In the forest test, models including parameter x_1, wind gust, and x_4, length of overhead line, in forest, are tested against models including parameter x_1, wind gust, and x_5, length of overhead line, not in forest. Similarly, the type of line test, tests models including parameter x_1, wind gust, and x_3, length of overhead line, not isolated, against models including parameter x_1, wind gust, and x_2, length of overhead line, isolated.

To see if the models better predict outages based on any one of the data sets or parameters, the average of each comparison metrics is calculated. This since the result from each speci c method is not of interest, but the relative importance of the variables are.

Chapter 6

Results

The aim of this chapter is to show concise results to highlight the important insights from the conducted analysis. The various tests in this chapter aim to answer the RQs, where Chapter 6.4-6.5 address RQ1, while Chapter 6.1-6.2 and 6.6 address RQ2.

The tables are truncated, due to limited space and ease of readability. The complete result tables can be found in Appendix for the interested reader.

6.1 Plots of wind gusts and outages

To understand the relationship between outages and wind gusts, plots of wind gusts against outages are produced. In Figure 6.1, it can be seen that outages have happened on a wide spread of wind gusts, from less than 5 m/s to as high as 30 m/s. In addition, outages have occurred during the whole period, but with two large numbers of outages in January 2015 and December 2015. This coincides with two large wind gust peaks above 25 m/s. However, observe that there is a wind gust peak above 25 m/s in February 2016, but without the large number of outages seen during 2015. In addition, the black horizontal line in Figure 6.1 marks the 15 m/s wind gust limit used in splitting the data for other tests conducted in the thesis.

The same plot of wind gusts and outages is produced for each substation area, which can be seen in the subplots of Figure 6.2. Note that the wind gust distribution is not equally the same for all substations, but some substations have been exposed to a larger number of outages such as Substation 4, 6, 7, 8, 9 and 10. In January 2015, the storm *Egon* hit Sweden and as seen for each substation area at least one outage occurred at that time. When the storms *Gorm* and *Helga* hit Sweden in the end of November in 2015, all substations except Substation 1 and 3 had outages. In addition, observe that all substations had outages on a wide spread of wind gusts and during the whole period.

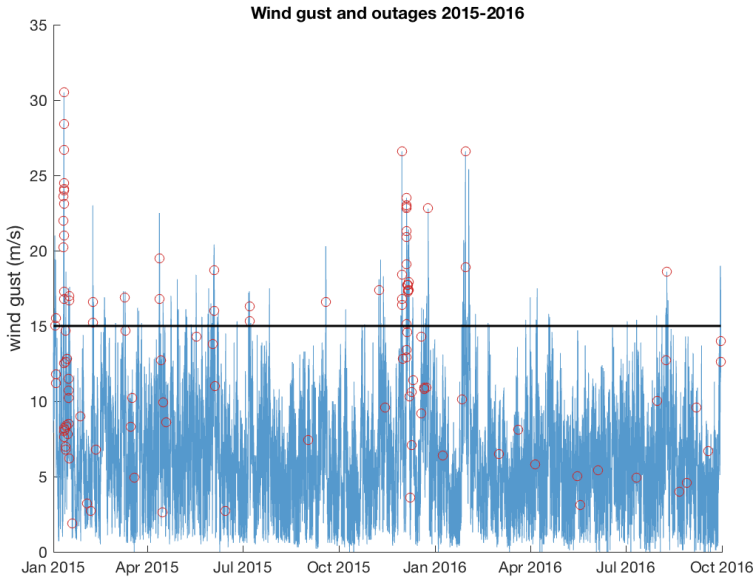

Figure 6.1: Plot of outages against wind gusts with the limit of 15 m/s marked as a black line.

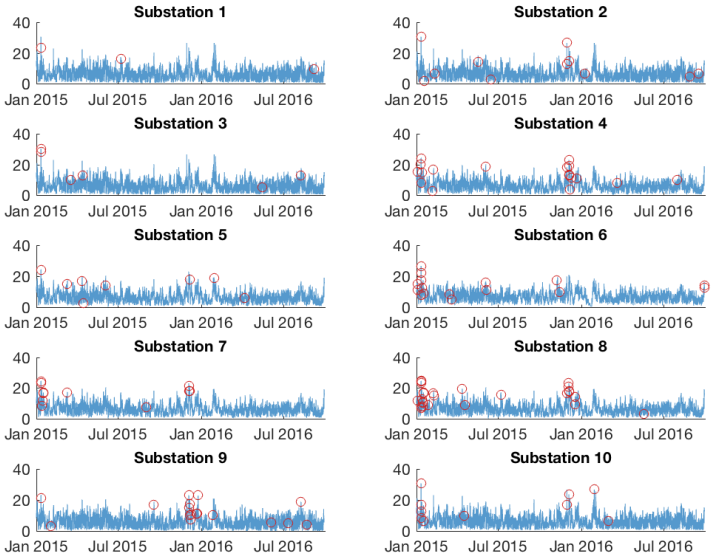

Figure 6.2: Plots of outages against wind gusts for each substation area.

6.2 Probability distribution function

Another way of presenting the relationship between wind gusts and outages is to consider the probability distribution function for wind gusts and outages and to visualise them. The histogram for wind gust is shown in Figure 6.3. Notice that the distribution of wind gust seems to be right-skewed Gaussian. This indicates that the likelihood for wind gusts to reach above 15 m/s or higher is relatively low. Of the 152,875 data points, only 1.6% were above 15 m/s and 0.1% were above 20 m/s.

The PDF for wind gusts and outages are visualised together in Figure 6.4. Taking into account that the frequency of strong wind gusts is low, the probability that an outage happens is relatively high for wind gusts above 15 m/s. With the same reasoning, the probablity of outages below 15 m/s is approximately zero, even though there are a total of 72 outages in this interval.

A 90% one-sided con dence interval is also calculated, which corresponds to the dashed line in Figure 6.4. Since the con dence interval is wide, this indicates that the shown probability distribution function cannot statistically be said to be true.

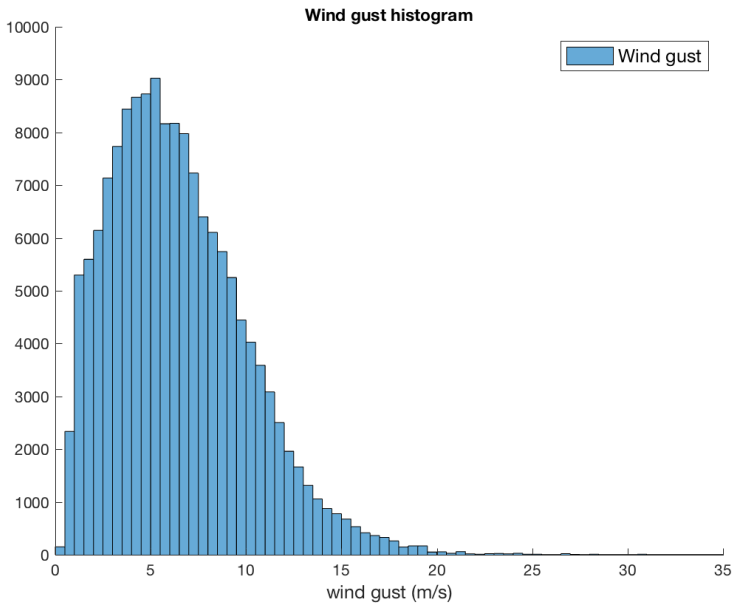

Figure 6.3: Histogram of the frequency of wind gust between 2015-01-01 and 2016-09-30.

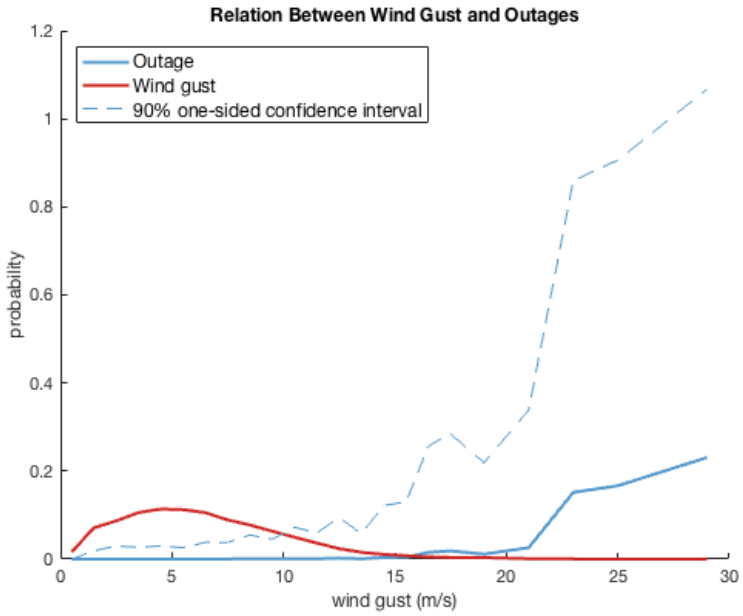

Figure 6.4: The probability distribution function for wind gusts and outages where the dashed line is a 90% one-sided con dence interval.

6.3 Overdispersed data

A Poisson regression model is developed for the data set with least number of zeros, i.e. the 24-hour data period with wind gust limit of 15 m/s. The Pearson-based dispersion statistic, = 1:5931, is calculated for the Poisson model. The result indicates that the data set is overdispersed.

6.4 Model comparison metrics

The three best models and the worst are displayed together with the VHI model for both the objective function, the adjusted objective function and the true positive rate. The tables consist of eight columns, where Method describes the type of method used, such as VHI, DT, NB and LOG. Time describes the time period of one, six and 24 hours and Wind describes the wind limit of 15 m/s or none. The last four columns show each model's confusion matrix. All tests are sampled three times and the average result is shown below. This was due to that TP, FP, TN and FN are generated by drawn samples for the negative binomial and logistic probability distribution. Hence, each sample changes slightly for each simulation.

In tables 6.1 and 6.2, the best models for the objective function and the adjusted objective are the same. It is important to make a remark on the

models' confusion matrices. The best model predicts one true positive count while the second and third best models have zero true positive counts. The false negative counts are expensive and hence the second and third model are more expensive despite having fewer false positive counts. The magnitude of the objective and adjusted objective function is shown from the worst model's result. Notice that the VHI model is among the top four models for the objective function, while this is not the case for the adjusted objective function.

Table 6.1: The three best models, the worst and the VHI model based on the objective function.

Method	Time	Wind	Objective function	TN	FP	FN	TP
VHI	24h	15	400,000	142	0	4	1
DT	1h	15	350,000	723	15	2	1
LOG	1h	15	410,000	727	11	3	0
NB	1h	15	420,000	726	12	3	0
			...				
LOG	1h	0	2,470,000	65392	47	20	0

Table 6.2: The three best models, the worst and the VHI model based on the adjusted objective function.

Method	Time	Wind	Adjusted objective function	TN	FP	FN	TP
VHI	24h	15	400,000	142	0	4	1
DT	1h	15	14,583	723	15	2	1
LOG	1h	15	17,083	727	11	3	0
NB	1h	15	17,500	726	12	3	0
			...				
DT	24h	0	2,280,000	2651	68	16	4

The best models for both the objective and the adjusted objective function are not included as one of the three best models for the last metric, the true positive rate. In Table 6.3, the best models are a mix of DT and NB models.

Table 6.3: The three best models, the worst and the VHI model based on the true positive rate.

Method	Time	Wind	TPR	TN	FP	FN	TP
VHI	24h	15	0.200	142	0	4	1
DT	24h	15	0.800	91	51	1	4
DT	6h	15	0.500	200	43	2	2
NB	24h	15	0.400	123	19	3	2
			...				
LOG	1h	0	0.000	65392	47	20	0

Notice that the best models for all three metrics are composed of the 15 m/s wind limit, while the worst model is composed of the 0 m/s wind. Comparing the di erent methods, decision tree is the best method for all three metrics. The di erence between the three methods is smaller for the objective functions than for the true positive rate metric. The VHI model is not the best model for any metric. In Table 6.4, the three models with time interval of 24-hour and wind limit are compared to the VHI model. Notice that the objective function for the VHI is better while the TPR is worse compared to at least two of the models. The objective function of NB is relatively close to the VHI and its TPR is twice as the VHI but it is still relatively low. DT has four times better TPR than the VHI but the objective function is 1.5 times worse.

Table 6.4: The models with 24-hour time period and wind limit of 15 m/s.

Method	Time	Wind	Objective function	TPR	TN	FP	FN	TP
VHI	24h	15	400,000	0.200	142	0	4	1
NB	24h	15	490,000	0.400	123	19	3	2
LOG	24h	15	560,000	0.200	126	16	4	1
DT	24h	15	610,000	0.800	91	51	1	4

6.5 ROC curves

ROC curves are plotted to compare the models with each other and with the VHI model. In gures 6.5-6.8, each method has a speci c colour where di erent shades of green, red and blue describe decision tree, negative bi-

nomial regression and logistic regression, respectively. The VHI model is viewed as a red dot in all plots.

In Figure 6.5, all models are plotted. The main interpretation is that no single model greatly outperforms any of the other models. The di erent line styles in the gure relate to the three time periods: solid line for one, dashed for six and dotted for 24 hours.

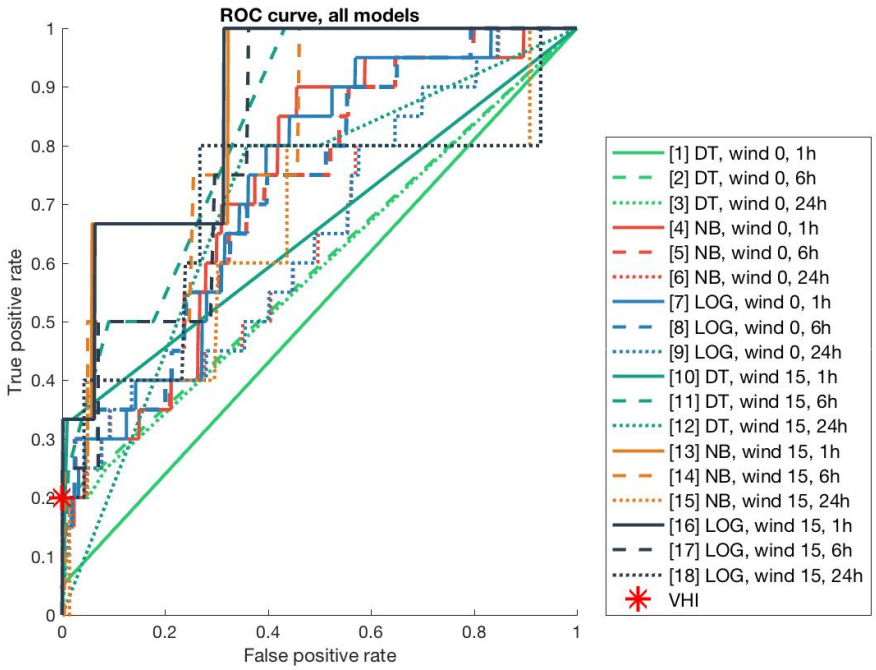

Figure 6.5: ROC curves for all models, where the red dot is the performance of the VHI. None of the models greatly outperform the others. Model 13 and 16 seem to be closest to the left-top corner.

Figure 6.6 shows a portion of Figure 6.5, where the performance of the VHI compared to the models is more clearly shown. With the same false positive rate as the VHI, there is only one model with a greater true positive rate, Model 17. This model has a rate of 0.25, which is 0.05 higher than the VHI model. Model 18 has the same performance as the VHI. Both Model 13 and 16 have approximately a zero false positive rate with a true positive rate of 0.33 which is over 1.5 times better than the VHI. All other models show a worse performance in the zero false positive rate point.

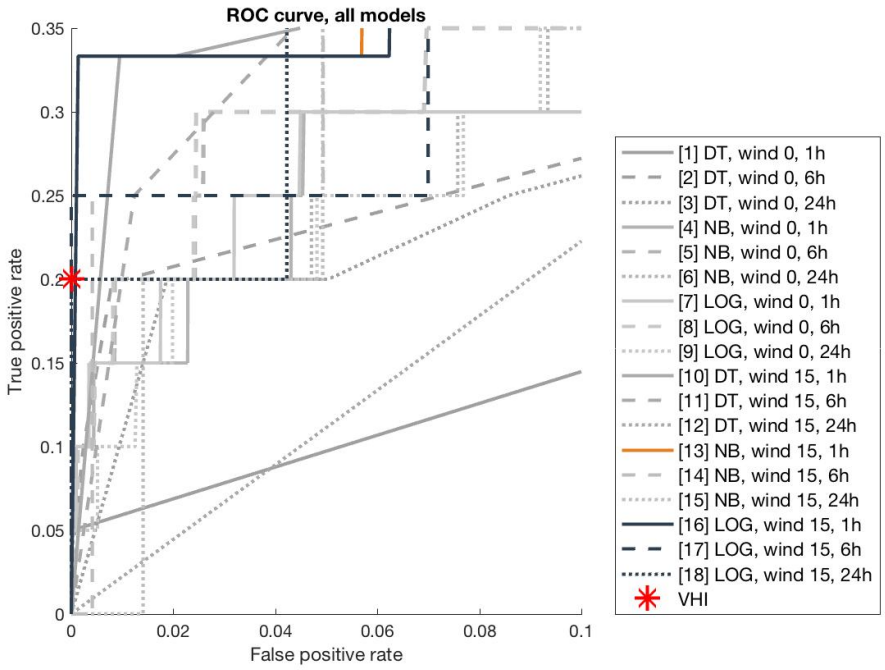

Figure 6.6: Portion of ROC curves for all models. At the same FPR, Model 17 has better performance and Model 18 has the same performance as the VHI. Model 13 and 16 have approximately a zero FPR and therefore outperform all other models, including the VHI, with a TPR of 33%.

Figure 6.7 comprises three subplots where each subplot shows all models where the data is aggregated into one specific time period. Throughout the various subplots, the red and blue lines representing NB and LOG, show very similar performance. This is especially true for the solid lines which represent the models with no wind limit.

NB and LOG models are in most cases better than the DT models. This is especially true for the one-hour time period, where both green lines are below the red and blue lines. The green solid line, representing DT without wind limit, has almost the same performance as if one would classify randomly. In the six-hour and 24-hour time period, the DT model with wind limit performs better than NB and LOG without wind limit. In the one hour time period, LOG and NB perform better than both the DT models.

The difference between models with a wind limit of 15 m/s and models without is relatively large. The models with wind limit perform better since they are closer to the top-left corner for all time periods. None of the models outperform the other models and NB and LOG is only significantly better than DT for the one hour time period.

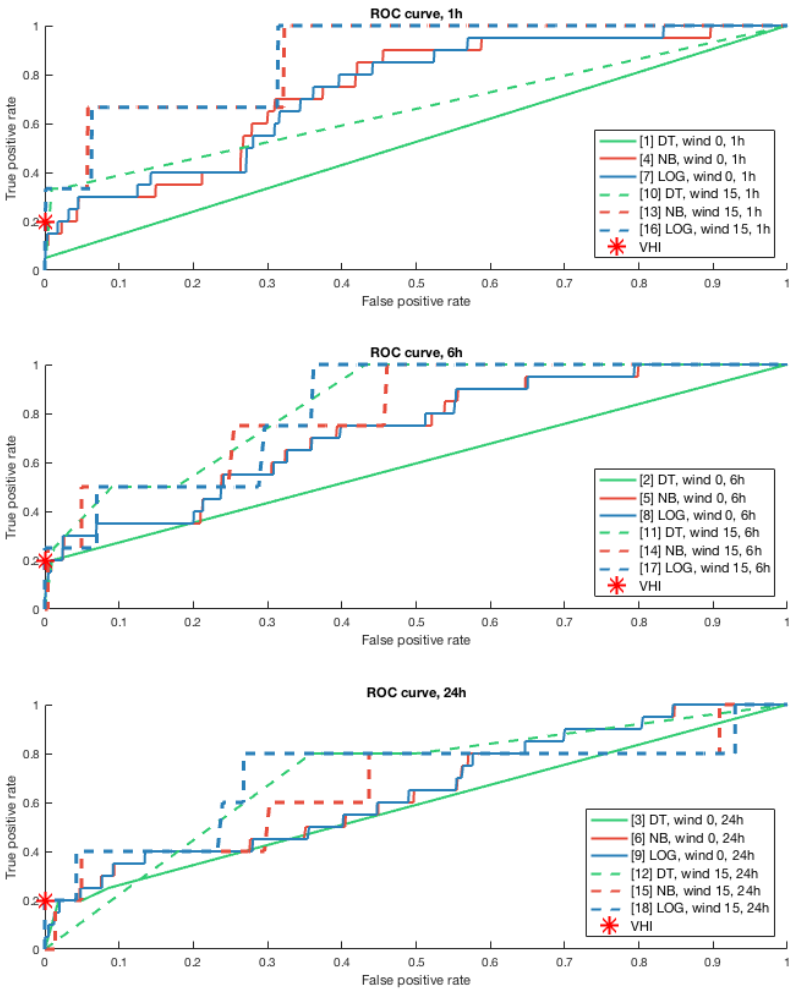

Figure 6.7: ROC curve for models with one-, six- and 24-hour time period. Note that NB and LOG show a similar performance which decrease for higher time periods, while DT performs better, especially the dashed line. The dashed lines are more left-corner centred and thus, the models with wind limit perform better.

In Figure 6.8, all models without a wind limit and all models with a wind limit of 15 m/s are shown in the top and bottom subplot, respectively. NB and LOG are better for all the time periods for the models without a wind limit. The one-hour time period gives the best performance for these models while the six-hour time period gives the best performance for DT.

This is also true for the models with a wind limit. An interesting finding is that NB and LOG perform better with a lower time period.

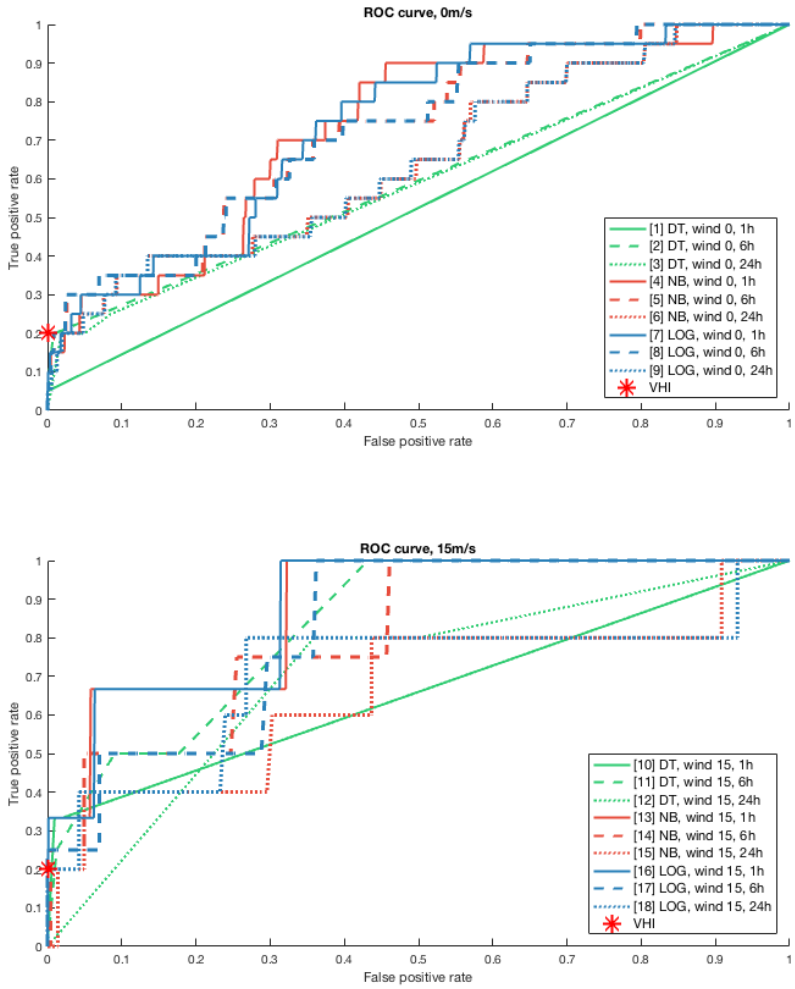

Figure 6.8: ROC curve for models with 15 m/s as wind limit and without. Note that the NB and LOG have similar performance and outperform DT in the data set without wind limit. With wind limit, none of the models distinguishably outperform the others, where the best model for DT is the six-hour data set while the one-hour data set is best for LOG and NB.

6.5.1 AUC

In Table 6.5, AUC is measured for the three best models and the worst. Note that the three best models are based on the one- and six-hour data set with wind gusts above 15 m/s. These have very similar AUC values where the LOG model performs slightly better than the NB model.

Table 6.5: The three best models and the worst based on AUC.

Model	Method	Time	Wind	AUC
13	LOG	1h	15	0.874
16	NB	1h	15	0.873
11	DT	6h	15	0.833
...	...	...	...	...
1	DT	1h	0	0.525

The VHI model has a zero false positive rate, and therefore high values of false positive rates are probably irrelevant. In order to adjust for this, AUC is calculated up to a false positive rate of 0.1. The three best models and the worst based on the smaller AUC are summarised in Table 6.6. The same models as for the total AUC give the best performance. LOG and NB models have the largest area under their ROC curves. Notice that Model 13 and 16 had approximately a zero false positive rate and outperformed the VHI model, as seen in Figure 6.6 in the earlier chapter. Those models also have the highest AUC up to the false positive rate of 0.1.

Table 6.6: The three best models and the worst based on AUC 0.1.

Model	Method	Time	Wind	AUC 0.1
13	NB	1h	15	0.047
16	LOG	1h	15	0.045
11	DT	6h	15	0.036
....	...	...	...	...
10	DT	1h	15	0.010

6.6 Further analysis

To further investigate the variables' impact on outages, different test are carried out and tabulated in 6.7-6.10. The first test focuses on analysing the relationship between wind gusts and the occurrence of outages, while the other tests focus on the relationship between grid structure, ground condition and the occurrence of outages.

The metrics used for comparison are the mean of the model comparison metrics and the AUC metrics. It is important to remember that a low value of the objective functions is preferred while a high AUC value is preferred.

These should all be considered in relation to the TPR which indicates the accuracy of the models, where a higher value is preferred. Mean number of outages in each data set is also included for the first two tests. All tests were sampled three times respectively and the average result is shown below.

The first test is summarised in Table 6.7. In the test, the performance of the classification and regression methods are measured based on data with wind gusts below 15 m/s and data with wind gusts above 15 m/s. The $(15; \infty)$ interval outperforms the $[0; 15]$ interval for all metrics. Since the objective functions are lower and the AUC value is higher, the methods are better at predicting outages in data with strong wind gusts than in the lower interval. In addition, the mean number of outages is higher for the $(15; \infty)$ interval, which also confirms that the methods are better at predicting outages in data with strong wind gusts.

Table 6.7: Mean comparison metrics for wind gusts [0,15] and $(15; \infty)$ respectively.

Wind gust	[0,15]	$(15, \infty)$
Mean number of outages	55	64
Mean for objective function	5,835,556	4,897,778
Mean for adjusted objective function	1,973,611	1,873,380
Mean TPR	0.011	0.352
Mean for AUC	0.690	0.808

The second test investigates if there is a difference between substations which have experienced many and few outages, respectively. There were two substations which had experienced the same amount of outages and hence they were separated by chance. The result is summarised in Table 6.8. The substations with many outages have higher objective and adjusted objective function and TPR but a slightly lower AUC. It is important to notice that this data set includes 65 outages on average whereas the other data set with the few outages includes only 27 outages on average. The metrics should be higher for the larger data set.

Table 6.8: Mean comparison metrics for substations based on number of outages.

Substation, number of outages	Few	Many
Mean number of outages	27	65
Mean for objective function	2,328,333	5,266,111
Mean for adjusted objective function	953,472	1,800,579
Mean TPR	0.251	0.337
Mean for AUC	0.821	0.800

In Table 6.9, the result from the comparison between large and small amounts of overhead line in forest is summarised. The third test is conducted by including di erent parameters in the models on the same data set. The two di erent parameters show very similar results and hence neither parameter x_4, length of overhead line, in forest, and x_5, length of overhead line, not in forest can be said to be more important than the other.

Table 6.9: Mean comparison metrics for substations based on length of overhead line in forest.

Substation, length of overhead line	Forest	Not forest
Mean for objective function	8,097,222	8,023,889
Mean for adjusted objective function	2,976,713	2,983,681
Mean TPR	0.236	0.260
Mean for AUC	0.799	0.798

To compare the relative importance of the type of overhead lines, a fourth test is conducted by including di erent types of overhead lines parameters in the models. The result is summarised in Table 6.10 and the conclusion is similar to the previous test. None of the two parameters x_2, length of overhead line, isolated, and x_3, length of overhead line, not isolated, is more important than the other.

Table 6.10: Mean comparison metrics for substations based on type of overhead line.

Substation, Isolated or not isolated	Isolated	Not isolated
Mean for objective function	5,516,667	5,621,667
Mean for adjusted objective function	2,042,361	2,030,093
Mean TPR	0.304	0.287
Mean for AUC	0.796	0.800

Chapter 7

Discussion

In order to answer if an automatic prediction program can increase the prediction in comparison to current manual prediction methods, several automatic models are evaluated and compared to the VHI model. In addition, several analyses are made to investigate the relationship and impact between the considered variables. Theoretically, the outages registered as caused by wind in E.ON's database should indeed be caused by wind. However, this does not explain why and when it occurs, nor the impact of the grid structure and ground condition for weather related disturbances.

7.1 Impact of variables

The impact of wind gust is interesting since wind gust is the current basis for E.ON's resource allocation and is also the most commonly analysed variable in previous research. There seems to be no clear relationship between wind gusts and the spread of outages over time. Outages occur throughout the whole year and for wind gusts of different severity, as seen in Figure 6.1. A correlation between peaks of wind gust and outages may exist. In the data set, three peaks of wind gust occurred whereof the first two caused several outages.

The lack of outages in the latter could be because most of the dangerous trees had already fallen in the previous peaks. In Figure 6.2, the substations with large numbers of outages in 2015 did not have any outages at the peak in 2016, while the substations with outages in 2016 had very few in the previous peaks. This finding could indicate a time dependency between outages for different peaks of wind gust, but this was not possible to further analyse due to the few peaks in the current data set.

The result from Chapter 6.2, also indicates a relationship between peaks of wind gust and outages even though it is not significant. The probability of outages are highly based on the frequency of the peaks of wind gust, while it is approximately zero for wind gusts below 15 m/s.

By comparing the performance of the classi cation methods for two different data sets, wind below and above 15 m/s, respectively, we hope to con rm that stronger wind gust is a better predictor than lower wind gust. The result in Chapter 6.6 shows that the methods are able to better predict an outage in the higher interval, (15*; 1*) for all evaluation metrics. Therefore, using wind gusts below 15 m/s as a predictor is not as good as using wind gusts above in explaining why outages occur.

In addition, from a business perspective and in terms of compensation costs included, outages below a wind limit of 15 m/s are less important than for the higher interval. This is since the VHI is only concerned about the warnings from *Blast och Sn*. Nevertheless, had the models below the wind limit performed better, we would have recommended E.ON and other power distribution companies to consider wind gusts below as predictor.

Another interesting analysis is to understand the relationship between ground condition, power grid structure and outages. When di erent parameters are included in the models, the mean metrics show similar result as seen in Chapter 6.6. The x_3 and x_5 give a slightly better performance but the di erence is very small. It can be interpreted such as that the parameter x_3, length of overhead line, not isolated, is more sensitive to disturbances. The same applies for x_5, length of overhead line, not in forest. The result is somewhat contradicting reality since most of the disturbances are believed to be caused by trees falling on the lines and hence parameter x_4, length of overhead line, in forest, should provide better results. The parameter x_3 could indeed provide better results since the not isolated lines are in reality more sensitive than the isolated lines. Since the di erences are too small between the di erent parameters on a relatively small data set, a larger data set would be needed to con rm this observation.

In the case of the substation test, the data set with fewer outages provides su ciently better results for the AUC and the objective functions. In relation to the TPR, the data set with more outages is better. Due to the large di erence in number of outages, the results are probably skewed for the metrics. A larger number of outages should increase both the objective functions and the TPR as well as decrease the AUC. To be able to draw any conclusions, the skewness needs to be considered. By weighting the metrics based on for example the number of outages the skewness could be solved. Other weighting methods could also be used.

7.2 Model selection and comparable metrics

There are various methods to use when comparing models based on the same data set. The problem increases when models based on di erent data sets are compared, since the results often need to be transformed to be comparable. The adjusted objective function considers this by weighting the objective

function based on the di erent time periods, but both gave the same results. However, none of the other metrics were weighted nor did any consider the di erences in number of outages, which may a ect the results as mentioned previously. To interpret the values is therefore di cult as well as to conclude which model actually provides the best results and should be recommended. However, it can be concluded that negative binomial regression is to be preferred over the simpler Poisson regression. This is because the data set with the least number of zero counts is overdispersed and hence all other data sets should result in worse dispersion statistics.

When developing the models, we have strived for simplicity over complexity. This applies to both the statistical and machine learning methods. The assumption is that a simpler model is preferable to a more complex, as the simpler is easier to understand and interpret. The methods used show su cient results given their complexity but more advanced methods, such as the zero-in ated regression and black box algorithms, could be more precise. Those models would however be more di cult to understand and interpret, and hence also to implement.

The statistical models use AIC to nd the optimal model and the DT models are designed for the same reason with a maximum number of splits of 10. A higher splitting would most likely give more accurate results but the models would be harder to apply on other data sets. A lower splitting would cause higher errors but be easier to interpret. Judging from the results, the splitting rule of 10 seems su cient, since DT is more accurate than NB and LOG in terms of TPR but also has higher errors.

In order to select models, good model comparison metrics need to be used. In order to interpret the models from not only a perspective of minimising errors, the TPR is used as metric. There are several metrics which can be applied on the confusion matrices where TPR is only one of them. TPR is good since it optimises the number of TP, which is important for implementation. ROC curves and AUC values are good for understanding how the TPR changes in relation to the FPR. Furthermore, using both graphical and numerical metrics provide easier interpretation of the result.

7.3 Evaluation of models

Deciding which model performs the most satisfying result is di cult both because several comparison methods and di erent data sets are used. The three time periods have di erent number of outages which a ect the result and the metrics are not adjusted for this. The one-hour data set provides the best result for both the objective and adjusted objective function. However, both the six- and 24-hour data sets generate models with only $100{,}000$ higher cost than the three best models, which equals a few FP or one FN. This is still a relatively small cost for E.ON. A missed outage could a ect

the result between a better and worse performing model. The objective function is of course aimed at punishing FN since the cost associated with a missed outage can be quite large.

By instead looking at TPR for the models, the 24-hour data set provides the best results. The best models for objective and adjusted objective function has quite low TPR or even a zero rate. This is because the functions do not reward TP counts but instead only penalise errors. As the opposite, TPR only rewards instead of punishes. As a consequence, Model 11 which has the highest TPR of 80%, can still have 51 FP and a high objective function. Therefore, it is important to consider both metrics to find models that combine penalty and reward.

To achieve models that both reduce errors and predict more accurately, ensemble methods could be used where different parts of models are combined to achieve optimal results. In addition, the models with wind gust limit seem to predict more accurately than those with no wind gust limit. Therefore, a model with wind gust limit above 15 m/s seems reasonable.

Regarding the two statistical methods, they achieve very similar result, which is easily seen in the ROC curves in Chapter 6.5. The differentiator is the response variable, where LOG is better for binary and NB is better for count. The advantage of NB is that the response variable describes the magnitude of the outage and could be used for prioritising different regions. This was not tested since aggregation of the response variable was necessary for comparison reasons.

7.4 Comparison to the VHI model

One remark when analysing the different models with the reference point, i.e. the manual predicition by VHI, is that the models are more accurate than the VHI in terms of region. The models predict for each substation while the VHI predicts on a higher aggregated level. Thus, the errors of the models are a combination of time and station while the VHI can only falsely predict the time since all substations are within the same subcontractor area. For this reason, the VHI cost of 400,000 SEK from Chapter 6.4, should not be considered as an absolute value but as a reference. Some of the models might perform better in reality but have a higher objective function. Nevertheless, the models that have a lower objective function should in theory perform better than the VHI.

Another remark is that that the VHI has zero FP counts. The interpretation is that the VHI might be cautious of ordering extra workforce, which was a surprising result after discussions with employees at E.ON whose beliefs were the opposite. The zero point could be interpreted as that high values are not of consideration and by using the total AUC, the result could be misleading. A threshold for AUC could be used and in this thesis is

selected to be 0.1, meaning that one out of ten FP counts is considered reasonable. In discussions with E.ON, this threshold aligns with their expectations but other thresholds might provide better results.

Increasing the threshold, the FPR, will generate models with either the same or a higher TPR than previously, while decreasing the threshold will permit fewer errors but probably at a cost of lower TPR. This is because the ROC curve can only be an increasing monotone linear or concave function. In a cost perspective, higher values of FPR and TPR mean more false positive errors and less false negative errors which is less expensive than the opposite. Therefore, the threshold should rather be too big than too small. There could exist models, which have the same AUC value but have completely di erent ROC curves. A carefully chosen threshold would distinguish these types of models. In addition, using both AUC values and ROC curves is an advantage since the graphs can display these di erences and help in providing a reasonable threshold value.

Deciding the threshold value is relative to E.ON's objectives and will probably change over time and be adapted to the company's sensitivity of risk. For this reason, the models should be compared with both the AUC, AUC 0.1 and the cost function to nd satisfying results. By doing this, there seems to be two models that outperform the rest. Model 16 and 13 are the second and third best model for the objective functions and have the highest AUC and AUC 0.1 values. On the other hand, they have a zero TPR, but since the AUC considers this exact point it should not be any critical concern.

7.5 Business implementation

To implement and integrate the models in a business context, the models need to some extent decide the magnitude of each outage. This is since the magnitude is the basis for predicting future needs of resource allocation. The current data and models don't distinguish one outage from another.

It is important to remember that an outage in the models and an outage for the VHI might not have the same scale of impact. The models predict all weather related outages and not only those which had an impact longer than 12 hours. A modelled outage may in reality be an outage that the subcontractors can handle without the need of extra workforce. In addition, the extra workforce is only ordered during non-normal work hours, which the models do not take into account. In conclusion, the models cannot capture the real world perfectly.

Another important remark is that there can be simultaneous outages for the same substation during a certain time period. As the response variable for the negative binomial regression is countable, the number of parallel outages can be estimated. LOG and DT can also estimate the number of

outages, but only if the substations are aggregated to a larger region such as a subcontractor region. This is especially useful for estimating future needs of resource allocation, since a subcontractor is only contracted to handle a speci c number of parallel outages. To implement the models, the countable response variable needs to be used for the negative binomial regression or the result needs to be aggregated to a larger region. In this thesis, the countable response variable is converted to a binary variable to allow comparison with other models. Future research could focus on predicting number of outages and evaluate the performance of the countable variable of NB.

In addition, the one- and six-hour data sets might not be relevant for E.ON, since the VHI needs to make a decision about reinforcements 48 hours in advance. An automatic prediction decision tool should be based on the same forecast periods as the VHI to make the tool as e ective and easy to use as possible. Out of the three di erent time periods, the 24-hour data set is the most suitable for an implementation perspective.

In this thesis, both the models and the tests are developed to facilitate business implementation. No explanatory variables for substations are used as predictors and therefore, testing on other substation areas could be done as well as implementing the models for the total power grid. The ndings only apply to the ten selected substation areas. The proven relation between strong wind gusts and outages might not apply to other substation areas as well as the small di erence in relative importance for the parameters. It would be very interesting to analyse the conditions for northern Sweden since that area has other characteristics than the selected substations. This needs to be analysed and taken into account, since a VHI covers all of E.ON's substation areas and the same applies for a future decision support tool.

Chapter 8

Conclusion

This thesis shows that it is possible to predict weather related disturbances on the power grid with statistical and machine learning methods and to achieve better prediction than with the current manual prediction method. The statistical and machine learning methods are able to maintain the same false positive rate as the manual prediction, and at the same time increase the true positive rate from 20% to 33%. Statistical methods are generally better than decision tree. This could be explained by the machine learning methods requiring a larger data set to be able to see the underlying patterns.

We also conclude that wind gust is the parameter that has the highest impact on weather related outages and that the models are better to predict outages during stronger winds. Other parameters, such as type of power line, might a ect the prediction and make it more accurate, but tests show no signi cant di erence. Further data analysis is needed to con rm the impact of these variables.

The results in this thesis are a great basis for a decision support tool for E.ON, but the models need to be further adjusted to be implemented for the total power grid. Therefore, further data analysis is warranted before an automatic prediction of outages and resource allocation can be implemented in a real world context.

Chapter 9

Future work

This thesis is a novel approach to automating prediction of outages and resource allocation for power distributors. However, before implementation in a real world context, more research needs to be conducted.

First, the relative importance of the variables and the time dependency can be analysed by including more historical weather data and more parameters, such as precipitation, temperature and wind direction. Interpolating the weather data can also be of interest. Furthermore, the research can cover more substation areas. The ndings and conclusions drawn are based on a subset of ten substations in Sweden, and might not apply for other areas. Outages are predicted on a substation level, and aggregating to a larger region might increase accuracy as seen in previous research.

Second, other aspects of the methods can be tested such as the countable response variable for NB. E.ON could then with this variable predict the areas that are most vulnerable and more e ciently allocate resources. In addition, the metrics can incorporate the di erences in numbers of outages between the data sets. More complex models can be of interest, such as zero-in ated negative binomial regression and neural networks. Using a black box method, such as neural networks, might decrease transparency, but with access to more data, it might increase the accuracy of the prediction.

Third, the models can be extended to incorporate the outage magnitude and the number of parallel disturbances. In addition, the presented models can be visualised by a map with coloured substation areas, where the colour indicates the future risk of outage in that particular area. This type of decision support tool would only have a accuracy of 33%, but might help to reduce the variance in the manual prediction between di erent VHIs.

Finally, using ensemble methods can increase the performance. This is because our models perform better at di erent time periods and wind gust limits and an ensemble model may combine the best of each of them. To summarise, a smorgasbord of work for a interested reader to dive into exists.

Appendix A

Tables

In the following pages, the complete tables for the tests are presented. The structure of the tables are described in the associated chapters, 6.4-6.6.

Table A.1 presents the result for the model comparison metrics in Chapter 6.4 and the numerical values in Chapter 6.5.1 for the nal models. Only the three best models, the worst and the VHI model are presented in Chapter 6.4 and 6.5.1 and then discussed in Chapter 7.2 and 7.3.

Table A.1: Result for nal models.

Model number	Method	Time	Wind	Objective function	Adjusted objective function	TPR	AUC	AUC 0.1	TN	FP	FN	TP
1	DT	1h	0	2,050,000	85,417	0.050	0.525	0.010	65424	15	19	1
2	DT	6h	0	2,060,000	515,000	0.100	0.595	0.023	10880	26	18	2
3	DT	24h	0	2,280,000	2,280,000	0.200	0.588	0.020	2651	68	16	4
4	NB	1h	0	2,160,000	90,000	0.000	0.734	0.024	65423	16	20	0
5	NB	6h	0	2,320,000	580,000	0.050	0.726	0.028	10864	42	19	1
6	NB	24h	0	2,160,000	2,160,000	0.100	0.647	0.022	2683	36	18	2
7	LOG	1h	0	2,470,000	102,917	0.000	0.736	0.025	65392	47	20	0
8	LOG	6h	0	2,310,000	577,500	0.050	0.726	0.028	10865	41	19	1
9	LOG	24h	0	2,090,000	2,090,000	0.100	0.647	0.022	2690	29	18	2
10	DT	1h	15	350,000	14,583	0.333	0.658	0.034	723	15	2	1
11	DT	6h	15	630,000	157,500	0.500	0.833	0.036	200	43	2	2
12	DT	24h	15	610,000	610,000	0.800	0.706	0.011	91	51	1	4
13	NB	1h	15	420,000	17,500	0.000	0.873	0.047	726	12	3	0
14	NB	6h	15	480,000	120,000	0.250	0.809	0.036	225	18	3	1
15	NB	24h	15	490,000	490,000	0.400	0.658	0.026	123	19	3	2
16	LOG	1h	15	410,000	17,083	0.000	0.874	0.045	727	11	3	0
17	LOG	6h	15	550,000	137,500	0.000	0.819	0.031	228	15	4	0
18	LOG	24h	15	560,000	560,000	0.200	0.705	0.030	126	16	4	1
VHI	VHI	24h	15	400,000	400,000	0.200	0.600	0.024	142	0	4	1

Table A.2 presents the result for the test when the models are developed based on two di erent data sets of with wind gusts up to and above 15 m/s, respectively. In Chapter 6.6 the mean of each set, [0;15] and (15;∞), is presented and then discussed in Chapter 7.1.

Table A.2: All models used for the wind test.

Method	Time	Wind	Objective function	Adjusted objective function	TPR	AUC	AUC 0.1	TN	FP	FN	TP
DT	1h	[0;15]	6,930,000	288,750	0.014	0.596	0.020	149513	3	69	1
DT	6h	[0;15]	5,560,000	1,390,000	0.036	0.589	0.020	24389	16	54	2
DT	24h	[0;15]	3,820,000	3,820,000	0.051	0.717	0.024	5777	12	37	2
NB	1h	[0;15]	7,600,000	316,667	0.000	0.757	0.023	149456	60	70	0
NB	6h	[0;15]	6,150,000	1,537,500	0.000	0.732	0.024	24350	55	56	0
NB	24h	[0;15]	4,290,000	4,290,000	0.000	0.667	0.019	5750	39	39	0
LOG	1h	[0;15]	7,730,000	322,083	0.000	0.757	0.023	149443	73	70	0
LOG	6h	[0;15]	6,190,000	1,547,500	0.000	0.732	0.024	24346	59	56	0
LOG	24h	[0;15]	4,250,000	4,250,000	0.000	0.667	0.019	5754	35	39	0
DT	1h	(15;∞)	4,590,000	191,250	0.453	0.766	0.045	3231	109	35	29
DT	6h	(15;∞)	3,930,000	982,500	0.556	0.839	0.039	877	113	28	35
DT	24h	(15;∞)	2,220,000	2,220,000	0.906	0.877	0.039	335	162	6	58
NB	1h	(15;∞)	6,240,000	260,000	0.109	0.809	0.038	3286	54	57	7
NB	6h	(15;∞)	5,450,000	1,362,500	0.222	0.796	0.034	935	55	49	14
NB	24h	(15;∞)	5,130,000	5,130,000	0.281	0.795	0.033	444	53	46	18
LOG	1h	(15;∞)	5,920,000	246,667	0.156	0.809	0.039	3288	52	54	10
LOG	6h	(15;∞)	5,510,000	1,377,500	0.206	0.792	0.033	939	51	50	13
LOG	24h	(15;∞)	5,090,000	5,090,000	0.281	0.795	0.033	448	49	46	18

Table A.3 presents the result for the test when the models are developed based on two di erent data sets of substations with the highest and the lowest numbers of outages, respectively. In Chapter 6.6 the mean of each set, OUTAGE and NOT_OUTAGE, is presented and then discussed in Chapter 7.1.

Table A.3: All models used for the substation test.

Method	Time	Wind	Set	Objective function	Adjusted objective function	TPR	AUC	AUC 0.1	TN	FP	FN	TP
DT	1h	0	OUTAGE	7,910,000	329,583	0.211	0.712	0.045	76344	41	75	20
DT	6h	0	OUTAGE	6,470,000	1,617,500	0.494	0.761	0.051	12451	227	42	41
DT	24h	0	OUTAGE	3,320,000	3,320,000	0.657	0.869	0.061	3032	92	24	46
NB	1h	0	OUTAGE	10,030,000	417,917	0.042	0.871	0.054	76293	93	91	4
NB	6h	0	OUTAGE	8,090,000	2,022,500	0.120	0.868	0.056	12599	79	73	10
NB	24h	0	OUTAGE	5,850,000	5,850,000	0.243	0.868	0.058	3069	55	53	17
LOG	1h	0	OUTAGE	9,710,000	404,583	0.074	0.871	0.054	76293	91	88	7
LOG	6h	0	OUTAGE	8,030,000	2,007,500	0.120	0.869	0.056	12602	73	73	10
LOG	24h	0	OUTAGE	6,090,000	6,090,000	0.214	0.868	0.058	3067	59	55	15
DT	1h	15	OUTAGE	3,060,000	127,500	0.426	0.784	0.041	1268	36	27	20
DT	6h	15	OUTAGE	2,320,000	580,000	0.957	0.790	0.033	149	212	2	44
DT	24h	15	OUTAGE	790,000	790,000	1.000	0.870	0.040	91	79	0	45
NB	1h	15	OUTAGE	4,370,000	182,083	0.149	0.756	0.035	1265	37	40	7
NB	6h	15	OUTAGE	3,700,000	925,000	0.283	0.678	0.026	323	40	33	13
NB	24h	15	OUTAGE	3,400,000	3,400,000	0.311	0.768	0.031	145	30	31	14
LOG	1h	15	OUTAGE	4,530,000	188,750	0.146	0.756	0.035	1258	43	41	7
LOG	6h	15	OUTAGE	3,950,000	987,500	0.234	0.678	0.026	329	35	36	11
LOG	24h	15	OUTAGE	3,170,000	3,170,000	0.383	0.768	0.031	144	27	29	18
DT	1h	0	NOT_OUTAGE	3,280,000	136,667	0.256	0.628	0.029	76386	38	29	10
DT	6h	0	NOT_OUTAGE	2,720,000	680,000	0.333	0.733	0.048	12685	32	24	12
DT	24h	0	NOT_OUTAGE	2,590,000	2,590,000	0.242	0.766	0.047	3153	9	25	8
NB	1h	0	NOT_OUTAGE	4,040,000	168,333	0.000	0.803	0.050	76409	14	39	0
NB	6h	0	NOT_OUTAGE	3,650,000	912,500	0.083	0.819	0.047	12684	35	33	3
NB	24h	0	NOT_OUTAGE	3,200,000	3,200,000	0.121	0.798	0.045	3130	30	29	4
LOG	1h	0	NOT_OUTAGE	4,170,000	173,750	0.026	0.820	0.050	76386	37	38	1
LOG	6h	0	NOT_OUTAGE	3,460,000	865,000	0.139	0.819	0.047	12678	36	31	5
LOG	24h	0	NOT_OUTAGE	2,990,000	2,990,000	0.182	0.800	0.044	3134	29	27	6
DT	1h	15	NOT_OUTAGE	830,000	34,583	0.813	0.949	0.076	1864	53	3	13
DT	6h	15	NOT_OUTAGE	790,000	197,500	0.647	0.933	0.061	579	19	6	11
DT	24h	15	NOT_OUTAGE	1,070,000	1,070,000	0.474	0.876	0.047	305	7	10	9
NB	1h	15	NOT_OUTAGE	1,510,000	62,917	0.125	0.868	0.055	1906	11	14	2
NB	6h	15	NOT_OUTAGE	1,440,000	360,000	0.235	0.854	0.047	584	14	13	4
NB	24h	15	NOT_OUTAGE	1,570,000	1,570,000	0.263	0.798	0.034	294	17	14	5
LOG	1h	15	NOT_OUTAGE	1,530,000	63,750	0.125	0.869	0.055	1898	13	14	2
LOG	6h	15	NOT_OUTAGE	1,310,000	327,500	0.294	0.854	0.047	588	11	12	5
LOG	24h	15	NOT_OUTAGE	1,760,000	1,760,000	0.158	0.799	0.034	295	16	16	3

Table A.4 presents the result for the test when the forest and not forest parameter was included in each model. The mean of each set, FOREST and NOT_FOREST, is presented and discussed in Chapter 6.6 and 7.1.

Table A.4: All models used for the forest test.

Method	Time	Wind	Set	Objective function	Adjusted objective function	TPR	AUC	AUC 0.1	TN	FP	FN	TP
DT	1h	0	FOREST	11,660,000	485,833	0.194	0.713	0.044	152723	86	108	26
DT	6h	0	FOREST	9,820,000	2,455,000	0.210	0.735	0.048	25353	42	94	25
DT	24h	0	FOREST	6,250,000	6,250,000	0.612	0.825	0.054	6061	225	40	63
NB	1h	0	FOREST	14,460,000	602,500	0.022	0.858	0.053	152673	136	131	3
NB	6h	0	FOREST	11,750,000	2,937,500	0.101	0.852	0.053	25290	105	107	12
NB	24h	0	FOREST	9,750,000	9,750,000	0.136	0.837	0.052	6201	85	89	14
LOG	1h	0	FOREST	14,140,000	589,167	0.052	0.858	0.053	152665	144	127	7
LOG	6h	0	FOREST	11,990,000	2,997,500	0.084	0.852	0.053	25286	109	109	10
LOG	24h	0	FOREST	9,480,000	9,480,000	0.165	0.837	0.052	6198	88	86	17
DT	1h	15	FOREST	4,570,000	190,417	0.397	0.689	0.037	3144	77	38	25
DT	6h	15	FOREST	4,220,000	1,055,000	0.397	0.806	0.036	917	42	38	25
DT	24h	15	FOREST	2,230,000	2,230,000	0.984	0.841	0.033	269	213	1	63
NB	1h	15	FOREST	6,330,000	263,750	0.079	0.797	0.038	3168	53	58	5
NB	6h	15	FOREST	5,710,000	1,427,500	0.175	0.764	0.031	908	51	52	11
NB	24h	15	FOREST	5,720,000	5,720,000	0.188	0.782	0.028	430	52	52	12
LOG	1h	15	FOREST	6,520,000	271,667	0.078	0.797	0.038	3155	62	59	5
LOG	6h	15	FOREST	5,700,000	1,425,000	0.148	0.764	0.031	909	50	52	9
LOG	24h	15	FOREST	5,450,000	5,450,000	0.219	0.782	0.028	436	45	50	14
DT	1h	0	NOT_FOREST	11,760,000	490,000	0.187	0.593	0.023	152723	86	109	25
DT	6h	0	NOT_FOREST	9,570,000	2,392,500	0.269	0.735	0.047	25308	87	87	32
DT	24h	0	NOT_FOREST	6,570,000	6,570,000	0.602	0.785	0.052	6039	247	41	62
NB	1h	0	NOT_FOREST	13,700,000	570,833	0.000	0.859	0.054	152779	30	134	0
NB	6h	0	NOT_FOREST	12,240,000	3,060,000	0.067	0.856	0.053	25281	114	111	8
NB	24h	0	NOT_FOREST	9,680,000	9,680,000	0.146	0.842	0.052	6198	88	88	15
LOG	1h	0	NOT_FOREST	14,010,000	583,750	0.045	0.859	0.054	152688	121	128	6
LOG	6h	0	NOT_FOREST	11,880,000	2,970,000	0.092	0.856	0.053	25287	108	108	11
LOG	24h	0	NOT_FOREST	9,860,000	9,860,000	0.126	0.842	0.052	6200	86	90	13
DT	1h	15	NOT_FOREST	4,520,000	188,333	0.460	0.771	0.045	3109	112	34	29
DT	6h	15	NOT_FOREST	3,860,000	965,000	0.556	0.810	0.037	853	106	28	35
DT	24h	15	NOT_FOREST	2,620,000	2,620,000	1.000	0.842	0.037	220	262	0	64
NB	1h	15	NOT_FOREST	6,030,000	251,250	0.127	0.803	0.038	3168	53	55	8
NB	6h	15	NOT_FOREST	5,720,000	1,430,000	0.190	0.773	0.031	897	62	51	12
NB	24h	15	NOT_FOREST	5,210,000	5,210,000	0.266	0.777	0.029	431	51	47	17
LOG	1h	15	NOT_FOREST	6,410,000	267,083	0.094	0.804	0.039	3156	61	58	6
LOG	6h	15	NOT_FOREST	5,590,000	1,397,500	0.190	0.773	0.031	912	49	51	12
LOG	24h	15	NOT_FOREST	5,200,000	5,200,000	0.266	0.777	0.029	432	50	47	17

Table A.5 presents the result for the test when the isolated and not isolated parameter was included in each model. The mean of each set, ISOLATED and NOT_ISOLATED, is presented and discussed in Chapter 6.6 and 7.1.

Table A.5: All models used for the type of line test.

Method	Time	Wind	Set	Objective function	Adjusted objective function	TPR	AUC	AUC 0.1	TN	FP	FN	TP
DT	1h	0	ISOLATED	7,800,000	325,000	0.237	0.593	0.023	152750	40	74	23
DT	6h	0	ISOLATED	6,430,000	1,607,500	0.326	0.736	0.048	25331	43	60	29
DT	24h	0	ISOLATED	4,160,000	4,160,000	0.697	0.786	0.051	6067	146	27	62
NB	1h	0	ISOLATED	9,150,000	381,250	0.000	0.861	0.054	152772	25	89	0
NB	6h	0	ISOLATED	7,770,000	1,942,500	0.134	0.855	0.053	25285	67	71	11
NB	24h	0	ISOLATED	6,540,000	6,540,000	0.211	0.840	0.052	6199	54	60	16
LOG	1h	0	ISOLATED	9,250,000	385,417	0.087	0.860	0.054	152684	85	84	8
LOG	6h	0	ISOLATED	7,740,000	1,935,000	0.157	0.855	0.053	25284	74	70	13
LOG	24h	0	ISOLATED	6,210,000	6,210,000	0.243	0.840	0.052	6194	61	56	18
DT	1h	15	ISOLATED	3,000,000	125,000	0.549	0.756	0.044	3116	70	23	28
DT	6h	15	ISOLATED	2,730,000	682,500	0.558	0.810	0.038	895	43	23	29
DT	24h	15	ISOLATED	1,530,000	1,530,000	1.000	0.834	0.032	253	153	0	64
NB	1h	15	ISOLATED	4,340,000	180,833	0.114	0.800	0.038	3156	44	39	5
NB	6h	15	ISOLATED	3,740,000	935,000	0.277	0.767	0.032	906	34	34	13
NB	24h	15	ISOLATED	3,340,000	3,340,000	0.362	0.776	0.028	430	34	30	17
LOG	1h	15	ISOLATED	6,180,000	257,500	0.109	0.8011	0.0381	3172	48	57	7
LOG	6h	15	ISOLATED	4,220,000	1,055,000	0.178	0.7764	0.0320	906	52	37	8
LOG	24h	15	ISOLATED	5,170,000	5,170,000	0.238	0.7756	0.0288	440	37	48	15
DT	1h	0	NOT_ISOLATED	7,880,000	328,333	0.255	0.593	0.023	152723	58	73	25
DT	6h	0	NOT_ISOLATED	6,560,000	1,640,000	0.284	0.735	0.047	25356	26	63	25
DT	24h	0	NOT_ISOLATED	4,410,000	4,410,000	0.593	0.784	0.053	6150	91	35	51
NB	1h	0	NOT_ISOLATED	9,170,000	382,083	0.000	0.860	0.054	152773	27	89	0
NB	6h	0	NOT_ISOLATED	8,050,000	2,012,500	0.110	0.856	0.054	25282	75	73	9
NB	24h	0	NOT_ISOLATED	6,520,000	6,520,000	0.181	0.843	0.052	6201	62	59	13
LOG	1h	0	NOT_ISOLATED	9,710,000	404,583	0.043	0.859	0.054	152679	91	88	4
LOG	6h	0	NOT_ISOLATED	7,980,000	1,995,000	0.133	0.856	0.054	25277	78	72	11
LOG	24h	0	NOT_ISOLATED	6,820,000	6,820,000	0.162	0.843	0.052	6199	62	62	12
DT	1h	15	NOT_ISOLATED	3,050,000	127,083	0.558	0.771	0.045	3109	75	23	29
DT	6h	15	NOT_ISOLATED	2,710,000	677,500	0.549	0.820	0.037	898	41	23	28
DT	24h	15	NOT_ISOLATED	1,610,000	1,610,000	0.919	0.835	0.026	316	111	5	57
NB	1h	15	NOT_ISOLATED	3,950,000	164,583	0.182	0.807	0.039	3166	35	36	8
NB	6h	15	NOT_ISOLATED	3,930,000	982,500	0.217	0.780	0.031	910	33	36	10
NB	24h	15	NOT_ISOLATED	3,280,000	3,280,000	0.388	0.785	0.030	437	28	30	19
LOG	1h	15	NOT_ISOLATED	6,300,000	262,500	0.095	0.8078	0.0391	3163	60	57	6
LOG	6h	15	NOT_ISOLATED	5,780,000	1,445,000	0.169	0.7800	0.0309	906	38	54	11
LOG	24h	15	NOT_ISOLATED	3,480,000	3,480,000	0.333	0.785	0.031	436	28	32	16

Bibliography

[1] G. Claeskens and N. L. Hjort, *Model selection and model averaging*, vol. 330. Cambridge University Press Cambridge, 2008.

[2] T. Fawcett, \An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861{874, 2006.

[3] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classi cation and regression trees*. CRC Press, 1984.

[4] J. M. Hilbe, *Negative binomial regression*. Cambridge University Press, 2011.

[5] SMHI, \Sveriges meteorologiska och hydrologiska institut." `https://www.smhi.se/`, 2017. Accessed: May 4, 2017.

[6] SVK, \Svenska kraftnat." `https://www.svk.se/`, 2017. Accessed: May 4, 2017.

[7] Energimyndigheten, \Stormen Gudrun { Vad kan vi lara av naturkatastrofen 2005," 2005.

[8] Energimyndigheten, \Funktionskrav inom elforsorjningen," 2009.

[9] Gartner, \Gartner's top 10 strategic technology trends for 2017," 2016.

[10] Gartner, \Magic quadrant for advanced analytics platforms," *Gartner Report G*, vol. 270612, 2015.

[11] ABB Communications, \Leading utility adopts Ventyx software from ABB for smart grid control center in Sweden," 2013.

[12] Y. Zhou, A. Pahwa, and S.-S. Yang, \Modeling weather-related failures of overhead distribution lines," *IEEE Transactions on Power Systems*, vol. 21, no. 4, pp. 1683{1690, 2006.

[13] R. Billinton, C. Wu, and G. Singh, \Extreme adverse weather modeling in transmission and distribution system reliability evaluation," in *Power Systems Computation Conf.(PSCC)-2002, Spain*, 2002.

[14] K. Alvehag and L. Soder, \A stochastic weather dependent reliability model for distribution systems," in *Probabilistic Methods Applied to Power Systems, 2008. PMAPS'08. Proceedings of the 10th International Conference on*, pp. 1{8, IEEE, 2008.

[15] 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, no. 4, pp. 1170{1175, 2002.

[16] H. Liu, R. A. Davidson, D. V. Rosowsky, and J. R. Stedinger, \Negative binomial regression of electric power outages in hurricanes," *Journal of infrastructure systems*, vol. 11, no. 4, pp. 258{267, 2005.

[17] S.-R. Han, S. D. Guikema, S. M. Quiring, K.-H. Lee, D. Rosowsky, and R. A. Davidson, \Estimating the spatial distribution of power outages during hurricanes in the Gulf Coast region," *Reliability Engineering & System Safety*, vol. 94, no. 2, pp. 199{210, 2009.

[18] E.ON, \Primarapparater, kraftsystemuppbyggnad," 2017.

[19] Energimyndigheten, \Elnat i fysisk planering," 2014.

[20] SMHI, \Produktblad: Varningstjanst for kraftig blast och sno," 2013.

[21] The R Foundation, \What is R?." `https://www.r-project.org/about.html`, 2017. Accessed: May 16, 2017.

[22] Salford Systems, \Gini criterion." `https://www.salford-systems.com/images/site/gini.gif`, 2017. Accessed: May 4, 2017.

[23] Salford Systems, \Twoing splitting rule." `https://www.salford-systems.com/images/site/gini4.gif`, 2017. Accessed: May 4, 2017.

[24] R. Timofeev, *Classi cation and regression trees (CART) theory and applications*. PhD thesis, Humboldt University, Berlin, 2004.

[25] Salford Systems, \Do splitting rules really matter?," 2017.

[26] P. A. Flach and S. Wu, \Repairing Concavities in ROC Curves.," in *IJCAI*, pp. 702{707, Citeseer, 2005.

[27] I. Var, \Multivariate data analysis," *vectors*, vol. 8, no. 2, pp. 125{136, 1998.

[28] P.-N. Tan, M. Steinbach, and V. Kumar, \Introduction to Data Mining," 2006.

[29] SMHI, \SMHI Open Data API Documentation." http://opendata.smhi.se/apidocs/, 2017. Accessed: May 4, 2017.

[30] SMHI, \Hur mats vind?." https://www.smhi.se/kunskapsbanken/meteorologi/hur-mats-vind-1.5924, 2017. Accessed: May 4, 2017.

[31] Lantmateriet, \Abous us - Land Survey." http://www.lantmateriet.se/en/About-Lantmateriet/About-us/, 2017. Accessed: May 16, 2017.

[32] E.ON, \Hur stor ar avbrottsersattningen?," 2017.