Wires-Down Predictive Modeling and Preventative Measures Optimization
by
Gregory D. Eschelbach

Submitted to the MIT Sloan School of Management and the Department of Engineering Systems Design on May 6, 2016 in partial Fulfillment of the Requirements for the Degrees of Master of Business Administration and Master of Science in Engineering Systems Design.

Abstract

In 2012, the Pacific Gas and Electric Company (PG&E) identified overhead wires-down failure events as an important metric for safety and reliability. These events occur when power lines contact the ground and are caused by a host of reasons including trees falling on wires, animals creating problems, car accidents into poles, or older assets deteriorating and separating. To mitigate these wires-down events, PG&E has a portfolio of preventative measures including tree trimming, routine inspections, and replacement of older assets. This thesis hypothesized that a statistical approach could be used to identify the key drivers in wires-down events and predict future failure locations, thereby helping to direct the preventative measures.

In the data aggregation phase of the project, ten datasets describing potential causal factors in wires-down events were integrated into a single database. Key contributions included utilizing clustering analysis to create standardized weather “environments” based on numerous weather factors, and developing an algorithm that used millions of individual tree trimming records to describe the vegetation environment surrounding each line segment.

Statistical regressions and machine learning algorithms were then applied to the dataset to model the wires-down events. Out-of-sampling testing determined the best approach, and the predicted timeframe was iteratively rotated to show robustness throughout time. The best results were achieved using logistic regression models. These models predicted hot spot locations of future failures for specific failure modes and provided insight into the key factors that were causing the failures.

To implement the models’ findings, the discrete line segment failure probabilities were aggregated into sections of line that were several miles in length to calculate cumulative failure ratings. These groupings were the appropriate length for preventative measure projects at PG&E, and allowed PG&E to prioritize their tree trimming and re-conductoring efforts based on the future failure risk. Ultimately, this project simplified decision-making for PG&E experts by focusing their efforts onto a more manageable and perceivable portion of their complex network.

Thesis Supervisor: Georgia Perakis
Title: William F. Pounds Professor of Management Science
MIT Sloan School of Management

Thesis Supervisor: James Kirtley
Title: Professor of Electrical Engineering
Department of Electrical Engineering and Computer Science
Acknowledgements

My experience at PG&E was exceptional, and it was largely due to the support from the PG&E team. The PG&E management from my direct manager up to the senior vice president gave me the freedom and resources to explore this project. The expert advice from the Asset Management team, the Vegetation Management team, the GIS team, and countless others provided timely guidance, and our daily interactions made this project enjoyable.

Additionally, I would like to thank my advisors, Professor Georgia Perakis and Professor James Kirtley, for their time, expertise, and support on this project. I owe much gratitude to Erick Corona (MIT LGO 2013) who was the catalyst for this thesis, and who provided daily guidance and motivation from the start that led to this successful project.

Finally, I need to thank my mom and dad, Amy, Danielle, and my friends for all of their support and love during my time at MIT. These last two years have been a special time in my life, and I owe a lot to all of you who made it possible.
# Table of Contents

1 **Introduction to the Wires-Down Project at Pacific Gas & Electric** .......................................................... 11  
1.1 Thesis Motivation .......................................................................................................................... 11  
1.2 PG&E Network Statistics ............................................................................................................. 11  
1.3 Wires-Down Statistics .................................................................................................................. 13  
1.4 Literature Review .......................................................................................................................... 15  
1.5 Thesis Hypothesis .......................................................................................................................... 16  
1.6 Thesis Outline & Contributions .................................................................................................... 16  
1.7 Important Terms ............................................................................................................................ 17  

2 **Data** ............................................................................................................................................... 18  
2.1 Summary of Data Sources ........................................................................................................... 18  
2.2 Additional Unused Data Sources ................................................................................................. 20  
2.3 Model Granularity ......................................................................................................................... 21  
2.4 Weather Clustering Methodology ................................................................................................. 21  

3 **Model #1: General Wires-Down Prediction Model** ...................................................................... 24  
3.1 Model Methodology .................................................................................................................... 24  
3.2 Logistic Regression Formulation .................................................................................................. 24  
3.3 Out of Sample Testing ................................................................................................................ 25  
3.4 Splice Significance ....................................................................................................................... 26  
  3.4.1 Splice Statistical Significance T-Test ...................................................................................... 26  
  3.4.2 Splice Categories .................................................................................................................... 27  
3.5 Ranking Variable Significance ..................................................................................................... 27  
3.6 General Model Results ................................................................................................................ 28  
3.7 Plot of Predicted Results .............................................................................................................. 29  
3.8 Limitations of the General Model ............................................................................................... 30  
3.9 General Model Conclusions ......................................................................................................... 31  

4 **Model #2: Vegetation-Specific Wires-Down Prediction Model** .................................................. 32  
4.1 Model Formulation ....................................................................................................................... 32  
4.2 Ranking Variable Significance .................................................................................................... 33  
4.3 Vegetation Model Results ............................................................................................................ 33  

5 **Implementation of the Vegetation Model Results** ...................................................................... 35  
5.1 Aggregating Predictions .............................................................................................................. 35  
5.2 Identifying Critical Locations ..................................................................................................... 36  
5.3 Implementation into PG&E Operations ....................................................................................... 38
List of Figures

Figure 1: Highlighted PG&E Service Territory................................................................. 12
Figure 2: PG&E Distribution Lines.................................................................................. 12
Figure 3: Wires-Down Events since 2012.................................................................. 14
Figure 4: Principal Component Analysis of Weather Variables ............................... 22
Figure 5: Number of Weather Cluster Optimization.................................................... 23
Figure 6: General Prediction Model ROC Curve ......................................................... 28
Figure 7: Map of Predicted Failure Locations from General Model ......................... 30
Figure 8: ROC Curve for Vegetation-Related Wires-Down Prediction....................... 34
Figure 9: Critical Regions for Vegetation Wires-Down Failure.................................. 37
Figure 10: ROC Curve for Equipment-Related Wires-Down Prediction...................... 40
Figure 11: Critical Regions for Equipment-Related Wires-Down Failure.................... 43
Figure 12: Ranked Variable Significance from the General Model............................. 47
Figure 13: Ranked Variable Significance from the Vegetation-Specific Model .......... 48
Figure 14: Ranked Variable Significance from the Equipment-Specific Model......... 49
Figure 15: General Model ROC Curve using a Decision Tree..................................... 50

List of Tables

Table 1: Vegetation-Related Wires-Down Criticality Rankings..................................... 36
Table 2: Equipment-Related Wires-Down Criticality Rankings..................................... 42
1 Introduction to the Wires-Down Project at Pacific Gas & Electric

1.1 Thesis Motivation

The definition of a wires-down event is when an electrical power line (often termed “conductor” in the industry) comes in contact with the ground or a ground-based object. These downed power lines pose both a service reliability problem and a safety problem for PG&E due to the loss of electricity to a region and the potential for electrical shock or fire ignition.

In 2012, PG&E began a concerted effort to track, minimize, and ultimately prevent these wires-down events. To accomplish this, PG&E utilizes a portfolio of preventative measures that address various wires-down causal factors including vegetation and aging assets. The vegetation management program does preemptive tree trimming to mitigate the vegetation risk, while the re-conductoring program installs new power lines to reduce the impact of aging assets. Integrated in this approach are also a series of inspections, some mandated by regulation and others employed by PG&E for further risk mitigation, that periodically visually check the power lines to find potential issues. Limited by operating budgets and time, these programs must make decisions on where to apply these preventative measures in order to be most efficient and reduce the most wires-down risk. Historically, these decisions have been made using the expertise of the employees, relying heavily on their experience and the historical failure records.

Given the large territory that PG&E covers, the increasing age of its assets, and the ever-present budgetary constraints, this project sought to aid the human decision making process in determining where to implement these preventative measures for maximum effectiveness.

1.2 PG&E Network Statistics

The Pacific Gas and Electric Company (PG&E), headquartered in San Francisco, is an electric and gas utility provider for a significant portion of California, and the largest electrical utility in the United States based on the number of customers served. Their service territory, shown in Figure 1, spans most of central and northern California, totals over 70,000 square miles, and contains a diverse landscape of coastal, heavily forested, mountainous, and urban regions.
To provide electricity for the region, PG&E has over 113,000 miles of power lines that are held up by over two million power line poles and connected to over 365,000 devices (switches, fuses, etc.). The vast majority of this equipment occurs above ground and forms their electrical distribution network, which connects the high-voltage transmission lines to the end customers. This thesis focused only on this overhead distribution network since those lines comprise such a large percentage of the system and are the most susceptible to wires-down failures.
1.3 Wires-Down Statistics

The wires-down dataset that has accumulated since the start of PG&E’s more detailed tracking program in 2012 determined the timeframe of this project. Specifically, the dataset ranged from January 2012 to March 2015. Over those 3.5 years, there has been ~10,000 wires-down events throughout PG&E’s electrical network. Of that total, approximately 6,000 of those events occurred outside of major storms, resulting in a mean failure rate of a handful of wires-down events on a normal day. During major storms, this failure rate is historically much higher in the localized region of the storm.

The wires-down dataset contained detailed descriptions of each event and categorized the failures into one of the governing failure modes delineated below. Figure 3 maps the locations of the wires-downs since 2012 and their primary cause. For this project, the vegetation and equipment failure modes were the only modes targeted because they represent the vast majority of failures and are the most easily addressed with the preventative measures.

Wires-Down Failure Modes

- Animal – Failures caused by an animal; typically due to an animal bridging the terminals on a device, causing a short, or over-weighting a line segment
- Company Initiated – Infrequent events caused by PG&E personnel
- Equipment – Failures related to splices separating or a line segment disconnecting from a device
- 3rd Party – Failures caused by non-PG&E personnel (other companies, car accidents, etc.)
- Vegetation – Failures due to vegetation falling into the power lines
As a first-order analysis of Figure 3, the failure patterns of the vegetation and equipment-related wires-down appear different. The vegetation failures, highlighted in green, are more concentrated, occur in only certain regions of the network, and appear biased to the coast and the eastern side of the territory. In comparison, the blue equipment failures are more uniformly spread throughout the network (often obscured by the vegetation failures). The reasons for these different failure patterns will be examined further throughout this thesis, and are correlated to the underlying vegetation environments, weather patterns, and network density surrounding each line segment.

Figure 3: Wires-Down Events since 2012
1.4 Literature Review

The growing popularity of data analytics has produced a breadth of research in the past decade related to data manipulation, visualization, and model-based prediction. The utility industry is a prime target for the application of these analytical techniques because their systems are often large and complicated for humans to fully perceive. In addition, their company budgets are usually fixed, increasing the need to make optimized data-supported decisions. Finally, the public has an increased awareness of outages and the aging utility system, making reliability and responsiveness more important. For this project, the literature review focused on two areas: 1) the prediction of outages, to understand the past successes in modeling utility failures, and 2) methods to account for weather variability in determining overall system reliability.

The first segment of literature on outage prediction detailed a host of studies that have applied analytical modeling techniques in attempts to predict power line failures. Recent work in [1] and [2] demonstrated considerable accuracy in predicting the location of storm outages due to hurricanes by utilizing a maximum likelihood estimator technique and classification trees. These works also highlighted the beneficial effects of aggregating individual predictions into geographic regions to boost the prediction accuracy.

The application of predictive models to utilities was first discussed in [3] and [4] when Poisson regression models were applied to outages to understand the most significant variables. Later, the work of [5] and [6] advanced these statistical approaches by using a maximum likelihood model or estimator technique and a Poisson regression model to predict future weather failures.

In recent years, other modeling techniques besides regression models have expanded the literature on failure prediction. Generalized linear models [7] and spatial generalized linear mixed models [8] were utilized to predict hurricane-related outages. These works also discussed the techniques of principal component analysis (PCA) to minimize the over-dispersion of the dataset and suggested important data sources including the National Land Cover Database, the STATSGO Soil Database, and the Standard Precipitation Index for estimating drought effects. [8] also aggregated the prediction results in small geographic regions [3 square km] to improve accuracy. Finally, Artificial Immune Recognition System (AIRS) models were attempted in [9] and [10] as an alternative method to identify failure causes. The AIRS approach was shown to be an effective way of combating the common data imbalance issue with utilities where only a small number of infrequent failures exist relative to the vast network scale.

For the majority of the failure prediction literature, the models were focused on a short time-horizon and were created to run regularly as weather systems moved through an area in order to anticipate outage locations and their magnitude. These applications are particularly useful for utilities to locate and scale their emergency response in real-time, but they differ from the intended application of this thesis. The intention of the PG&E application was to run semi-annually during the yearly planning cycles to locate potential hot spots.
and plan preventative maintenance projects to address them. Therefore, this thesis adds to the existing applications in the literature by utilizing prediction models to evaluate longer-term system risks and improve maintenance planning.

The second portion of the literature review focused on measuring system reliability specifically due to weather variability. In works [11], [12], and [13], the weather is categorized into three severity states (normal, adverse, and extreme) to account for the impact of weather on the system's reliability. These categories allow the model to account for weather effects without introducing numerous weather variables. In [14], several datasets including the “Daily Surface Summary Data” from the U.S. National Oceanic and Atmospheric Administration (NOAA) and the Palmer Drought Severity Index (PDSI) are suggested. Additionally, [14] introduces the technique of normalizing the yearly system reliability targets based upon the year’s weather events. The method of year-end reliability metric normalization was expanded in [15] in which the number of storms combined with the average storm impact was used to modify the baseline reliability target. This thesis expands on the system reliability literature by quantifying the impact of weather on wires-down failure probabilities, helping to justify the use of normalizing techniques to adjust yearly metrics based on the weather.

1.5 Thesis Hypothesis

Based on the literature findings and guidance from PG&E, the hypothesis of this thesis was that certain factors surrounding each power line increase the likelihood of a wires-down event. By understanding these significant factors and then investigating where else they occur within the PG&E system, the probability of future wires-down events can be calculated across the network to highlight susceptible locations. The preventative measures can then be optimally applied in a directed manner towards these critical spots to best mitigate the risk of future wires-down failures.

1.6 Thesis Outline & Contributions

This thesis generated three significant contributions that expand on the previous academic research and directly benefit PG&E and their preventative maintenance operations. These contributions are:

1. Created a comprehensive dataset that synthesized many unique data streams describing the PG&E electrical network that was utilized for this project, and applicable to many additional analytics projects in the future.
2. Expanded the use of machine learning into predicting wires-down failure events, and demonstrated a high degree of predictive capability.
3. Generated a ranking of variable significance in wires-down failures that increased the industry understanding about which factors increase the likelihood of these events.
All of these contributions enhance PG&E’s ability to proactively address potential wires-down hot spot locations and aid in prioritizing their preventative maintenance projects. The potential cost savings from this project are substantial given the severity of every wires-down event.

This thesis begins by presenting the available data sources and the data manipulation methods utilized to create a cohesive, mineable dataset. This section contains multiple contributions in terms of the unique approaches taken to implement the data that PG&E had available. First, k-means clustering was applied to generate weather environments from numerous weather variables, minimizing the consumption of the models’ degrees of freedom by the weather. Additionally, a vegetation algorithm is detailed that utilized millions of tree-trimming records to generate more specific descriptor variables about the vegetation around each line segment.

The wires-down predictive models are then described in the following sections, starting with a general model and then moving to the specific vegetation and re-conductoring models. The general model uses all of the historical wires-down outages to predict the location of future failures and better understand the significant variables. Later sections then propose models based on specific failures modes (vegetation and equipment failure) to predict their unique failure locations.

Based on the results, the sections also detail how the appropriate mitigating preventative maintenance can then be applied to the predicted critical locations, optimizing their effectiveness.

1.7 Important Terms

- **Wires-down Event** – A failure which occurs when an electrical power line comes in contact with the ground or a ground-based object.
- **Conductor** – An industry term for a power line. Re-conductoring is the process of replacing old or failed power lines with new power lines.
- **Line Segment** – A term used in this analysis to denote the length of a line that spans between two power poles. The typical length in the PG&E distribution network is approximately 250 feet.
- **splice** – A mechanical join of two conductors that is usually necessitated when installing a new length of conductor or when repairing a split line. Different methods for splicing are available, but they all introduce an anomaly along the conductor whose strength is dependent on many variables.
2 Data

The initial portion of the project focused on finding and aggregating a host of data sources that would form the foundation of the predictive models. To determine the necessary data streams, interviews with company experts and suggestions from the literature review were used to formulate a list of potential variables in wires-down events. Then, internal and external data sources were identified to represent these variables in the models.

While PG&E has made concerted efforts in recent years to collect data and utilize it in their decision-making, most of these data streams are housed only in their immediate department (e.g. tree trimming records are owned and used only by the vegetation department). Therefore, each of the individual data streams had to first be collected and then merged into a cohesive dataset. Often, the individual sources required data cleaning to remove null or default values, manipulation to match the models’ level of granularity, and geospatial analysis to assign spatial data to each line segment.

2.1 Summary of Data Sources

The final dataset consisted of ten data sources that described the daily weather, the local vegetation environment, and the physical properties of each line segment combined with the outage history.

Weather – An external dataset from the U.S. National Oceanic and Atmospheric Administration (NOAA) called the “Daily Surface Summary Data” was utilized, which provided daily summary data for each of approximately 250 weather stations in California. The weather from the nearest station was associated with each line segment using the shortest straight-line distance. This daily data consisted of 35 variables including average/max rainfall, average/max temperature, and average/max wind speed. Additional features were added to capture cumulative weather effects (e.g. sum of weekly rain, sum of weekly wind speed) and lag effects (e.g. days since the last major wind day) because the PG&E experts hypothesized that storms may weaken the system but failures may not occur until the following days. These variables were intended to capture those lag and cumulative effects if they existed in the data.

Drought – An external dataset called the “US Drought Monitor” was incorporated which provided weekly, county-level drought data based on six severity levels. During the early portion of the dataset in 2012 and 2013, the drought stage was varied amongst the six severity levels across the PG&E network which provided a source of variance in the drought variable which aided in capturing the drought effect. However by the later years of 2014 and 2015, a large percentage of the region was in the most severe drought stage and therefore provided less data insight for comparing the locations once they all reached the most critical stage.
Conductor Properties – An internal dataset that contained the physical properties of the line segments, including the size, phase, and material of the conductor.

Splices – An internal dataset that provided the GPS location and circuit of every line splice. The GPS location was associated with the nearest line segment on the correct circuit using straight-line distance. If the distance between the nearest line segment and the splice location was an extreme outlier, the splice was cleaned from the dataset and assumed to be a data collection error.

Vegetation Land Cover – An external GIS dataset from the US Department of Interior called the “National Land Cover Database” was utilized to classify each line segment into one of 33 different land cover types (e.g. Conifer Forest, Desert). This dataset was the standard in the literature review and provided a fairly coarse description of the vegetation environment around each line. The utilization of the vegetation trim records was a contribution that succeeded in augmenting this coarse variable with more granularity.

Vegetation Trim Records – An internal dataset that contained over four million tree trimming records was incorporated into the database that detailed the tree species, height, and width of every tree that PG&E has trimmed over the past ten years. To better understand the specific vegetation environment at the line segment level, an algorithm was created that associated each trim record with the nearby line segment, and then calculated features unique to those trim records and that line segment. This data provided additional vegetation specificity along the line segment including the dominant tree species, the maximum and average height of those trees, and the density of the trees.

Location-Based Properties – An internal GIS dataset was extracted by overlaying the line segments onto geographic-based data, providing population density from the U.S. census data, land elevation, and classifying the lines by district and county. Additionally, PG&E has designated corrosion zones for severe corrosion regions which were incorporated with this GIS dataset.

Outages – An internal dataset containing the dates and line segment of the wires-down outages was included in the dataset which detailed the binary variable of whether a failure occurred that day and provided the failure mode (e.g. vegetation, animal, etc.).
2.2 Additional Unused Data Sources

Besides the ten data streams incorporated into the final database, there were additional datasets that were located but ultimately left out of the final version. Usually this was due to difficulty in merging the data or the incompleteness of the source. It is likely that all of these data sources would improve the models' performance, but the timeline of the project and the effort required for incorporation did not allow their use. This section lists those excluded data sources, the reasoning they were left out, and what steps are required to successfully incorporate the data.

**Age Data** – The age or installation date of each asset is a potential critical variable in estimating the level of asset degradation. This data would be especially beneficial in calculating equipment failure probabilities because that failure mode is assumed highly correlated to the age and decay of the asset over time. An incomplete dataset of installation data that covered only a portion of time was available, but the source was too incomplete to utilize. A potential solution was to use the installation date of the power-line poles as a proxy for the conductor installation date, but the pole data ultimately was unusable as well. In order to improve the equipment failure predictions, PG&E must uncover a more complete set of installation records or succeed in incorporating the pole data to use as an age proxy.

**Pole Data** – The pole dataset described the location and type of poles used to support the conductors. These poles can range in their height and thickness, their support geometry, and their material (wood, metal); all variables that potentially have an impact on their strength and ability to resist a wires-down failure. PG&E has a complete dataset that contains the locations and properties of all two million poles in their system. However, the data is based on a geospatial location (latitude and longitude), and not associated with the specific line segment the pole is supporting. During this project, an algorithm was created to assign the poles to their nearest line segments, but this program had limited success due to mismatches between the lines and pole locations. Often the locations of the two features differed significantly, and produced inaccurate pole assignments which could not be manually corrected given the number of poles. Several GIS projects at PG&E aimed at cleaning and synthesizing datasets are ongoing, and these should address this mismatch issue, allowing the use of pole data in the near future.

**Loading Profiles** – The loading profiles of the conductors could highlight a link between overloading of circuits and failures. While this hypothesis was mentioned in the literature review, no models were found that actually used this data in a prediction model. A large dataset was collected at PG&E from the group who monitors the system's load profiles, but ultimately this dataset was not incorporated into the model. The data was too coarse to provide immediate benefit and was the lowest priority of the sources to implement. Further
research into the anticipated benefit of the loading data should be done before investing the time to sync this data source.

**Inspection Records** – PG&E is continually conducting inspections on their conductors to satisfy regulations and for additional failure prevention. Knowledge of when a line segment was last inspected could improve the prediction models since inspections are designed to spot hazardous vegetation and visible equipment issues. A dataset was aggregated that provided the last inspection timeframe for each line segment. However, this model improvement arose near the end of the project, and once issues were encountered trying to associate this dataset in both time and space, this inspection data was ultimately excluded from the model. For next generations of the prediction models, this data source can be easily incorporated and should improve the performance.

### 2.3 Model Granularity

The granularity of the database was set at the “daily, line segment”-level, meaning the dataset contained all the variables to depict the system at the line segment level (span of line between two power poles) on each day. In the PG&E system, the line segment span was typically 250 feet, and provided the highest level of fidelity available from the given datasets. The daily-level was chosen to capture the unique weather events of each day while avoiding the unnecessary complexity of hourly data only required for a real-time failure prediction model.

Additionally, by having the initial data segmented at a very discretized level, the prediction results could then be aggregated at higher levels to provide better accuracy. A common method, determined from the literature review, is to group the lines by circuits or by proximity (e.g. all lines within a 2 mile by 2 mile region). This topic is discussed later in this paper in terms of operationalizing the modeling results.

### 2.4 Weather Clustering Methodology

The daily summary weather variables and the drought variables combined for twenty-nine predictor variables. This represented a significant portion of the model variables, and posed an interesting dilemma for the model formulation. The significance of weather in causing wires-down failures is intuitive, and thus an accurate prediction model needed to contain weather features. Yet the aim of the models was to direct preventative measures more associated with the other, more mitigable variables in the model. Additionally, the ability to predict future weather with the same high level of specificity provided in the summary weather data was impractical (e.g. difficult to forecast the future wind speed, temperature, and rain amount), and thus would hamper efforts to predict future failure locations if included at the same level of granularity.
Therefore, in order to include the significance of the weather but reduce its impact on the model's
degrees of freedom, the weather and drought variables were combined into a single feature using k-means
clustering. Each of the clusters can be envisioned as a “daily weather climate”, a combination of the
underlying twenty-nine variables that accounts for the temperature, wind, drought, etc. When the model is
executed, the weather clusters are developed from the training data, and each line segment is assigned to the
weather cluster that best approximates the weather that they experienced that day. Those same cluster
centroids (weather environments) are then used to associate the line segments in the test data to the most
approximate weather environment for each day, maintaining continuity between the weather environments
created with the training data and assigned to the test data.

Before implementing the clustering algorithm, the weather variables were scaled so that one variable
did not dominate the feature, since the range of each variable differed significantly (temperature was originally
between 0°-100° while rain depth was only 0”-2”). Additionally, principal component analysis (PCA) was
performed to help reduce the dimensionality of the data. PCA identified the variables that contributed the
most to the variance of the weather variables, and allowed the non-influential variables to be removed,
thereby eliminating variables that would consume the model's degrees of freedom without significantly
differentiating the weather state. Based on the PCA findings shown in Figure 4, the top seven components
were the most influential to the variance and were retained for the weather clustering.

![Figure 4: Principal Component Analysis of Weather Variables](image)
For the k-means clustering approach, an analysis was conducted to determine the optimal number of weather clusters to create to sufficiently capture the unique weather environments experienced throughout the PG&E territory each year. Theoretically, an unlimited number of clusters could be created that exactly replicated each unique weather environment on each day. However, the goal of the weather clustering was to find some commonality amongst the weather environments, and to reduce the dimensionality of the weather variables. Therefore, this analysis was necessary to find the best balance between differentiating between the various climates while still minimizing the overall complexity of the dataset.

Ultimately, the number of weather clusters to generate was based on an iterative calculation that examined the “Within Sum of Squares” error relative to the amount of clusters, essentially measuring how well the clustered results approximated the actual data. Figure 5 details the decrease in the error as the number of clusters increases, illustrating that more clusters better approximate the actual data but with a decreasing marginal benefit for each additional cluster. For the model formulation, one hundred clusters was chosen as an appropriate amount of weather clusters since one hundred clusters reduced the majority of the error and any additional cluster provided minimal beneficial impact.

Figure 5: Number of Weather Cluster Optimization
3 Model #1: General Wires-Down Prediction Model

3.1 Model Methodology

The goal of the general wires-down prediction model was to use all of the historical failure events to predict the location of future failures and to determine the key inputs in making those predictions. While not all of these inputs may be causal of a wires-down event, a better understanding of which variables are typically present could be useful in tailoring the type and location of wires-down preventative maintenance. This general model incorporated all wires-down failures, agnostic to the specific failure mode, relying only on the binary response of failure or no failure for a particular line segment on a particular day.

The modeling techniques that were tested were logistic regressions, decision trees, and random forests; each chosen because of their previous promising results in the literature. Multiple iterations of each model were run which adjusted the variables included and the models' parameters to achieve the best prediction results. Based on those results, the logistic regression technique provided the best approach per out-of-sample testing and was implemented in the final model formulation. The following sections present the formulation and the results from the logistic regression model. For comparison purposes, Appendix B contains the results from the highest-performing decision tree model.

3.2 Logistic Regression Formulation

The logistic regression model formulation [5.1] is the standard logistic regression equation adapted to the wires-down dataset variables.

\[
\ln \left( \frac{p}{1-p} \right) = b_0 + b_w X_w + b_a X_a + \cdots + b_s X_s
\]

where \( b_w \) and \( X_w \) refer to the weather cluster, \( b_a \) and \( X_a \) refer to the line properties, and \( b_s \) and \( X_s \) refer to the splice property. The probability of a wires-down event is calculated by solving for \( p \) [5.2].

\[
P = \frac{e^{(b_0+b_w X_w+b_a X_a+\cdots+b_s X_s)}}{1+e^{b_0+b_w X_w+b_a X_a+\cdots+b_s X_s}}
\]

The variables included in the final model [5.1] are presented below. Some of these variables are continuous (example: the surrounding tree density) while others are categorical (the material of the conductor). Each variable was explicitly designated as either continuous or categorical, and the R software package handled them accordingly.
1) Response variable: Outage probability (between 0 and 1)

2) Predictor variables:

- COASTAL_INTERIOR: designation whether in PG&E-defined coastal or interior zone
- CORROSION: designation if line segment is in a PG&E-defined corrosion zone
- DENSITY: population density from US Census data
- VegClass: vegetation designation from National Landcover Database
- Elevation: elevation of the terrain at the line segment
- k_wx: different weather clusters (environments)
- splice_Qty_Ind: splice quantity categories
- MATERIAL: material of the wire conductor
- PHASE_CD: phase of the wire conductor
- SIZECAT: thickness of line segment
- PHASE_COUNT: number of phases of the wire conductor
- TYPE_CONSTR: PG&E code that describes material, size, and phase of line segment
- DIST_CODE: geographical district code

Vegetation variables calculated from the historical trim records:

- tree_AvgDBH: average DBH along line segment. DBH is the diameter at breast height (1.4m off the ground) of the tree trunk.
- tree_AvgHeight: average height along line segment
- tree_Density: density of trees along line segment
- tree_MaxDBH: max tree DBH along line segment.
- tree_MaxHeight: max tree height along line segment

3.3 Out of Sample Testing

To compare the modeling techniques and to measure the accuracy of their predictions, out-of-sample testing was utilized in which the total dataset was divided into a training and test set based on a 70%-30% split. In this approach which minimized over-fitting of the model, the model was built using only the training set, and then predictions were made on the test set. The predicted results were then compared to the actual outcomes to judge the accuracy of the model. This out-of-sample testing simulated actual use by PG&E where the model would be trained on historical data and then used to predict future outages.

Initially, this subdivision of the data into a training and test set was done completely randomly with respect to time and location. However, some concern was raised about the true independence of outages from the same storm event as it moved across the territory, and whether this potential lack of true independence could artificially enhance the prediction capability if failures from the same storm were
included in both the training and test set. To prevent this occurrence, the subdivision of the dataset was changed so that training and test sets were not divided randomly with respect to time but rather separated into continuous time chunks (e.g. training set from: 1/1/12 – 6/30/14 and test set from 7/1/14 – 5/19/15). Later in the project, the time span of the test dataset was iteratively shifted to demonstrate the model was robust over time.

3.4 Splice Significance

As shown in Appendix 1 with the variable significance rankings from both the general and equipment models, the presence of splices on a line segment is highly indicative of a future failure. This section details why the splice data is nuanced and how it was cautiously implemented.

The splice dataset is comprised of inspection records that detail the frequency and type of splices on the line segments. However it fails to contain any installation date for when the splice was actually installed on the line segment. The lack of temporal installation data presents a prediction challenge. Namely, if the splices are assumed already installed at the temporal beginning of the dataset, the model could actually be using a splice to predict a failure that actually caused that splice to be installed. This occurrence would artificially skew the prediction results towards higher accuracy.

Recognizing this issue, two steps were taken to address it in this project. First, the statistical significance of splices in predicting failure was tested to validate that splices were actually significant. If they were not, then the splice variable could simply be removed from the models. And secondly, the splice data was incorporated into the model as a categorical variable that minimized the impact of any one individual splice.

3.4.1 Splice Statistical Significance T-Test

To test the statistical significance of splices to wires-down failures, a one-sample T-test was used that compared the mean and standard deviation of the splices in the failure sample set to the mean and standard deviation of the splices in the entire population of power lines. The null hypothesis was that no significant difference existed between the mean number of splices in the failure sample set and in the population.

Splices in the Wires-down Failure Sample Set:
Sample mean = 2.677 splices/line
Sample std dev = 3.3015 splices/line
Sample size = 3105 lines

Splices in the Population Sample:
Pop mean = 5.004 splices/line
Pop std dev = 0.1575 splice/line
Based on the t-test, the two-tailed P-value was less than 0.0001, signifying that splices were statistically significant in failures with 95% confidence. This result confirmed that splices should be in the models to improve their predictive capability.

### 3.4.2 Splice Categories

The next step was to determine how to include the splice data knowing that some of the splices were due to historical failures the model was trying to predict. To mitigate the importance of any one splice, the models incorporated the frequency of splices on a line segment as categorical: 0 splices, 1-3 splices, 4-6 splices, and 7 or more splices. This limited the possibility that a single splice in the dataset could artificially skew the model into predicting its failure because now the model was considering the summation of splices on the line segment. Only if that splice would change which splice category the line would be characterized by, would the model be slightly skewed. While this approach did not eliminate the concern, it was a compromise between fully eliminating a variable that was shown to be significant and letting that variable have an artificially high influence on the results.

### 3.5 Ranking Variable Significance

One of PG&E’s primary intentions for the predictive models was to better understand which variables were driving the wires-down events and causing specific failure modes. The general model’s ability to answer these questions was limited because the model was inclusive of all failures and therefore could not differentiate between the causes of a vegetation failure versus an equipment failure. Additionally, caution was used in labeling any variable as causal because the model may have been highlighting variables that were simply present at most failures but not actually causing the failures. With those caveats known, ranking the variables’ significance was still highly valuable to better understand the results and look for possible preventative maintenance improvements.

To rank the variable significance of the logistic regression model, three different methods were tested: 1) standardized coefficients, 2) p-values of a Wald Chi-Square test, and 3) a measurement of the individual impact of each variable on the AIC and residual deviation of the model.

- **Standardized coefficients:** All coefficients are adjusted to have the same scale so that the magnitude of the coefficient illustrates the significance of the variable.

- **P-values of a Wald Chi-Square test:** The p-value of this test indicates if there is evidence of an association between the variable and the outage. However, it does not describe the magnitude of its effect.

- **Individual variable impact:** Manually, each variable was individually removed from the model, the model re-fit, and the fit statistics were calculated to judge the change in the fit based on that one variable.
The ranking results from each of these three methods largely matched except for variables with low levels of significance. Therefore, the simpler standardized coefficient method was chosen as the primary method to utilize in the general model and the subsequent models. Appendix A contains the full ranking of variable significance for the general model. Broadly, the key variables were certain weather environments, line segments containing three or more splices, and several geographic districts.

3.6 General Model Results

The receiver operating characteristic (ROC) curve for the general model is depicted in Figure 6. This curve graphs the false positive rate versus the true positive rate based on different discrimination thresholds, which are plotted as points on the curve. The discrimination threshold is the cutoff value at which the failure probability calculated in equation 5.2 is either rounded down to 0 predicting “no failure” or rounded up to 1 predicting “failure”. As Figure 6 demonstrates, by adjusting the discrimination threshold, the ratio of true positives to false positives shifts, tailoring the model for different results and purposes.

A rough visualization of the model’s performance can be seen by comparing the ROC curve to the gray line in Figure 6. The gray line represents a random guess at where the wire-down failures would occur, and any points above that line indicate better performance than that completely random guess. This comparison illustrates that the general model is able to extract useful inferences from the data to make exceedingly accurate predictions compared to a random guess as the curve is tucked into the upper left-hand corner.

Figure 6: General Prediction Model ROC Curve
PG&E’s application of these prediction models is to highlight potential hot-spot locations in their system so preventative measures can be applied to the most critical locations. Therefore, the models were tailored to over-predict failures locations rather than under-predict by setting a lower discrimination threshold of 0.08, because the cost of over-predicting was negligible and only helped to highlight more problematic locations. In other scenarios, where the cost of over-prediction is higher and detrimental, a more conservative threshold could be set.

Based on a discrimination threshold equal to 0.08:

\[
\text{Sensitivity} = \frac{\text{Number of True Positives}}{\text{Number of True Positives} + \text{Number of False Negatives}} = 0.87
\]

\[
\text{Specificity} = \frac{\text{Number of True Negatives}}{\text{Number of True Negatives} + \text{Number of False Positives}} = 0.90
\]

The sensitivity measures the true positive rate and indicates the model’s ability to identify all the true positives (accurately predict failure locations) and avoid many false negatives (actual failures that are not predicted). Counter to that, the specificity finds the true negative rate, measuring how many true negatives (accurately predict non-failure locations) are identified and how good the model is at avoiding false positives (non-failures that are predicted as failures). Since these metrics can have a maximum of 1.0, the relatively high performance of both of them indicates that the model is appropriately identifying a large majority of the test set.

3.7 Plot of Predicted Results

The two maps presented below in Figure 7 show a comparison of the actual failure locations and the predicted failure locations from the general model. This side-by-side evaluation was useful to see where the model struggled to make accurate predictions. Per the maps, the model appeared to perform best along the coastal regions and in areas with a high number of failures, but found it difficult to predict the more sporadic failures in the central valley region (the lower right hand portion of the map).

The inference for why the model under-performed in some regions has to do with the different patterns between the vegetation and equipment failures. As the following sections will illustrate, the vegetation failures are highly correlated to the localized tree environment, and therefore these failures occur more concentrated in the regions with significant vegetation. The dataset contains numerous variables to describe this vegetation around each line. Conversely, the equipment failures are largely due to the aging of assets that occurs uniformly over the entire network, and results in more randomized failure locations. In the dataset, no age data is available to help predict these equipment failures. Therefore, the general model
performed well in the high failure locations because those are predominantly vegetation failures described by many variables in the dataset, but struggled to capture the more sporadic failures that are often equipment failures for which limited predictor variables exist.

3.8 Limitations of the General Model

Several limitations of the general predictive model have already been highlighted related to variable significance and its ability to predict certain failures more accurately than others. Another key limitation is that the model cannot be utilized as a one-for-one predictor of the number of failures over a given timeframe. Put in terms of PG&E, the model cannot be used to generate a year-end performance metric for the number of wires-down events anticipated over the next year. This is due to the sheer number of predictions generated each day, meaning even a small over-prediction error rate will estimate the number of failures magnitudes above the actual amount.

*Illustration of the Over-Prediction:*

- Number of Line Segments = 1.2 million
- If each line segment is predicted to fail or not fail each day for one year:
  
  Number of predictions = 1.2 million x 365 days = 438 million predictions / year
- If model over-predicts by even 2%:
  
  2% x 438 million predictions = 8.7 million failures  
  
  [Actual failures = a few thousand]
Hypothetically, a factor could be applied to adjust the predicted failure count down in hopes of generating a usable metric. But the magnitudes of the numbers are so different and the necessary precision of the failure metric so small that any factored metric would be highly questionable. Therefore the recommendation is to not use the model’s output to formulate metrics, but rather to serve its intended role of highlighting critical failure locations and critical variables.

3.9 General Model Conclusions

The general model demonstrated successfully the ability to predict wires-down failure locations at a level well above a naive guess and at a very granular, line-segment level. Analysis of the results highlighted relative underperformance in the area of equipment failures, and attributed it to a lack of age data in the dataset. The significant variables confirmed the experts’ intuitions that specific weather, dense vegetation, and multiple splices are highly correlated to failures, but the model offered little insight into which variables caused which failure modes. Finally, the model’s value was reiterated to be its ability to forecast hot-spot failure locations and not its use in yearly metric creation.
4 Model #2: Vegetation-Specific Wires-Down Prediction Model

4.1 Model Formulation

Unlike the general model which accounted for all historical wires-down failures to train the prediction model, the next iteration of models, described in Sections 4 and 6, utilized only the vegetation and equipment-related failure modes respectively. The purpose of these more specific models is to predict the location of failures due to specific failure modes so that the appropriate preventative maintenance can be applied to mitigate the risk.

The basic model formulation for the vegetation prediction model is identical to the general model, using a logistic regression approach. In this case, the failure set only includes vegetation-related failures, as designated in the internal failure reports. Additionally, to avoid over-fitting, the predictor variables were limited to vegetation and location-based variables. This decision was made to prevent non-causal variables from potentially artificially enhancing the predictive capability.

The list of the predictor variables included in the final iteration of the vegetation logistic regression model is shown below. As in the general model, each variable was designated as either continuous or categorical, and the R software accounted for each accordingly.

1) Response variable: Vegetation outage probability (between 0 and 1)
2) Predictor variables:

   COASTAL_INTERIOR............designation whether in PG&E-defined coastal or interior zone
   VegClass..........................vegetation designation from National Landcover Database
   Elevation............................elevation of the terrain at the line segment
   DIST_CODE..........................geographical district code

Vegetation variables calculated from the historical trim records:

   tree_AvgDBH ..................average DBH along line segment. DBH is the diameter at breast height (1.4m off the ground) of the tree trunk.
   tree_AvgHeight ................average height along line segment
   tree_Density ....................density of trees along line segment
   tree_MaxDBH ...................max tree DBH along line segment.
   tree_MaxHeight ..................max tree height along line segment

A noted absence from the list is the weather variable which was omitted from this prediction model. While it was demonstrated that weather is a significant variable in predicting failures, the ability to future forecast the weather is limited. Therefore, to create a model that PG&E could realistically implement, weather was omitted at the cost of reduced predictive power. It is important to recognize however that the location
variables have an underlying component of weather because of the historical weather patterns each area experiences. So while no weather variable is explicitly included, the location features do incorporate an obstructed linkage to the weather.

4.2 Ranking Variable Significance

Similar to the general model, the predictor variables were ranked in order of significance using the standardized coefficients method as part of the post-processing of the results. Appendix A contains the full ranked list. Based on the list, the most hazardous vegetation environment is the conifer forest, followed by shrubs, hardwood forest, and then hardwood woodland. Shrubs are an unlikely inclusion in this list, but the hypothesis is that some critical coastal areas prone to bad weather are labeled as shrubs and therefore that vegetation class shows high correlation to wires-down events even though the failure cause is actually bad weather. In terms of highly significant locations, Section 5 analyzes the predicted hot spots, and examines how these locations compare to the traditional areas that PG&E has targeted.

4.3 Vegetation Model Results

The ROC curve for the vegetation-related wires-down prediction model is depicted in Figure 8. When compared to the ROC curve from the general model shown in Figure 6, this vegetation model curve is shallower and demonstrates less predictive capability. However, the model still shows substantially better results than a naïve guess shown by the gray line. The decrease is primarily due to the omission of several highly correlated variables in the general model such as the weather and the splice count, and is a necessary sacrifice in order to make a model focused only on vegetation wires-down failures.
For the vegetation model, the discrimination threshold was set at 0.09 to balance the true positive and false positive rate. Based on this discrimination threshold:

\[
\text{Sensitivity} = \frac{\text{Number of True Positives}}{\text{Number of True Positives} + \text{Number of False Negatives}} = 0.67
\]

\[
\text{Specificity} = \frac{\text{Number of True Negatives}}{\text{Number of True Negatives} + \text{Number of False Positives}} = 0.83
\]

In comparison to the sensitivity and specificity of the general model, the vegetation model struggled to minimize the number of false positives as the true positive rate increased. With the reduced number of variables included in the vegetation model, there was less variance within the descriptive data between a failure and non-failure, thereby providing fewer distinguishing features for the model to base its predictions. These performance metrics still demonstrated a substantial amount of predictive capability that will be useful in identifying critical vegetation locations.
5 Implementation of the Vegetation Model Results

5.1 Aggregating Predictions

PG&E predominately utilizes tree trimming to mitigate vegetation wires-down failures, sending a small crew onsite to cut back the limbs around a line to a regulated clearance. Besides the crew, these operations usually involve a lift and can occur in a variety of environments from alongside the road to more difficult settings like in the forest. Therefore, when a tree-trimming project is prescribed, it is normally completed on at least several miles of lines to minimize the impact of travel and setup time to the site.

The prediction model generated failure probabilities for each line segment, with the typical length averaging 250 feet. While the optimal approach would be to specifically target each discrete line segments rated most critical with tree trimming, this method was operationally inefficient given the travel and setup time mentioned previously. Therefore, the individual line segment failure probabilities were aggregated into groupings more consistent with the current method. For this project, the lines were aggregated up to their source side device (SSD), essentially their upstream protective device, resulting in continuous segments that averaged several miles in length.

\[
C_j = \frac{\sum p_i}{\sum length_i} \times 1000 \quad \text{forall lines in SSD } j
\]
5.2 Identifying Critical Locations

The criticality ranking \([6.1]\) for each SSD \((C_i)\) was the summation of the individual line segment failure probabilities \((p_i)\) normalized based on their length to prevent biasing towards longer segments. Table 1 shows the number of SSD segments and their associated line miles based on certain criticality thresholds; the higher the criticality ranking, the higher the likelihood of a wire-down vegetation failure on that segment. The mean criticality ranking for the system was 0.083.

The main conclusion from Table 1 is that a small percentage of the network has significantly higher failure probabilities relative to the typical line segment, and offer the best opportunities to mitigate wires-down vegetation failures. In a network covering over 100,000 miles of lines, just over 200 miles of those lines are highlighted as extremely vegetation-failure critical. This is the best contribution from the vegetation model: the ability to simplify the decision-making for PG&E by focusing their efforts onto a more manageable and perceivable amount of the network.

**Table 1: Vegetation-Related Wires-Down Criticality Rankings**

<table>
<thead>
<tr>
<th>Criticality Ranking</th>
<th>Number of SSD Segments</th>
<th>Total Linefeet</th>
<th>Total Miles</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 4.0</td>
<td>3</td>
<td>1840</td>
<td>0.35</td>
</tr>
<tr>
<td>&gt; 3.0</td>
<td>11</td>
<td>10706</td>
<td>2.0</td>
</tr>
<tr>
<td>&gt; 2.0</td>
<td>39</td>
<td>54759</td>
<td>10.4</td>
</tr>
<tr>
<td>&gt; 1.0</td>
<td>434</td>
<td>1087579</td>
<td>206.0</td>
</tr>
<tr>
<td>&gt; 0.083</td>
<td>28983</td>
<td>210520328</td>
<td>39871.3</td>
</tr>
</tbody>
</table>
To visualize the locations of critical SSDs, Figure 9 plots the top 20,000 SSD segments onto the geographic regions of California. The heat map is the density of the critical SSDs in each region, which was calculated by taking the number of critical SSDs per region divided by the total number of SSDs in that region. Based on the plot, the vegetation critical areas are heavily concentrated and predominantly coastal. Instead of being uniformly distributed across the territory, the critical regions are relatively few, signaling that many vegetation failures are occurring in only a small subset of the network. Upon further investigation, these regions are heavily forested and susceptible to bad weather, highlighting the strong correlation in vegetation failures between the local tree environment, the local weather pattern, and the wire-down probability.

Figure 9: Critical Regions for Vegetation Wires-Down Failure
5.3 Implementation into PG&E Operations

To utilize the vegetation model results, the SSDs were sorted in order of their criticality ranking. As previously concluded, a small percentage of the lines, totaling 206 miles, had very high criticality rankings. This length of line represented a fairly small percentage of the yearly tree trimming mileage that PG&E performs, and therefore offered the potential that PG&E could immediately decrease their system’s vulnerability within one year. For 2016, the vegetation department was investigating these 206 miles of critical line rated over 1.0 to validate the model’s results, and mitigating the vegetation risk with tree trimming when confirmed.
6 Model #3: Equipment-Specific Wires-Down Prediction Model

6.1 Model Formulation

The equipment-related prediction model paralleled the vegetation model effort to examine a specific failure mode to better understand its significant variables and the critical regions for this failure type. In the instance of equipment-related wires-downs, the preventative maintenance is increased inspections to find poor connections and to spot degradation, combined with the more drastic step of re-conductoring.

The logistic regression formulation for the equipment-specific prediction model was identical to the vegetation model except for the inclusion of only equipment wires-down events in the failure set and the usage of only location-based and equipment-related predictor variables. This model also excluded the weather predictor variable to simplify future predictions.

1) Response variable: Equipment-related outage probability (between 0 and 1)

2) Predictor variables:
   - COASTAL_INTERIOR designation whether in PG&E-defined coastal or interior zone
   - CORROSION designation if line segment is in a PG&E-defined corrosion zone
   - DENSITY population density from US Census data
   - Elevation elevation of the terrain at the line segment
   - splice_Qty_Ind splice quantity categories
   - MATERIAL material of the wire conductor
   - PHASE_CD phase of the wire conductor
   - SIZECAT thickness of line segment
   - PHASE_COUNT number of phases of the wire conductor
   - TYPE_CONSTR PG&E code that describes material, size, and phase of line segment
   - DIST_CODE geographical district code

6.2 Ranking Variable Significance

The predictor variables were ranked in order of significance using the standardized coefficients method. Appendix A contains the full ranked list for the equipment model. As in the general model, the occurrence of splices on the line segment was a significant indicator of future failures. Since most splices are created by a past failure, the significance of splices indicates that repeat failures are happening frequently on lines, and that certain line segments are failure hot spots that have experienced multiple failures over time.
Additionally, the significance of splices could highlight a splicing repair problem where the repaired condition is not as strong as the original wire. Future analysis must be conducted to determine if the equipment failures are due to splices separating over time, and if so, then the splicing repair procedures must be examined to ensure they return the wire to its prior un-spliced strength.

6.3 Equipment Model Results

As with the previous models, a ROC curve was generated for the equipment prediction model and used to find the optimal discrimination threshold. The ROC curve shown in Figure 10 demonstrates better predictive capability than the vegetation model in Figure 8, but still lags the general model since fewer predictor variables were included.

The shape of this ROC curve is unique compared to the other models because it has a distinct bend rather than a gradual arc. This sharp change signals that a large portion of the equipment failures have distinct features that make their prediction fairly easy, and then the remaining failures exhibit less distinctive features that require a trade-off between increases in the true positive and the false positive rate in order to capture.

![ROC Curve for Equipment-Related Wires-Down Prediction](image)

*Figure 10: ROC Curve for Equipment-Related Wires-Down Prediction*
The discrimination threshold was set at 0.05 which occurs at the sharp bend in the ROC curve since that resulted in a high true positive prediction rate with almost no false positives. With this discrimination threshold of 0.05:

\[
\text{Sensitivity} = \frac{\text{Number of True Positives}}{\text{Number of True Positives} + \text{Number of False Negatives}} = 0.83
\]

\[
\text{Specificity} = \frac{\text{Number of True Negatives}}{\text{Number of True Negatives} + \text{Number of False Positives}} = 0.98
\]

The sharp bend in the ROC curve and the sensitivity and specificity metrics allude to the presence of highly correlated variables that successfully predict equipment failures. In this model, that significant variable is the splice count on the line, which has been discussed in-depth during this thesis. This finding emphasizes the need to have a comprehensive dataset of installed splices, and to study the failure mechanisms of those splices when they break.
7 Implementation of the Equipment Model Results

7.1 Aggregating Results

The individual line segment failure predictions from the equipment model were aggregated to their upstream protective device (SSD) and normalized for their length similar to the vegetation results to find the cumulative criticality ranking of each SSD. Since the ultimate preventative step for equipment-related wires-down failure is re-conductoring, these SSD lengths provide a manageable length of line that is operationally efficient and is similar to their current project lengths. Table 2 shows the breakdown of the SSD criticality rankings; the higher criticality ranking signaling a higher likelihood of equipment-related wires-down failure.

Similar to the vegetation results shown in Table 1, a small subset of the network has criticality rankings that far exceed the mean ranking. However, the outliers appear less severe than the vegetation results since the max ranking just exceeds 2.0 from the equipment model rather than the 4.0 from the vegetation model.

Table 2: Equipment-Related Wires-Down Criticality Rankings

<table>
<thead>
<tr>
<th>Criticality Ranking</th>
<th>Number of SSD Segments</th>
<th>Total Linefeet</th>
<th>Total Miles</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 2.0</td>
<td>3</td>
<td>1735</td>
<td>0.33</td>
</tr>
<tr>
<td>&gt; 1.5</td>
<td>6</td>
<td>5649</td>
<td>1.1</td>
</tr>
<tr>
<td>&gt; 1.0</td>
<td>36</td>
<td>33690</td>
<td>6.4</td>
</tr>
<tr>
<td>&gt; 0.5</td>
<td>311</td>
<td>413957</td>
<td>78.4</td>
</tr>
<tr>
<td>&gt; 0.0193</td>
<td>19666</td>
<td>184933618</td>
<td>35025</td>
</tr>
</tbody>
</table>

7.2 Identifying Critical Locations

A density plot of the top 20,000 most critical SSDs in PG&E’s system is shown in Figure 11. The heat map is the density of the critical SSDs in each region, which was calculated by taking the number of critical SSDs per region divided by the total number of SSDs in the region. Comparing this equipment-related plot versus the vegetation-related plot in Figure 9, the critical regions are more evenly dispersed for equipment failure than vegetation failure. As discussed earlier, the hypothesis for this dispersion is that equipment failure is largely due to aging infrastructure which occurs more uniformly on all assets, while vegetation failures are more concentrated due to their dependence on the local vegetation.
Figure 11: Critical Regions for Equipment-Related Wires-Down Failure

7.3 Implementation into PG&E Operations

The outcome of this modeling and aggregation is a ranked list of critical SSDs for equipment-related failure. Since the process of re-conductoring is time-intensive and costly, any optimization on the most beneficial locations to re-conductor could yield significant positive results for PG&E. Therefore, this prioritized ranking of segments has been incorporated into the tools for the PG&E experts to help better focus their attention on a smaller, more critical subset of the network and aid their location-selection process.
8 Conclusions and Improvements

8.1 Conclusions

The work presented in this thesis demonstrates the capability that PG&E has to predict wires-down failures despite the presence of variability, infrequency of failures, and its large scale. The initial sections outlined the available datasets and described several novel approaches to incorporate weather and extract more value from their vegetation trim data. The three models each demonstrated substantial predictive capability given the variability, and they highlighted critical factors and line segments for PG&E to investigate. This work acknowledged that many improvements can be made for better results and usability, but overall presented a hopeful outlook for future failure prediction and more data-supported decisions. The models' ability to highlight critical portions of the network reduced the magnitude of the system and will allow the PG&E experts to focus their attention.

This thesis generated three significant contributions to the academic research and to PG&E: 1) a uniquely comprehensive dataset that was utilized in this work and can form the foundation of future analytics projects, 2) the successful implementation of machine learning techniques to predict wires-down events, and 3) a better understanding of the major factors in these wires-down events from the rankings of variable significance from each predictive model.

8.2 Implementation of the Models

This project was conducted as an initial effort to explore PG&E’s available data, test if valid failure predictions could be extracted from the dataset, and understand what are the challenges and appropriate next steps. Given the early-stage maturity of the project, most the data collection and manipulation was done using comma separated value (CSV) documents, which were static extracts from the various data sources. These CSVs were manipulated and merged using Python scripts into the final, combined dataset that was also a CSV. The modeling was produced in R, and the results were exported as CSV files that can be viewed in Microsoft Excel. Finally the results mapping was plotted with Python libraries.

PG&E has begun to utilize the ranked lists of critical SSDs from the vegetation and equipment models to make more data-supported decisions. The critical vegetation lines are being investigated by their vegetation experts and checked in the field, with the goal of examining all SSDs rated over 1.0 for vegetation criticality. Additionally, the asset management division of PG&E has continued to grow their analytics focus, bringing on staff to continue this effort and other similar projects.
8.3 Improvements to the Dataset and Prediction Models

Since this project was a static proof-of-concept, many improvements can be made to boost results and usability. The most significant improvement is the data: incorporating more predictor variables and automating the data aggregation process. The current dataset excludes the likely causal variables of equipment age and loading data as well as any temporal inspection data. This data either does not exist or was too complex to include in the given timeframe. Efforts to gather and incorporate this data should increase the predictive capability of the models, especially for the equipment model. Additionally, the current dataset is a static extract from multiple sources. While some of this data is largely constant over time, other variables are more dynamic and could benefit from an automated method to periodically update.

Besides the data improvements, the other major necessity is to validate the models’ results thru on-site visits. This effort was done on a small scale at the conclusion of the project by comparing the modeling results to historic trim locations and discussing them with PG&E experts, but the effort needs more depth to understand how well the models’ output aligns with the on-site observations.
Appendix A: Ranking of Variable Significance from Each Prediction Model

The following figures show the ranking of variable significance for the three prediction models using the standardized coefficients method. This method adjusts the variables so that the resulting variance of all variables equals one, allowing comparison of dissimilar units and scales of measurement. A description of the variable names is also listed below.

Index of Variable Names
* signifies a part of the variable name that changes based on the value

- k_wx* ................. different weather clusters (environments)
- splice_Qty_Ind>* .......... splice quantity categories
- DIST_CODE* .................. geographical district code
- VegClass* ................. vegetation designation from National Landcover Database
- MATERIAL* .................. material of the wire conductor
- PHASE_CD* .................. phase of the wire conductor
- elevation .................... elevation of the terrain at the line segment
- CORROSION* ............... designation if line segment is in a PG&E-defined corrosion zone
- DENSITY* .................... population density from US Census data
- TYPE_CONSTR* ............ PG&E code that describes material, size, and phase of line segment
- SIZECAT* ................... thickness of line segment
- COASTAL_INTERIOR .......... designation whether in PG&E-defined coastal or interior zone.

Vegetation variables calculated from the historical trim records:
- tree_AvgHeight ................ average height along line segment
- tree_MaxDBH .................. max tree DBH along line segment. DBH is the diameter at breast height (1.4m off the ground) of the tree trunk.
- tree_AvgDBH .................. average DBH along line segment
- tree_Density .................. density of trees along line segment
- tree_MaxHeight .............. max tree height along line segment
Figure 12: Ranked Variable Significance from the General Model
Figure 13: Ranked Variable Significance from the Vegetation-Specific Model
Figure 14: Ranked Variable Significance from the Equipment-Specific Model
Appendix B: Decision Tree Results from General Model

As discussed in Section 3, multiple prediction algorithms were tested to find the best predictive performance. This appendix contains a summary of the decisions tree method for comparison to the logistic regression model which was ultimately utilized. The primary splits in the decision tree were based on the weather cluster, the number of splices on the line segment, and the surrounding vegetation class.

![Figure 15: General Model ROC Curve using a Decision Tree](image)

The ROC curve shown in Figure 15 performed much better than a random guess but lagged the performance of the logistic regression method. One difficulty with the decision tree method was that only a few variables were actually used in the prediction. As the depth of the decision tree was grown with more variables, the models experienced issues with over-fitting because new variables introduced noise into the prediction from the training set. This created a decision tree that fit the noise unique to the training set, but that performed poorly on the test set with a different subset of noise. By under-utilizing all the variables in the dataset, this problem was mitigated, but these results lagged the logistic regression method.
Bibliography


