RaDA Modeling Algorithms and Methodologies Version 1 Documentation

Prepared By: PG&E Risk and Data Analytics Team

Rev	Date	Comment
0	08/27/2024	Initial Release
1	03/12/2025	Added new cover page

Table of Contents

Executive Summary

The Risk and Data Analytics (RaDA) team produces a variety of event probability, consequence, and risk models in support of risk mitigation work prioritization and planning. As these models have evolved, the RaDA team has adopted several core algorithms and methodologies for model development and delivery. This document provides an overview of the core RaDA model development tools.

Modeling Algorithms

The RaDA team explores a variety of machine learning algorithms during event model development to determine which algorithm best suits a particular model's objectives. An overview is provided for the three model algorithms currently used by the RaDA team for production event probability models:

- Maximum Entropy (MaxEnt)
- Extreme Gradient Boosting Classification (XGBoost)
- Logistic Regression

Model Performance Metrics

Several performance and feature metrics are considered during model development to determine when predictive model performance is optimal. Overview descriptions and guidance on interpretation of results are provided for the most commonly used metrics.

Model Results Methodologies

The RaDA team probability, consequence, and risk models have been developed to support risk mitigation work prioritization and planning. Frequently, work planning teams need the RaDA model data results to be transformed to match with the particular mitigation program's planning requirements. The RaDA team has developed the following model result transform methodologies in support of mitigation program work planning requirements:

- System Hardening Circuit Segments
- Model Results Aggregation
- Model Results Compositing

Future Plans

New algorithms and methodologies are developed as needed for new model development and to fulfill new user requirements. In response to user requests, there are several new methodologies planned for the aggregation of model results to Isolation Zone, Support Structure, and Regional views.

1 Introduction

The Risk and Data Analytics (RaDA) team produces a variety of event probability, consequence, and risk models in support of risk mitigation work prioritization and planning. The RaDA team has adopted several core algorithms and methodologies for model development and delivery. This document provides an overview of the core RaDA model development tools, including:

- Modeling Algorithms used to match model objectives with machine learning capabilities.
- Model Performance Metrics used to compare and improve model performance.
- Model Results Methodologies for delivering model outputs that support user needs.

1.1 Document Suite

This document is part of the documentation suite for:

- WDRM v4 Documentation
- Wildfire Fire Consequence Model v4 Documentation
- Distribution Network Event Probability Models v4 Documentation

This document will be updated to support additional RaDA team products as they are released, including the Reliability and Public Safety Risk Models currently under development.

This document presents the shared algorithms and methodologies used for development of event probability, consequence, and risk models produced by the RaDA team.

This document provides the lay reader with a broad understanding of the applicability and use of the described algorithms and methodologies. The document does not provide comprehensive detailed mathematical or scientific descriptions.

2 Modeling Algorithms

The RaDA team explores a variety of machine learning algorithms during event model development to determine which algorithm best suits a particular model's objectives. Currently, three model algorithms are used by the RaDA team for production event probability models:

- MaxEnt
- Extreme Gradient Boosting Classification (XGBoost)
- Logistic Regression

The following sections provide overviews of each of the three modeling algorithms and their applicability to RaDA event probability models.

2.1 MaxEnt

Many event probability models use a maximum entropy algorithm, a modeling technique developed by computer scientists at AT&T research and Princeton University for the modeling of ecological prevalence of species. RaDA team event probability models using the maximum entropy approach have been implemented using MaxEnt software.

2.1.1 Overview

MaxEnt probability models are trained on outage locations and their associated environmental and asset attribute data organized onto a set of 100m by 100m pixels that cover PG&E electrical grid. In MaxEnt terminology, event locations are *presence* observations, and the entire set of electrical grid pixels are the model *domain*. To accrue enough data to identify spatial patterns in a set of events, training data typically spans several years. Although it is possible to filter training data by coarse temporal criteria, the typical MaxEnt model is dedicated to spatial, not temporal, patterns. MaxEnt models produce raw relative values that correspond with the relative number of likely grid events by pixel location. Given known event rates, the raw relative event values can be calibrated through a logistic mapping into pixel-based event probabilities.

2.1.2 Constructing MaxEnt Models

MaxEnt delivers an approach for estimating the relative event occurrence rate given a modest number of known event locations. MaxEnt fits a statistical distribution of covariate values for event locations that is consistent with the values at known event locations, but otherwise as similar as possible to the distribution of values found among all modeled locations. The similarity criteria are enforced using a metric called the *relative information entropy* between the distributions of covariate values associated with event locations and all modeled locations. The larger relative information entropy, the more similar are the two distributions. Hence, the overall approach is referred to as a *maximum entropy* estimation of the relative event occurrence rate.

For the distribution event probability models the goal is to use MaxEnt to model the occurrence of ignition and failure/outage events using distribution asset attributes and environmental conditions as model covariates.

The MaxEnt software normalizes its input and target output variables, and therefore produces a normalized output where all predicted probabilities for the domain of a model sum to a value of 1. A scaling parameter is used to calibrate the sum of model domain outputs to match historical fire season event counts.

Maximum entropy as an optimization target has been broadly applied to many types of optimization problems. Therefore, online searches for key words such as “maximum entropy models” may yield results that are not related to the approach implemented by MaxEnt.

2.1.3 Applicability to the Event Probability Models

An important distinction between MaxEnt and many other Machine Learning classification methods is that MaxEnt is a presence-only modeling tool. A presence-only modeling approach trains a model using the locations of known events without requiring ground truth absence data – identification of locations where events cannot occur. Instead, presence-only modeling like MaxEnt compares known event location attributes with the attributes at all other locations.

At first glance, this might seem like an odd choice – whether there is an event or not on the grid is a clear criterion – however, there are several aspects of the wildfire modeling problem that benefit from the choice.

The following aspects of MaxEnt make it useful for spatially oriented contact from object event probability models:

- Contact from object events are typically environmentally driven and rely on spatial data to understand causal paths
- MaxEnt is open-source software with a robust community of users and peer review of its applications
- The MaxEnt software includes support for:
  - Spatial data sets and outputs
  - Feature generation from covariates such as interactions, hinges, thresholds, etc.
  - Regularization that adaptively down weights features that lack or degrade predictive power
  - Support to produce train and test performance metrics

2.1.4 Implementation

The MaxEnt software used for event probability model development is sourced from the Stanford Center for Conservation Biology open source library.

2.1.4.1 Model Data Preparation

There are three types of data that need to be acquired to build an event probability model using MaxEnt:

- Asset attributes for geometry and classifying characteristics such as material, size, age, etc.
- Known event locations used for training, testing, and validating models
- Covariate Values that change over time and space

Some covariates, such as weather and other time-based variables, require derivation, as the data should represent the conditions under which known events have occurred. To support multi-year event datasets, the derivation typically requires statistical extracts of multi-year sequences of weather data. Typical derivations for time-based data such as wind speed and fuel moisture include average values, percentile values, counts of days beyond threshold limits, etc.

All collected data is assigned by its location onto a common set of 100m square pixels that overlay the geometry of the distribution network grid. The grid pixel values are stored into raster datasets for MaxEnt modelling.

2.1.4.2 Iterative Model Process

The process for building an event probability model using MaxEnt is similar to that used for most machine learning (ML) technologies. MaxEnt models are built through an exploratory process managed by a data scientist, iteratively configuring and evaluating candidate models.

A MaxEnt model run configuration specifies options like:

- Source pixel raster datasets
- Attributes and covariates available for modeling events
- Event training, testing, and validation datasets
- Options for model fitting techniques and output scaling
- Options for model evaluation

Once a candidate model has been built from a configuration, MaxEnt produces a variety of reports to assist with model performance evaluation, including:

- Model output evaluation
- Predicted event probabilities
- Model performance metrics for train and test event data
- Model fit performance
- Jackknife charts
- Dropped known events due to insufficient attribute or covariate values

2.2 Extreme Gradient Boosting Classification (XGBoost)

Most distribution equipment asset event probability models use XGBOOST, a boosted decision tree modeling algorithm.

2.2.1 Overview

XGBoost is a powerful machine learning algorithm used for predictive modeling. XGBoost combines multiple weak models, typically decision trees, to create a combined, robust predictive model. The combined model is created using gradient boosting, which iteratively creates new combined model configurations to correct prediction errors produced by previous models.

Key features of XGBoost include:

- Handling of missing values – a common issue for real world data sets.
- Fast and efficient execution on large data sets
- Often provides superior performance against alternative algorithms.

2.2.2 Applicability to the Event Probability Models

The distribution equipment asset event probability models are classification problems and therefore best solved using a classifier modeling algorithm. Using a decision tree-based classifier algorithm provides model feature interpretability through which feature causality can be confirmed by Subject Matter Experts (SMEs). While a single tree model might work well for small data sets, tree-based classifier algorithms tend to overfit larger training data sets like those used for the event probability models, reducing predictive power against novel data sets. As a result, combinations of multiple tree models are typically used for large data sets.

There are two common approaches for combining tree models, bagging and boosting. Bagging algorithms train multiple decision tree models in parallel, while boosting trains models serially. Bagging, or parallel modeling, does not resolve the overfit issues experienced with single tree models. Boosting, which includes a process for pruning out specific tree models, is typically able to minimize model overfit by iteratively training to the residual error produced by its prior tree.

Another advantage for model boosting is that it produces better performance on imbalanced training data sets with very few actual events in comparison to the number of non-events.

There are multiple boosting algorithms available, such as:

- Adaptive boosting (Adaboost)
- Gradient Boosting (GB)
- XGBoost

Adaboost is a basic boosting algorithm and use decision stumps, which are decision trees based on a single feature, to compare performance to determine relative feature importance. Gradient Boosting works similarly to AdaBoost but uses fully trained decision trees rather than stumps. XGBoost follows the same algorithm as Gradient Boosting but uses advanced regularization techniques to suppress weights, prevent overfitting, and enhance its performance against novel data set. Through RaDA team model development research, XGBoost has been selected for building the distribution event probability models.

2.2.3 Implementation

As of document publication, XGBoost 1.7.6 is used for model development.

2.2.3.1 Model Data Preparation

There are three types of data that need to be acquired to build an event probability model using XGBoost:

- Asset attributes for geometry and classifying characteristics such as material, size, age, etc.
- Known event locations used for training, testing, and validating models
- Covariate Values that change over time and space.

Some covariates, such as weather and other time-based variables, require derivation, as the data should represent the conditions under which known events have occurred. To support multi-year event datasets, the derivation typically requires statistical extracts of multi-year sequences of weather data. Typical derivations for time-based data such as wind speed and fuel moisture include average values, percentile values, counts of days beyond threshold limits, etc.

Spatial data, like weather data, is assigned to an asset for XGBoost from a common set of 100m square pixels that overlay the geometry of the distribution network grid.

2.2.3.2 Iterative Model Process

The process for building an event probability model using XGBoost is similar to that used for most ML technologies. XGBoost models are built through an exploratory process managed by a data scientist, iteratively configuring and evaluating candidate models.

A XGBoost model run configuration specifies options like:

- Source datasets
- Attributes and covariates available for modeling events
- Event training, testing, and validation datasets
- Options for model hyperparameters to control regularization and prevent model overfit

Once a candidate model has been built for a configuration, its performance is evaluated using:

- Predicted event probabilities
- Model performance metrics for train and test event data
- Model feature importance
- SHAP plots

2.3 Logistic Regression

2.3.1 Overview

A logistic regression model predicts the probability of discrete outcomes from a set of covariate inputs. A discrete outcome can be any set of ordinal values, such as:

- Yes/No
- True/False
- High/Medium/Low
- Ignition/No Ignition

For any given set of inputs, a logistical regression model will assign a probability for each possible discrete outcome.

Logistic regression models are structured so that the log of the odds of each outcome is a linear combination of the covariate data, which means that once the “log-odds” transform is applied, the model is fit as a standard multiple linear regression. The probability that a binary variable y has a specific outcome given a matrix of all covariate values associated with all observations the binary variable is a logistic function over all covariates with the form:

$$p(y = 1|X) = \frac{1}{1 + e^{-\beta_0 + \sum_{i=1}^{m} \beta_i x_i}}$$

Where:

Variable	Description
P(y=1|X)	Probability that an input event is associated with a modeled outcome
β_i	Covariate correlation coefficients
B_0	Covariate correlation coefficient mid-point
i	Covariate index
m	Number of covariates
X	Covariates matrix
x_i	Covariate vector of values

Logistical regression optimizes a set of logistical functions for each covariate. A generic 1-dimensional logistic function is presented in the *Figure 1*. The y-axis is the probability of the desired binary outcome (True, Yes, 1, etc.) and the x-axis represents the numerical values of one of the input covariates. In this example covariate values below -2 are associated with very low probability outcomes and values above 2 are associated with very high probability outcomes. The function assigns probability value of 0.5 (50/50 odds) for the desired outcome at the midpoint of the curve located at a covariate value of 0.

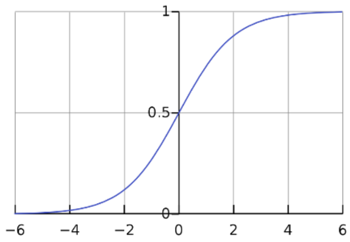


Figure 1 - 1-Dimensional Logistic Function

During the optimization of the logistical regression model, the location of the mid-point 0.5 value along the x-axis can also shifted to the left or right, formulaically expressed as $-\beta_0/\beta_i$. Likewise, the transition steepness of the y-axis from 0.0 to 1.0 can be rescaled during optimization, formulaically expressed as $1/\beta_i$.

2.3.2 Applicability to the Event Probability Models

The RaDA team uses logistical regression modeling to develop the Probability of Ignition given Outage Model, p(i|o), to predict the binary probability of whether an ignition event has occurred as the result of an electrical equipment asset failure or outage event. The model is trained with a data set of historical outage events for which includes a binary indicator of whether an ignition event also occurred. The outage data is combined covariate inputs for weather, fuel, and vegetation conditions as well as the type of failure that initiated the outage to construct the predictive model for probability of ignition given an outage.

2.3.3 Implementation

The DEPM v4 Probability of Ignition given Outage Model was developed using the scikit-learn machine learning framework, written in Python. The details of the modeling process, the model fit, and its predictive performance are discussed in Distribution Network Event Probability Models v4 Documentation.

3 Model Performance Metrics

Several performance and feature metrics are used when evaluating risk and event probability models as fit for purpose. The following sections provide an overview of the most commonly used model performance and feature metrics considered by the RaDA team.

3.1 Model Performance Metrics

Model performance metrics are used to gain insight into how a model performs on data its algorithm hasn't been exposed to during model development. Typically, an out-of-sample test data set is withheld from the data gathered for model training. After fitting a candidate model using a machine learning algorithm, the candidate is exercised to generate model predictions for the test data set. The candidate model predictions are compared against the actual test data set outcomes to produce performance metrics. The performance metrics commonly used by the RaDA team metrics used for evaluating machine learning models are:

- Receiver Operating Characteristic Area Under the Curve (ROC AUC)
- Precision-Recall Curves and Average Precision
- Top 20% Concentration Factor

The distribution event modeling framework produces probability values. The metrics described in the following section support the review of model performance across a full range of probability thresholds for determining event outcomes without the need to compute performance results at a specific probability threshold.

A Confusion Matrix, as shown in *Figure 2*, is used to understand model performance at a particular probability threshold. The Confusion Matrix compares the probabilities predicted by the model to the actual real-world outcome on an out-of-sample test data set. The probability decision threshold is used to determine whether a model probability value indicates a predicted event has occurred or not. Each prediction/test set event pair will fit into one of four Confusion Matrix outcome categories:

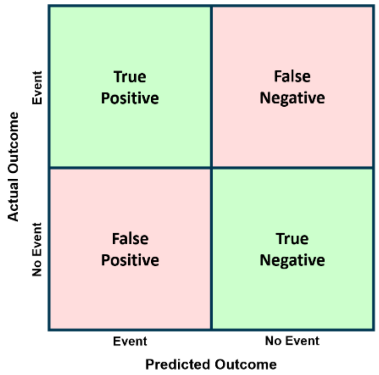


Figure 2 - Confusion Matrix

Table 1 - Confusion Matrix Outcome Categories

True Positive	Event predicted for an actual event
True Negative	No event predicted when there was no actual event
False Positive	Event predicted when no actual event occurred
False Negative	No event predicted when there was an actual event

The Confusion Matrix definitions are the foundation for understanding the model performance metrics described in the following sections.

3.1.1 Receiver Operating Characteristic Area Under the Curve (ROC AUC)

The Receiver Operating Characteristic Area Under the Curve (ROC AUC) measures the tradeoff between capturing True Positives and misclassifying False Positives across the range of thresholds for the predicted probabilities.

Figure 3 shows a sample ROC Curve showing the relationship between the True Positive Rate, also known as recall, and the False Positive Rate, for a given probability decision threshold. For a binary classification such as event / no event, rates are calculated as:

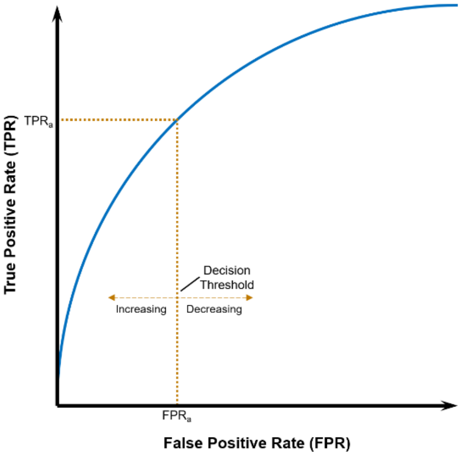


Figure 3 - Example ROC Curve

$$TPR_a = {}^{TP_a}\!/_{(TP_a + FN_a)}$$

$$FPR_a = {}^{FP_a}\!/_{(FP_a + TN_a)}$$

where:

a	Probability decision threshold
FN	False Negative – predicted no event where event occurred
FP	False Positive – predicted event where no event occurred
FPR	False Positive Rate
TN	True Negative – predicted no event where no event occurred
TP	True Positive – predicted event where event occurred
TPR	True Positive Rate, also know as Recall

Notice that the highest probability threshold occurs where the x- and y-axis meet. This portion of the chart shows the TPR and FPR for the highest probability values since only the highest probability values exceed that threshold.

Figure 4 shows the relationship between ROC and AUC. A ROC that renders as straight diagonal line has an area under the curve, or AUC, of 0.5 and indicates that a model algorithm cannot differentiate between events and non-events across all probability thresholds. In other words, the model performance is no better than random guessing. An ROC curve above the diagonal indicates that a model performs better than random guessing. The steeper the vertical start of the and the flatter the horizontal end of the ROC curve, the higher the AUC and the better a model is at predicting true positives while minimizing false positives for the highest probability thresholds.

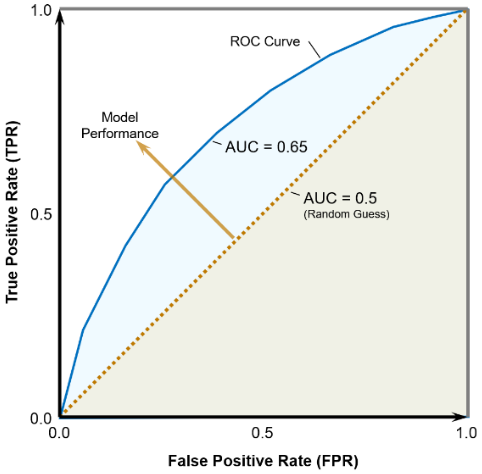


Figure 4 - Example ROC AUC

3.1.2 Precision-Recall Curves and Average Precision

The training data sets for equipment asset event probability models are highly imbalanced, meaning that there are very few event occurrences against a very large number of non-events. Such an imbalance will often lead to an ROC curve that gives too much credit for True Negatives, where an event did not occur and the proposed algorithm predicts no event. True Negatives represent the overwhelming majority of predictions for the event probability models, so it is important to consider the underrepresentation of False Negatives where an event occurred and the proposed algorithm predicts no event. Precision-Recall curves can provide insight into missed event underrepresentation.

Precision-Recall curves measure the tradeoff between converting False Negatives into True Positives at the expense of increasing False Positives. Precision and Recall are calculated across the same range of probability thresholds used to produce the ROC curve (*Figure 3*).

$$Precision_a = {}^{TP_a}/_{(TP_a + FP_a)}$$

$$Recall_a = {}^{TP_a}/_{(TP_a + FN_a)}$$

where:

a	Probability decision threshold
FN	False Negative – predicted no event where event occurred
FP	False Positive – predicted event where no event occurred
TN	True Negative – predicted no event where no event occurred
TP	True Positive – predicted event where event occurred
Precision	Fraction of True Positives against positive predictions, true or false
Recall	Fraction of True Positives against all positives, also known as True Positive Rate

Figure 5 presents a typical Precision-Recall curve. In this chart, to achieve 20% Recall (also known as True Positive Rate, Precision is reduced to 3.8%. Note that as Precision increases to account for more True Positive results, the Recall declines as the number of False Positive results increases. Therefore, a model that produces the steepest initial downward slope of preferable.

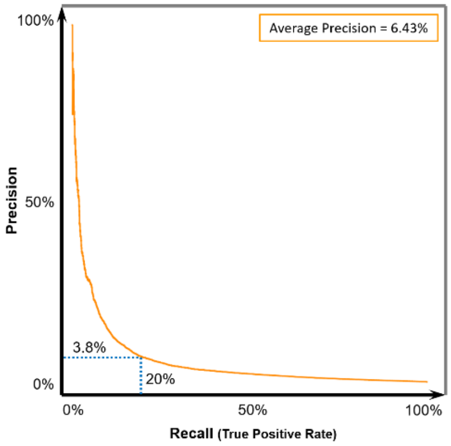


Figure 5 - Example Precision-Recall Curve

Average precision summarizes the precision-recall curve by calculating the weighted average precision across the entire range of probability prediction thresholds. The weight is the increase in recall from the previous threshold. This value is not interpolated or calculated using the trapezoidal rule with the Area Under the Precision Recall Curve, which can be overly optimistic. It is computed using the Average Precision Score method from scikit-learn, a Python machine-learning library.

3.1.3 20% Concentration Factor

Ultimately, the Distribution Event Probability Models, in conjunction with the Wildfire Consequence Model, are used to prioritize future wildfire mitigation work. Wildfire mitigation work planning tends to consider only the riskiest top 20% of all potential locations on the distribution grid. Therefore, a useful metric for understanding how a model will influence mitigation spend efficiency is to consider a 20% Concentration Factor, which measures how many actual distribution event locations are found in the top 20% of modeled locations when ranked in descending order of predicted event probabilities.

The Concentration Factor is calculated from:

$$Concentration\ Factor = \frac{Events_{Actual}/Events_{Total}}{Locations_{Ranked}/Locations_{Total}}$$

where:

20% Conc. Factor	Model concentration factor for top 20% of ranked locations.
$\text{Events}_{\text{Actual}}$	Number of events that occurred within the ranked list of locations
$\text{Events}_{\text{Total}}$	Number of total actual events
$\text{Locations}_{\text{Ranked}}$	Number of modeled locations considered in order of descending predicted event probability
$\text{Locations}_{\text{Total}}$	Total number of modeled locations

The 20% Concentration Factor provides insight into the effectiveness of mitigating risk for the highest priority locations or assets. The higher the concentration factor, the more likely that addressing the highest priority locations will truly address locations of significant risk. *Figure 6* provides an example where the top 20% of asset locations by ranked probability include the locations of 71% of all historical events, which equates to a very good concentration factor of 3.55. Therefore, it is likely that mitigation work to the top ranked locations will be very effective in reducing future events. On the other hand, a poor concentration factor closer to 1 would indicate that a model is only slightly better than a coin toss, and that other factors beyond model-ranked locations should be considered before performing mitigation work. Additionally, a model with a poor concentration factor is an obvious candidate for further model development.

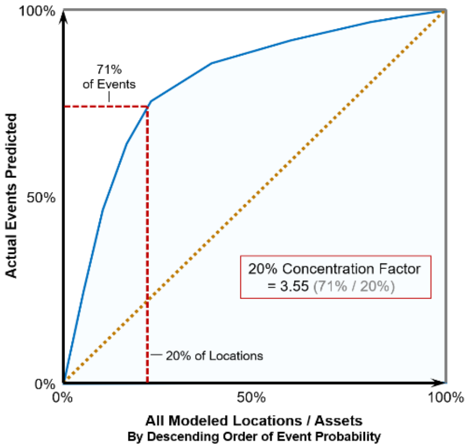


Figure 6 - Example 20% Concentration Factor

3.2 Feature Metrics

Features metrics provide insight into how helpful and influential the inputs are to a model algorithm. The metrics described in the following sections are used to help explain how a machine learning model is influenced by its covariates and hence why it predicts a particular set of probability results.

The most common feature metrics commonly used the RaDA team for evaluating machine learning models are:

- Feature Importance
- SHAP Plots

3.2.1 Feature Importance

A popular metric for evaluating a proposed machine learning model is to examine the relative importance of each input feature to the model output. Two types of feature importance are used for event probability development: permutation importance, use for MaxEnt based models, and total gain feature importance, used for XGBoost based equipment asset models.

Permutation importance for MaxEnt models is determined by sequentially replacing the historical data for each feature with randomly scrambled data and observing the subsequent difference in model performance.

Total gain feature importance is a part of the XGBoost algorithm library. The metric represents the cumulative performance gain across all the tree splits where that feature is used.

Figure 7 shows a typical feature importance chart for a model developed using eight features. The chart displays the features in order of importance to a model from most to least important. Additionally, the length of the bar for a feature represents its relative importance for the model when compared with the other features.

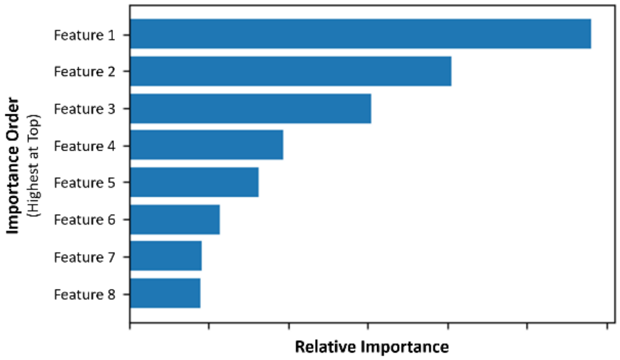


Figure 7 - Example Feature Importance Chart

3.2.2 SHAP Plots

The SHAP plot provides insight into the contribution of each feature to the performance of a model. The plot is based on individual data points in contrary to the total gains feature importance typically used for model feature evaluation. The SHAP plot provides local interpretability by disaggregating by individual data points and shows the contributions of each feature across different observations. In addition, the plot provides intuition about whether a feature has a generally positive or negative correlation with predict values.

SHAP values originate from cooperative game theory. The methodology aims to allocate optimal credit to each feature (emphasis on "cooperative") and make explanations local (so each test case/row can be explained). Not only does each test case have its own SHAP values, but the values across test cases are additive allowing for both individual and group evaluation of the test dataset.

An example of a SHAP plot is shown in *Figure 8*. The SHAP Beeswarm plot uses the local and additive nature of the SHAP values to provide three dimensions of insight to model inputs:

- The plot y-axis is similar to other feature importance bar charts and ranks features from the most important to least important. The feature order is selected according to the mean absolute value of the SHAP values for each feature. This order emphasizes broad impact and can undervalue rare but high magnitude impacts.
- The x-axis shows each SHAP value with a dot for each data set row. The dots aggregate along each feature row to show density, indicating when many rows have similar SHAP values, providing insight to a features SHAP values distribution.
- Color is used to display the relative original value of a feature, with higher values in red and lower values in blue. The color dimension gives insight into whether high or low values are associated with higher probability outcomes, providing a useful check on whether the model algorithm is using feature values as expected.

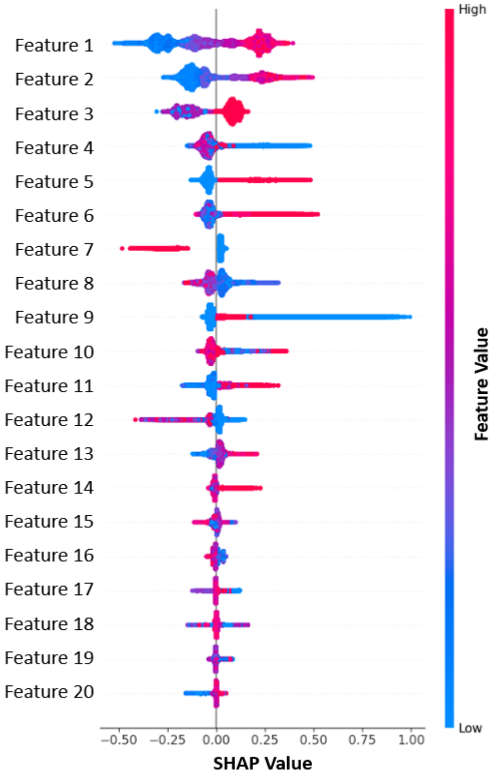


Figure 8 - Example SHAP Plot

4 Model Results Methodologies

The RaDA team probability, consequence, and risk models have been developed to support risk mitigation work prioritization and planning. Frequently, work planning teams need the RaDA model data results to be transformed to match with a particular mitigation program's planning requirements. The RaDA team has developed the following model result transform methodologies in support of mitigation program requirements:

- System Hardening Circuit Segments
- Model Results Aggregation
- Model Results Compositing

The following sections provide overviews of each of the three transform methodologies and their application to RaDA probability, consequence, and risk model results.

4.1 System Hardening Circuit Segments

Circuit Segments are an artificial risk mitigation work planning construct applied to the distribution grid.

A circuit on the distribution grid is the set of electrical grid assets downstream of a substation circuit breaker. However, an electrical circuit is often too large for planning and performing risk mitigation work. Therefore, the System Hardening/Undergrounding wildfire mitigation planning teams have chosen to organize their work by Circuit Segments.

4.1.1 Circuit Segment Definition

A Circuit Segment is the section, or segment, of a circuit and all its connected assets downstream of its closest recloser, or Dynamic Protective Device (DPD). Multiple reclosers on a circuit divide the circuit into various smaller segments such that a fault within any segment will only disrupt power to itself and any downstream segments. An example circuit broken into four Circuit Segments by three DPDs is shown in *Figure 9*.

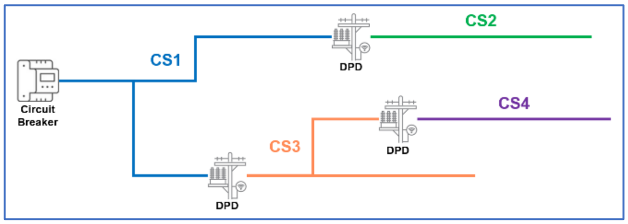


Figure 9 - Circuit Segmentation for Work Planning

There is one significant nuance when defining a work planning Circuit Segment. Some DPDs have switching capabilities that can alter the effective segment configuration during operation. For planning purposes, Circuit Segment configuration is set to match default, or as designed, switch positioning.

4.1.2 Circuit Segment Identification

Circuit Segments are an artificial work planning construct. Each Circuit Segment is identified through logical inspection of the Electrical Distribution GIS (EDGIS) datasets.

4.1.2.1 Circuit Segment Configuration Source Data

The work planning Circuit Segments are determined using grid asset data from three sources:

- Circuit Breaker EDGIS
- DPD EDGIS
- EDGIS Circuit Trace Table

The electrical grid is constantly changing due to maintenance, risk mitigation, and newly installed services. Therefore, it is necessary to use data that is sourced at a specific point in time to minimize configuration mismatches. Snapshot of all three datasets for defining the Circuit Segments are archived by RaDA on the first of each month along with other source datasets necessary for developing risk models. When RaDA begins development on a new set or risk models, all source datasets are synchronized to a common snapshot of data.

4.1.2.2 Protective Device List

As Circuit Segments are an artificial work planning construct, it is necessary to determine from available GIS datasets which electrical assets serve as segment defining protective devices. For System Hardening, this includes Circuit Breakers and DPDs that meet the following criteria:

- Circuit Breakers
  - Substation full circuit breakers
  - Default position configuration, if applicable
  - Parameter: subtypecd = “Source”
  - Parameter: enabled = “True”
- DPDs
  - Default position configuration
  - Parameter: subtypecd = “Recloser”
  - Parameter: status = “In Service”

For Circuit Breakers and DPDs that include configuration switching functionality, additional logic is applied to include only devices with default, as designed, switch settings.

4.1.2.3 Circuit Segment Protected Asset Identification

For every protective device that serves to identify a Circuit Segment, there is a set of electrical assets that the protective device safeguards. The EDGIS Circuit Trace Table is used to cross-reference each Circuit Segment with its protected assets using the global ID of its protective circuit breaker or DPD.

The Trace Table contains millions of relationships that define the distribution grid configuration. The relationships are inspected to determine the global ID of the closest upstream protective device, and hence the containing Circuit Segment, for each electrical asset on the grid.

Unfortunately, due to clerical errors or differences in update frequencies of the EDGIS tables, a very small percentage of assets cannot be mapped to a circuit breaker or DPD in the protective device list. These leftover assets are assigned to an “Unknown” Circuit Segment so that their risk can be accounted for during the modeling process and to facilitate investigation of GIS errors.

4.1.2.4 Circuit Segment Name Assignments

Circuit Segment names are created by combining the circuit name and the protective device operating number for a segment. *Figure 10* provides several examples of Circuit Segment naming for circuit "*El Dorado PH 2101*". The following table illustrates how the circuit name and device operating numbers from *Figure 10* are combined to form unique circuit segment names:

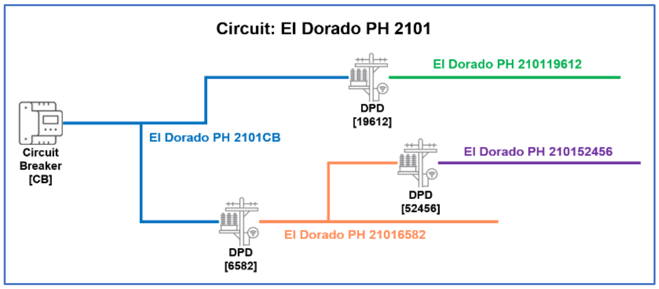


Figure 10 - Circuit Segment Naming

Circuit: El Dorado PH 2101		
Operating Number	**Device Type**	**Circuit Segment Name**
CB	Circuit Breaker	El Dorado PH 2101CB
19612	DPD	El Dorado PH 210119612
6582	DPD	El Dorado PH 21016582
52456	DPD	El Dorado PH 210152456

Table 2 - Example Circuit Segment Names

It is important to remember that Circuit Segment names reflect a fixed point in time. The grid is continuously evolving for reasons such as adding new services, inserting new protective devices for PSPS or EPSS, undergrounding portions of circuits, and replacing or removing circuit sections. Each time the grid configuration changes, one or more Circuit Segments may see a change in name, length, or protected assets. This is the primary reason why the RaDA risk modelling work is always anchored to a set of EDGIS datasets captured at a common date to ensure that mapping of risk results is consistent.

4.1.2.5 Circuit Segment Geometries

Multiple user requirements for the RaDA risk model results are satisfied through analysis of the Circuit Segment geometries. Circuit Segment shape geometries are determined via the cross-referenced protected primary and secondary conductor records from the EDGIS Circuit Trace Table. The geometry of each Circuit Segment protected conductor section is merged to create the total geometry.

The Circuit Segment geometries are required to satisfy two significant user requirements:

- Display risk model results on maps in Foundry as well as other systems such as ArcGIS.
- Determine Circuit Segment lengths in defined areas such as:
    - HFRA
    - HFTD
    - HFTD Tier 2
    - HFTD Tier 3
    - HH Zone 1
    - County
    - Region

Note that location of non-conductor assets, which have point locations rather than geometries, are not merged into the Circuit Segment shape geometry.

4.1.3 Circuit Segment Use for Risk Models

As required by the System Hardening and Undergrounding risk mitigation work planning teams, model results from the Distributions Event Probability Models, the Wildfire Consequence Model, and the Wildfire Distribution Risk Model are aggregated to Circuit Segments. Work planners consider the relative aggregated values for probability of ignition, wildfire consequence, and wildfire risk to prioritize the timing, type, and amount of risk mitigation work that will be performed on high-risk Circuit Segments, especially in high fire risk areas.

Aggregation and compositing of risk model results to Circuit Segments are described in Sections *4.2* and *4.3* of this document.

4.1.4 Circuit Segment Vintage Warning

The GIS datasets used to create a set of distribution Circuit Segments are constantly changing to reflect the current physical state of the grid. While it is possible to continuously update the Circuit Segments as the underlying GIS datasets are updated, this is not useful from a risk modeling perspective. Therefore, any risk model produced by the RaDA team will have a GIS vintage date associated with it. The GIS vintage date for a risk model is recorded as part of its provenance metadata. Users are advised to check and consider the Circuit Segment vintage date when trying to compare or merge risk model results with Circuit Segment based datasets produced outside of the RaDA team.

4.2 Model Results Aggregation

4.2.1 Introduction

RaDA produces both asset and spatial models. Asset models produce results that estimate event probabilities or risk for individual assets at point locations. Spatial, or grid pixel, models, product results that estimate event probabilities or risk within 100m by 100m square pixels that form a grid over the distribution and transmission service territories.

Many end users need to understand model results in a larger context than the direct model outputs. The most used context is Circuit Segment based values. Other contexts that have been requested include county and regional based values. Providing these values requires that the asset and spatial model results be aggregated to the desired context.

4.2.2 Circuit Segment Line Geometry Aggregation

A common user requirement is aggregating model results to a line geometry. While there are several permutations of line geometry possible for the electric grid, the only aggregation currently supported by the RaDA team is the System Hardening Circuit Segment. The aggregated model result for a Circuit Segment is the sum of two components, grid pixel model result values aggregation and asset model result values aggregation.

4.2.2.1 Grid Pixel Aggregation

Many of the RaDA risk models produce results for 100m by 100m grid pixels that overlay our service territory. Pixel model results are aggregated to line geometries like Circuit Segments by summing the model result value for each grid pixel that intersects spatially with the Circuit Segment geometry. *Figure 11* presents a single Circuit Segment that intersects with 14 grid pixels. The aggregated model result value for the Circuit Segment is the simple sum of all intersecting pixel values.

CS Aggregated Pixel Value = A5 + A6 + B1 + … + I7 + J8

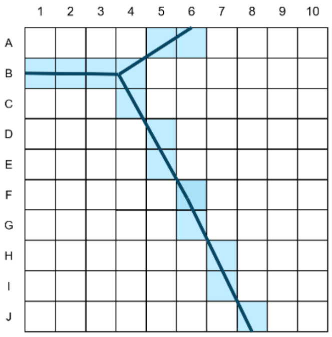


Figure 11 - Single Circuit Segment Pixel Aggregation

Aggregating model results is a bit more complicated when multiple Circuit Segments intersect with one or more shared pixels as shown in *Figure 12*. If the same model result value is summed to both Circuit Segments, then you end up with more summed total Circuit Segment result values then is modeled for the entire grid. In other words, the sum of all Circuit Segment results would be greater than the sum of all pixel results.

The model results aggregation is modified for shared pixels by dividing the model result for each shared pixel by the number of Circuit Segments that intersect it.

Blue CS Aggregated Pixel Value = A5 + … + F6/2 + … + J8

Orange CS Aggregated Pixel Value = G1 + … + F6/2 + … + E10

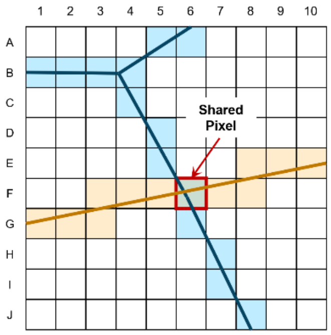


Figure 12 - Multi-Circuit Segment Pixel Aggregation

4.2.2.2 Asset Value Aggregation

Several event probability models produce by RaDA produce results by unique assets. Asset model results are aggregated to line geometries like Circuit Segments through parent-child relationships between assets kept in various system of record databases. Most assets relationships can be established using the EDGIS Circuit Trace Table. Unfortunately, support structure assets are not included in the Circuit Trace Table and their relationship to other assets must be inferred through multiple data sources.

4.2.2.2.1 EDGIS Circuit Trace Table Assets Assignment to Circuit Segments

The EDGIS Circuit Trace Table, which is used to identify and name Circuit Segments, also serves to relationally identify most electrical assets that are associated with a Circuit Segment. This includes:

- Capacitor Banks
- Dynamic Protection Devices (DPDs)
- Fuses
- Switches
- Transformers
- Voltage Regulators

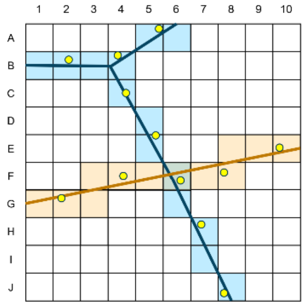


Figure 13 - Circuit Segment Assigned Assets

Unfortunately, Support Structures, which are not energized assets, are not included in the trace table.

Figure 13 presents two Circuit Segments showing both their assigned pixels and assets.

4.2.2.2.2 Support Structure Assignment to Circuit Segments

Support Structures, or poles, require specialized logic to determine their owning Circuit Segment(s). As noted in the prior section, Support Structures are not recorded in the EDGIS Circuit Trace Table as they are not energized assets. Currently, there is no comprehensive single data source that definitively relates Support Structures with Circuit Segments.

Support Structures to Circuit Segment(s) relationships are established through a cascaded search of two datasets in the following order of preference:

1. RaDA's Manual Assignment Dataset
2. Asset Knowledge Management (AKM) Pole to Conductor Dataset

4.2.2.2.2.1 RaDA Manual Assignments

There is a known issue with the AKM Pole to Conductor dataset for assets near electrical stitch points and circuit breakers where there are sometimes too many conductors, and by association circuit segments, assigned to a single support structure. For cases where there are four or more circuit segments linked to a support structure, the RaDA team manually use the EDGIS Web Viewer to review conductor connections and create a custom support structure to circuit segment lookup dataset.

Figure 14 provides an example of the potential need for a manual assignment. The magenta dots represent support structures where there is the potential for as many as four Circuit Segment that could be assigned to each of the poles. The RaDA Manual Assignment Dataset would be used to determine the Circuit Segment and Support Structure relationships.

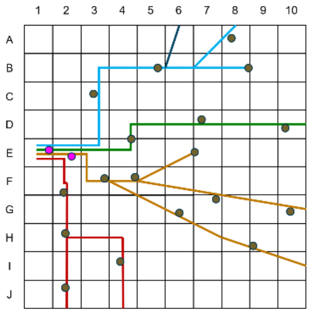


Figure 14 - Circuit Segment Support Structure Assignments

4.2.2.2.2.2 Asset Knowledge Management (AKM) Pole to Conductor Dataset

The AKM team maintains a dataset that is the most reliable source for understanding relationships link between distribution support structures and conductors. The AKM team actively works to maintain and enhance the dataset. The AKM dataset is used for establishing nearly all of the conductor to support structure relationships with the few exceptions originating from the RaDA Manual Assignments dataset described in the prior section. The AKM dataset is snapshotted monthly to support synchronization of model data for development.

4.2.2.2.3 Asset Model Result Aggregation to Support Structures

The following assets have one-to-one relationships with Circuit Segments:

- Capacitor Banks
- Dynamic Protection Devices (DPDs)
- Fuses
- Switches
- Transformers
- Voltage Regulators

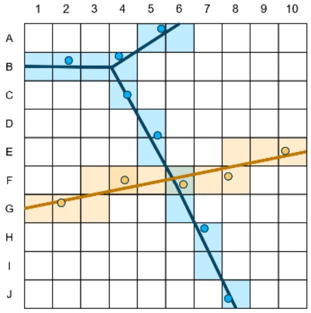


Figure 15 - Direct Asset Risk Aggregation to Circuit Segment

The risk results for these assets list above can be directly attributed to a Circuit Segment. *Figure 15* shows two circuit segments and their directly associated assets for aggregation.

4.2.2.2.4 Support Structure Result Aggregation to Circuit Segments

Support Structure and Primary Conductor risk results are attributed to specific poles. Many poles are associated with multiple Circuit Segment. Pole-based risk results are therefore apportioned equally to associated Circuit Segment depending on the number of connected Circuit Segments. *Figure 16* presents three Circuit Segments with a few shared poles. Their risk results would be assigned and summed as follows:

- Blue CS = 1/3 Red + 1/2 Magenta + 3 Blue
- Green CS = 1/3 Red + 1/2 Magenta + 2 Green
- Brown CS = 1/3 Red + 3 Brown

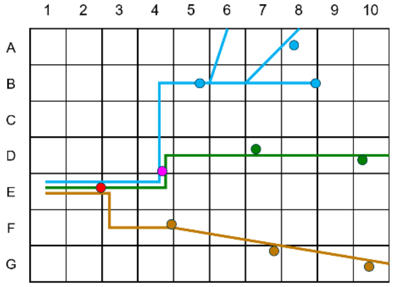


Figure 16 - Pole-based Asset Risk Aggregation to Circuit Segment

4.3 Model Results Compositing

The ultimate purpose for the RaDA risk models is to inform the prioritization of risk mitigation programs. The event probability model risk results can be flexibly composited to provide probability values, risk values, and priority rankings for specific mitigation programs. Using composited results, programs can prioritize mitigation of the highest total risks while using the contributing event probability models to understand the best mediation approach to handle the specific components of risk.

Risk can be composited for any combination of event probability models. Mitigation planners and Subject Matter Experts can focus on the drivers of risk for which they are responsible with confidence that their composited view is relevant to their work planning needs.

4.3.1 Compositing Methodology

An event probability model produces, by asset or pixel, a probability of ignition. Combining a probability of ignition with its consequence produces the wildfire risk. Probability of ignition and risk results can be composited to create total probability of ignition and total risk values. Compositing methodology has evolved as the distribution event probability models have matured and improved.

In producing the WDRM v3, all event probability model results were composited on a pixel basis and equipment asset results were spatially assigned to circuit segments. Unfortunately, a significant number of pixels contain multiple circuit segments. As risk results were attributed to a pixel, v3 compositing lacked a methodology for attributing asset risk at a pixel level proportionally to a specific circuit segment. Therefore, pixel risk would be divided equally between all circuit segments that crossed through a pixel.

For WDRM v4, most of the equipment asset models produce results on an individual asset basis and each asset's relationship with a containing circuit segment is traced through various GIS and SAP data sets. Therefore, equipment asset risk can be attributed directly to a circuit segment, eliminating the shared risk approach necessary for pixel-based results using for v3.

The compositing methodology used by the RaDA team in support of the WDRM is currently a three-step process:

1. Composite pixel model results
2. Composite equipment asset model results
3. Aggregation of pixel and asset composite values

4.3.1.1 Compositing Pixel Model Results

Models that produce pixel results are typically spatially oriented in the context of the risk than threatens the electrical grid network. *Figure 17* depicts a single pixel with multiple potential spatial threats. Individual event probability models for vegetation, animals, and third-party events have produced high, medium, and low risk results, respectively for this particular pixel. Mitigation programs are typically interested in comparing the total risk for the circuit segment that passes through this pixel relative to other circuit segments where work might be performed.

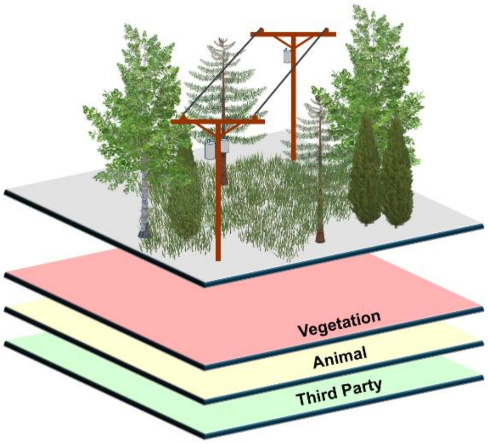


Figure 17 - Pixel Model Layers

Pixel model results are very straightforward to composite, this risk values from each of the contributing pixel models are simply summed to determine the composite risk for a pixel:

$$Composite\ Risk_{pixel} = \sum_{1}^{n} Pixel\ Model\ Risk$$

Figure 18 presents the composited risk value for the example pixel. For this example, a high vegetation risk result and a significant animal risk combine to produce a relatively high overall pixel model risk result for any circuit segment that passes through the pixel.

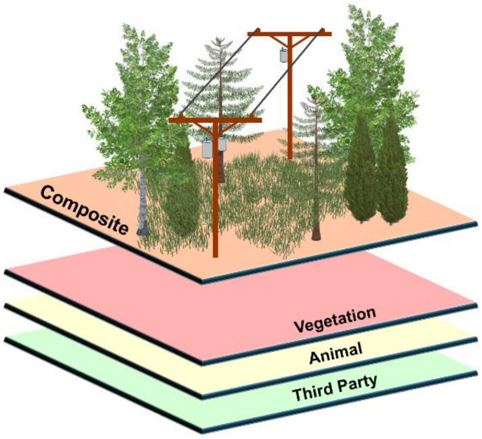


Figure 18 - Composited Pixel Result

4.3.1.2 Compositing Equipment Asset Model Results

Equipment asset models produce risk results that are specific to individual assets. *Figure 19* depicts a very simple example of a circuit segment. Even simplified, there are multiple modeled assets represented, including:

- Two support structures
- Two primary conductor spans
- Three attached transformers

Risk results are generated by the appropriate equipment asset model for each piece of equipment. *Figure 20* displays visually the risk results for each of the individual assets that make up the simple circuit segment. Note that the two support structures, the two conductor spans, and three transformers each have different levels of risk assigned to them.

The overwhelming majority of equipment assets are pole based. Therefore, equipment assets are composited to the support structure that holds the asset. Conductor spans are a special case in that they are supported by two poles, and hence, their risk must be distributed equally to their support structures.

Figure 21 depicts the compositing of equipment asset risk to the containing support structures. The composite risk indicator at the base of each support structure combines the risk for the pole, the equipment attached to the pole, and half of the conductor risk. Note that any pole with a significant number of attached equipment assets is likely to have a relatively high composite risk simply due to the number of assets.

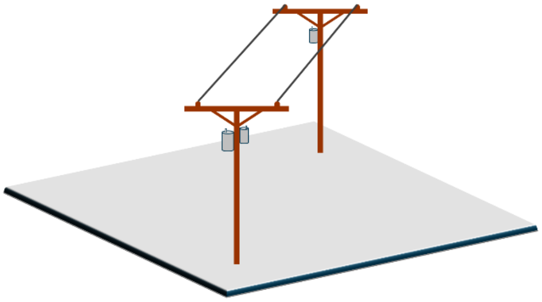


Figure 19 - Simplified Equipment Asset Example

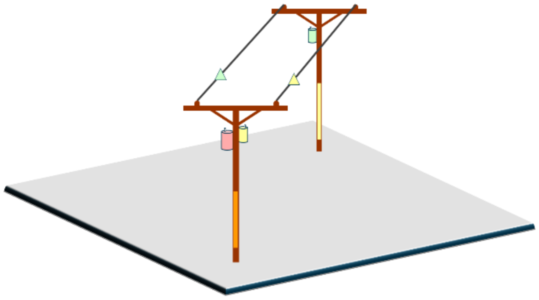


Figure 20 - Individual Equipment Asset Risk Results

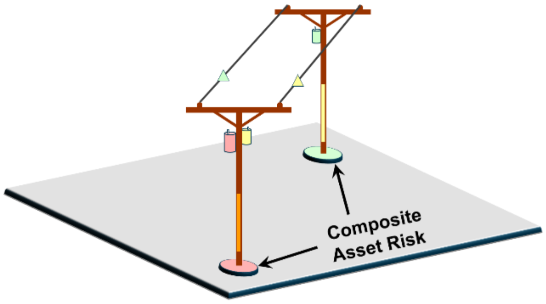


Figure 21 - Composited Asset Risk Example

4.3.1.3 Aggregation of Pixel and Asset Composite Results

Composite pixel and equipment asset results are typically aggregated to the context of circuit segment risk values for mitigation work planning.

Figure 22 depicts a single circuit segment that spans multiple grid pixel areas and has several equipment assets. The aggregated risk value for this segment is the combined sum of three composite pixel risk values and four composite asset risk values.

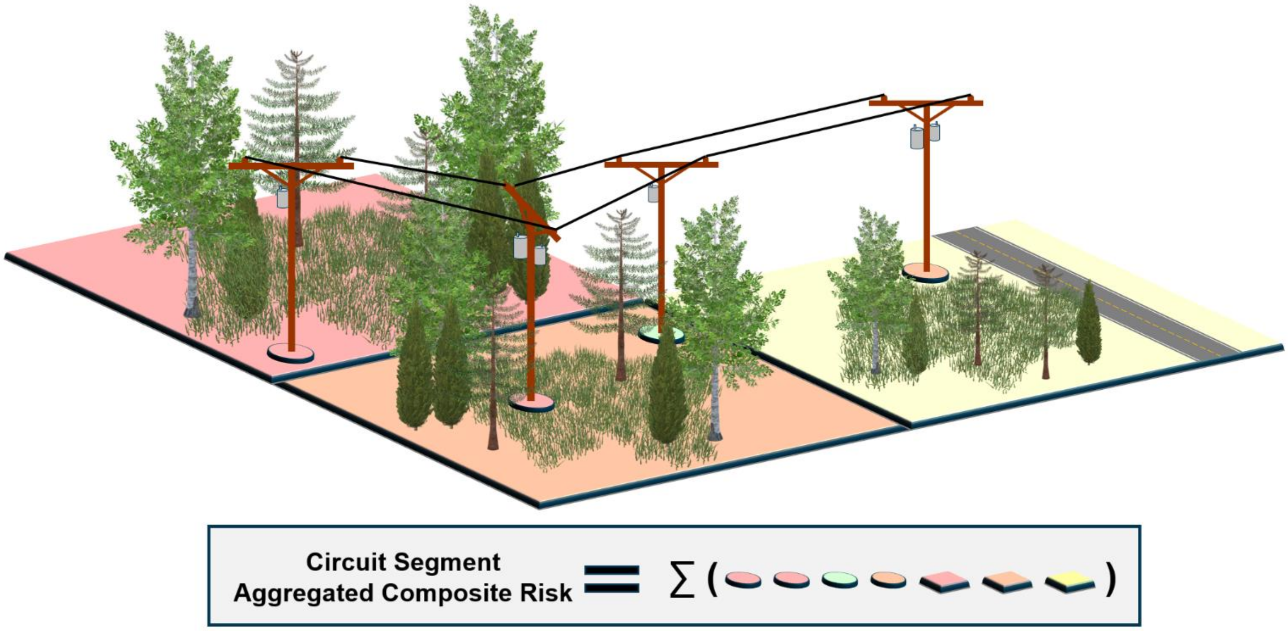


Figure 22 - Single Circuit Segment Aggregated Composite Example

The electrical grid network, however, is more complicated than the example present above. Sometimes, a support structure supports conductors for more than one circuit segment. Another common configuration issue is when multiple circuit segments pass through the same grid pixel. In these cases, some the aggregation must distribute the shared risk. *Figure 23* presents a configuration where two of the grid pixels are spanned by two distinct circuit segments. In this case, while asset risk can be directly attributed to its containing circuit segment, the grid pixel risk must be equally shared by the two segments.

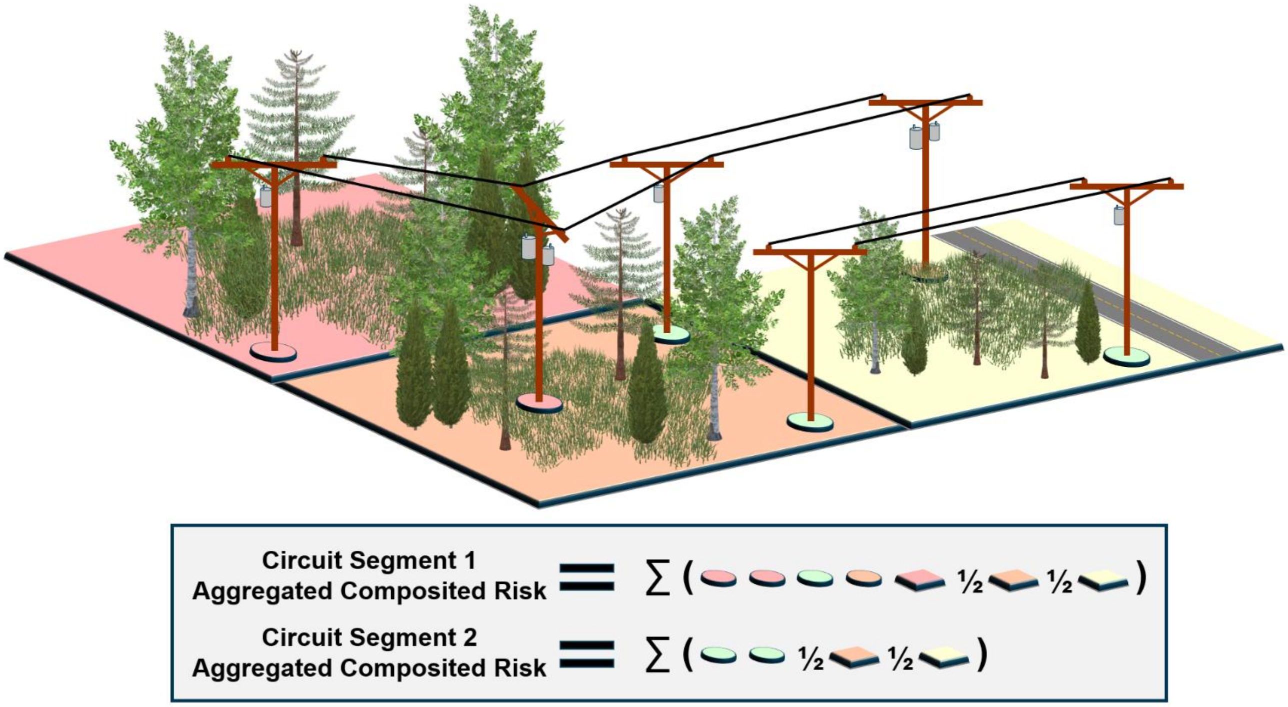

Circuit Segment 1 Aggregated Composited Risk = ∑(½ ½)
Circuit Segment 2 Aggregated Composited Risk = ∑(½ ½)

Figure 23 - Multiple Circuit Segment Aggregated Composite Example

4.3.1.4 WDRM v3 and v4 Circuit Segment Composite Aggregation Comparison

The aggregation of composited risk is a major difference between WDRM v3 and v4, and the updated aggregation results in more accurate relative circuit segment risk scores for v4. The single pixel circuit segment configuration provided in *Figure 24* will be used to illustrate the difference between v3 and v4 circuit segment risk values. Note that a set of example risk values have been associated with the icon colors in the figure.

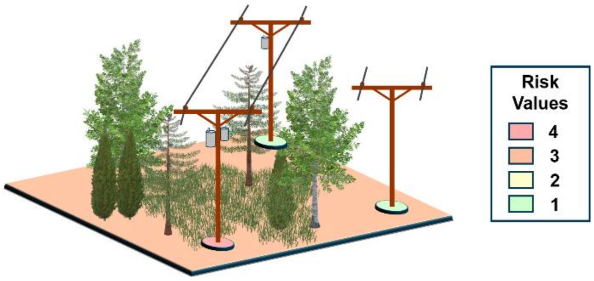


Figure 24 - v3/v4 Circuit Segment Risks Example

For v3, all event probability model risk results were recorded as pixel values. For the sake of brevity, the asset composite results presented in *Figure 24* will be assumed to be pixel level results. For v3, a composited pixel risk value was calculated and distributed equally to the two circuit segments that cross the pixel as shown in *Figure 25*. In the v3 example, each circuit segment was assigned a risk value of 4.5 and the two circuits would have been considered to be of equal priority for mitigation work.

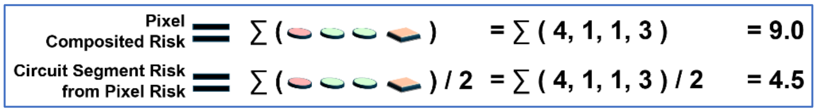


Figure 25 - WDRM v3 Circuit Segment Pixel Risk Assignment

In contrast, for v4 the equipment asset risks are directly assigned to their containing circuit segments and only pixel model risk is shared equally between the two circuit segments. *Figure 26* illustrates that the direct assignment of asset risk results in different circuit segment risk values, and hence, very different priorities for receiving potential mitigation work. This result makes sense, as one segment has an asset with considerably greater wildfire risk than the other.

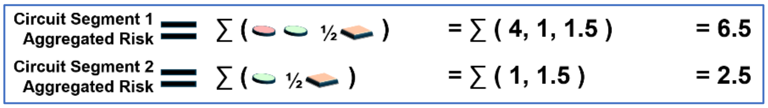


Figure 26 - WDRM v4 Circuit Segment Asset & Pixel Risk Assignment

4.3.2 Mitigation Work Composites

The number of event probability event models has grown with each new version of the WDRM to provide better causal modeling and interpretation of wildfire risk. An All model risk composite, depicted in *Figure 27*, is built for each WDRM version that combines the entire suite all equipment asset and pixel model event probability risk results into a total composite risk result for each circuit segment.

The All composite, however, is usually of interest only to the RaDA team for model calibration and quality control. Most mitigation program work planners are interested in only a partial set of event probability model risk results that quantify only the risks that can be addressed by a specific program.

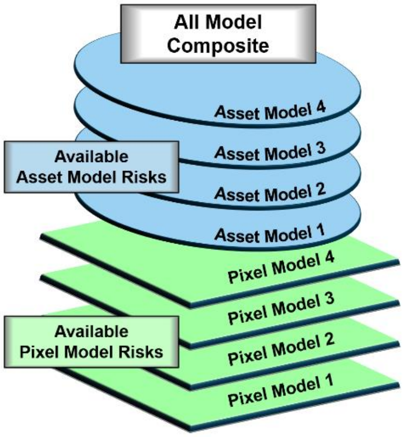


Figure 27 - All Model Composite

Custom composites are configured so that the total risk for only the applicable event probability models gets considered as part of the work planning and prioritization process. *Figure 28* illustrates how model selection can be used to configure a composite of selected event probability models for a specific work plan. In the example, only a sub-set of the available equipment asset and pixel model risk results are relevant to the mitigation work program.

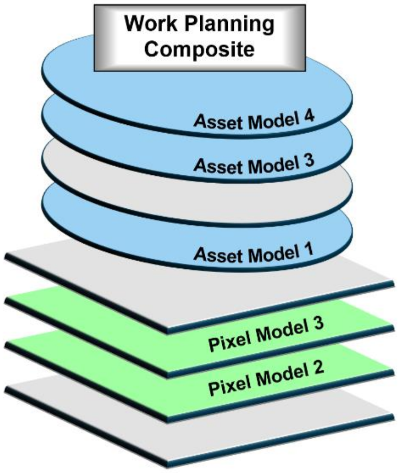


Figure 28 - Mitigation Work Planning Composite

4.3.3 Ignition-Weighted Consequence

The compositing of asset and pixel model results for WDRM v4 revealed that the relationship between aggregated wildfire risk, consequence, and probability of ignition is not straightforward. Simply multiplying an aggregated probability of ignition with an aggregated consequence value will not produce a correct aggregated risk value. To address this issue, it is necessary to back-calculate an aggregated consequence value from aggregations of probability of ignition and wildfire risk. The back-calculated consequence is called the Ignition-Weighted Consequence.

The example circuit segments shown in *Figure 29* will be used to explain how an Ignition-Weighted Consequence value is calculated and why it is necessary for aggregated composites. Note that the only difference in the two circuit segments is the pixel location of the center support structure and its associated assets.

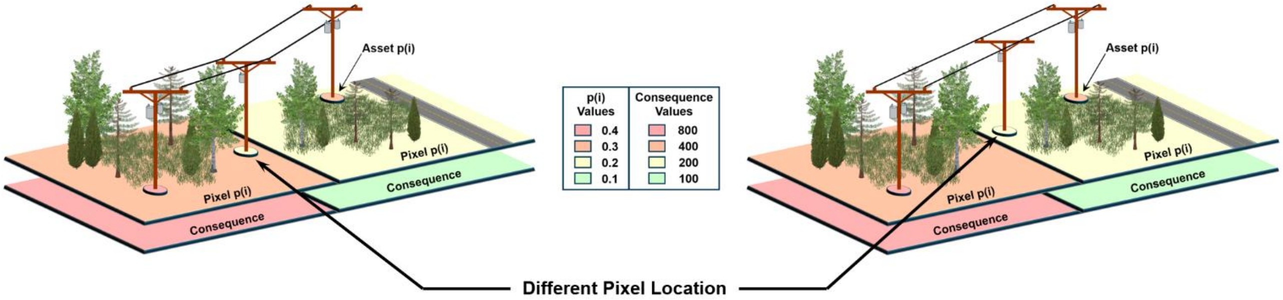


Figure 29 - Ignition-Weighted Consequence Explanation Circuit Segments

The box in the center of the example diagram provides color-coded numeric values for p(i) and wildfire consequence to be used for determining circuit segment risk and ignition-weighted consequence values.

A key understanding for the example is that both circuit segments have identical aggregated p(i) values and the same average consequence value, as shown by the calculations in *Figure 30*. Wildfire risk is defined as a p(i) value multiplied by a consequence value. However, an aggregated composite circuit segment risk value cannot be directly determined by multiplying the aggregated p(i) and the average circuit segment wildfire consequence. Aggregated composite risk must consider the location of the modeled assets to produce the correct aggregated risk value.

Circuit Segment Aggregated p(i) = ∑ (🟥 🟨 🟧 🟧 🟨)
= ∑ (0.4, 0.2, 0.3, 0.3, 0.2)
= 1.5

Circuit Segment Average Consequence = ∑ (🟥 🟩) / 2
= ∑ (800, 100) / 2
= 450

Circuit Segment Risk ≠ Aggregated p(i) × Consequence
≠ 1.5 × 450
≠ 675

Figure 30 - Incorrect Aggregated Composite Risk Example

Consider the sample circuit segment where the center support structure is located in the left-hand grid pixel (*Figure 31*). For this case the aggregated composite risk would be calculated as shown in *Figure 32*.

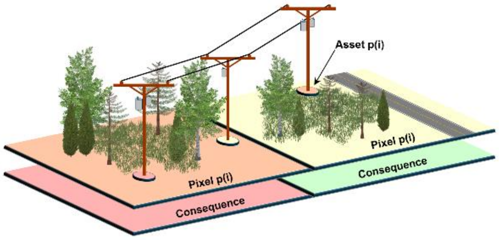


Figure 31- Example Circuit Segment Configuration 1

Circuit Segment Aggregated Risk = ∑ (🟥 🟨 🟧) × 🟥 + ∑ (🟧 🟨) × 🟩
= ∑ (0.4, 0.2, 0.3) × 800 + ∑ (0.3, 0.2) × 100
= 770

Figure 32- Configuration 1 Aggregated Risk Value

Contrast the results for Configuration 1 with Configuration 2, where the center support structure is located in the right-hand grid pixel (*Figure 33*). The right-hand grid pixel has a much lower consequence value and, as a result, the aggregated composite risk value for the circuit segment will be much lower, as demonstrated in *Figure 34*.

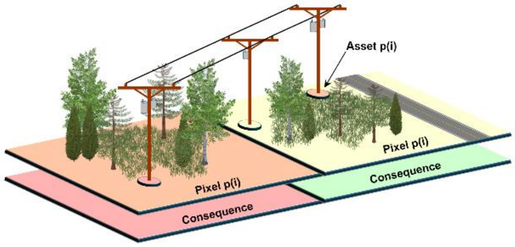
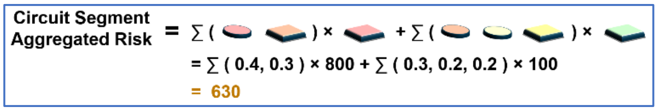


Figure 33 - Example Circuit Segment Configuration 2



Circuit Segment Aggregated Risk = ∑ (● ▰) × ▰ + ∑ (● ● ▰) × ▰

= ∑ (0.4, 0.3) × 800 + ∑ (0.3, 0.2, 0.2) × 100

= 630

Figure 34 - Configuration 2 Aggregated Risk Value

Note that neither of the aggregated composite risk values determined for the example circuit segment configurations match the risk value calculated from the summed p(i) and average consequence values presented in *Figure 30*.

Since the circuit segment asset configuration against the network grid pixel locations can strongly influence its aggregated composite risk value, then its relative consequence value must also consider the asset locations. This is achieved by dividing the aggregated composite risk value by its summed probability of ignition to yield an ignition-weighted consequence value.

Figure 35 provides the ignition-weighted consequence values for the two example circuit segment configurations. Note that neither of the ignition-weighted consequence values match the average pixel consequence value presented in *Figure 30*.

Ignition-Weighted Consequence = Aggregated Risk / Aggregated p(i)

Configuration 1 Ignition-Weighted Consequence = 770 / 1.5 = 513

Configuration 2 Ignition-Weighted Consequence = 630 / 1.5 = 420

Figure 35 - Configuration 1 and 2 Ignition-Weighted Consequence

5 Future Plans

New algorithms and methodologies are developed as needed for new model development and to fulfill new user requirements. In response to user requests, potential new developments include:

- Isolation Zone aggregation and compositing of event probabilities and risk.
- Support Structure aggregation and compositing of event probabilities and risk.
- Regional aggregation of event probabilities and risk.
- Multi-risk optimization for wildfire, PSPS, reliability, and public safety risks.