fire

MDPI

Article

An Empirical Modelling and Simulation Framework for Fire Events Initiated by Vegetation and Electricity Network Interactions

Roy Wilson, Rohan Wickramasuriya and Dean Marchiori *

Endeavour Energy, 51 Huntingwood Drive, Huntingwood, NSW 2148, Australia
* Correspondence: dean.marchiori@endeavourenergycom.au

Abstract: Electrical infrastructure is one of the major causes of bushfire in Australia alongside arson and lightning strikes. The two main causes of electrical-infrastructure-initiated fires are asset failure and powerline vegetation interactions. In this paper, we focus on powerline–vegetation interactions that are caused by vegetation falling onto or blowing onto electrical infrastructure. Currently, there is very limited understanding of both the spatio-temporal variability of these events and their causative factors. Bridging this knowledge gap provides an opportunity for electricity utility companies to optimally allocate vegetation management resources and to understand the risk profile presented by vegetation fall-in initiated fires, thereby improving both operational planning and strategic resource allocation. To bridge this knowledge gap, we developed a statistical rare-event modelling and simulation framework based on Endeavour Energy's fire start and incident records from the last 10 years. The modelling framework consists of nested, rare-event-corrected, conditional probability models for vegetation events and consequent ignition events that provide an overall model for vegetation-initiated ignitions. Model performance was tested on an out-of-time test set to determine the predictive utility of the models. Predictive performance was reasonable with test set AUC values of 0.79 and 0.66 for the vegetation event and ignition event models, respectively. The modelling indicates that wind speed and vegetation features are strongly associated with vegetation events, and that Forest Fire Danger Index (FFDI) and soil type are strongly associated with ignition events. The framework can be used by energy utilities to optimize resource allocation and prepare future networks for climate change.

Keywords: electricity network; fire risk; powerlines; vegetation

Citation: Wilson, R.; Wickramasuriya, R.; Marchiori, D. An Empirical Modelling and Simulation Framework for Fire Events Initiated by Vegetation and Electricity Network Interactions. *Fire* **2023**, *6*, 61. https://doi.org/10.3390/fire6020061

Academic Editors: Begoña Vitoriano and Jesús Barreal

Received: 2 December 2022
Revised: 16 January 2023
Accepted: 6 February 2023
Published: 8 February 2023

1. Introduction

Electricity transmission infrastructure is one of the four main causes of bushfires alongside lightning, arson and accidental fire escape [1,2]. Fires caused by electricity transmission infrastructure can be divided into those caused by equipment failure and those caused by interactions between powerlines and the surrounding vegetation. Although bushfires initiated by powerline–vegetation interactions are relatively infrequent compared with the other main causes of bushfires, they tend to occur on days with extreme fire weather conditions and are initiated closer to population centres, thus leading to comparatively larger burned areas and greater damage to lives and property [1,3–5]. Consequently, implementing measures to reduce bushfire risk associated with powerline–vegetation interactions has become a top priority for electricity network service providers. Such measures include improvements to network infrastructure (e.g., covered conductors) and actively managing vegetation on and near powerlines. However, committing to and implementing any of these intervention mechanisms requires meticulous planning due to costs, legal and socio-cultural issues, and the associated complexity of the powerline–environment interactions [6]. Vegetation management planning requires risk estimates that are calculated at sufficiently fine spatio-temporal scales to allow the risk variation over

networks to be considered. Direct estimation of the risks posed by powerline–vegetation interactions is difficult due to the rarity of these events; however, the problem can be approached by modelling the probability of events that led to fire events as these preceding events have considerably greater frequency. Consequently, in this paper, we focus on modelling ignition events that can be considered to have been the result of vegetation–powerline interactions.

The literature includes several attempts to model ignition events with a variety of causes at a fine spatio-temporal resolution. The general approach has been to formulate the problem as a bernoulli process, where the two mutually exclusive outcomes are ignition and no ignition [7]. One study [8] explored the potential of Logistic Regression and Decision Tree algorithms to convert satellite-derived Live Fuel Moisture Content (LFMC) into ignition probability for the Iberian Peninsula territory of Spain. Authors stated that Logistic Regression model, a form of Generalised Linear Model (GLM), performed the best as measured by a popular metric, Area Under the Curve (AUC), of over 0.65 for a section of the peninsula and an AUC of over 0.8 for the rest of the peninsula. Another study [9] compared the predictive performance of three such algorithms namely, GLM, Random Forest and Maximum Entropy (MaxEnt), using 16 years of ignition data and environmental data for the Huron-Maanistee National Forest in Michigan, USA. Authors found that the two machine learning algorithms (MaxEnt and Random Forest) performed slightly better than GLM, which is a statistical model. Another study [10] explored factors leading to ignition in the Sydney basin using a probabilistic modelling approach where a Generalised Additive Model (GAM) with a binomial distribution was the algorithm of choice. Authors compared the differences between known ignition locations and a set of randomly selected non-ignition locations in terms of topographic, vegetation and fire weather variables. The ability of GAMs to take into account the non-linear relationships between predictor variables was the main reason for its selection. The authors of [11] modelled both human-caused and lightning-caused ignitions in the Australian state of Victoria using Random Forest occurring in native vegetation and cleared/urbanized land. This study demonstrated the utility of machine learning models in identifying relationships between ignitions and patterns of landscape and weather variables, with the models achieving prediction accuracy of between 86.4% and 90.3%.

Among studies that explore ignition events at a fine spatio-temporal scale, there are a limited number of studies that specifically look at the probability of ignitions caused by electricity networks [12–15]. One such study [12] used data collected from multiple electrical distribution networks in Australia to calculate fault and subsequent ignition rates under different electrical infrastructure, landscape and weather scenarios. The study used empirical detection rates calculated from data categorized by different levels of explanatory variables. The study found that the variables most strongly associated with fault rates were wind speed and vegetation type and that the variable most strongly associated with ignition was Fire Danger Rating (FDR) and that this relationship was influenced by the cause of the fault. The importance of wind speed in determining fault rates for powerlines was also investigated by [13], who found that increased wind speed leads to considerable increase in outage probability. Another study [15] conducted a study of ignition modelling using data from the Pacific Gas and Electricity (PG&E) franchise in California at the feeder and day level of spatio-temporal resolution. The study investigated the predictive performance of logistic regression, random forests and gradient boosted trees together with different methods for dealing with the large class imbalance present in the data. The study found that the most effective algorithm and class imbalance method for modelling ignition probability was a combination of gradient boosted trees and majority class down-sampling and that the most influential features were various daily weather summaries derived from the Gridded Surface Meteorological dataset (gridMET) and vegetation features, as well as feeder length [15].

The objective of this study is to develop a modelling framework for the causal chain from weather events that occur in different landscape and network contexts through

to events that are characterized by vegetation falling onto or blowing onto powerlines (vegetation events) and subsequent ignition events. The framework is intended to be implemented in network management and planning activities and as such should provide event probability predictions that are calibrated to the actual rate of events on the network and that are calculated at a spatio-temporal scale that is convenient for the management and planning activities carried out by electricity distributors. An additional objective of the study is to identify the weather, landscape and network features that are related to the probability of both vegetation and ignition events across the network and to investigate the utility of Light Detection and Ranging (LiDAR) derived features in predicting these events.

We found that although ignitions are rare events, we were able to generate reasonably well performing models through adopting a two-step conditional probability approach that involved modelling vegetation events first and then modelling ignition events for all data for which vegetation events were observed. For the approach to succeed we also required two other technical modelling adjustments namely: majority class down-sampling and rare event correction. The framework allows us to estimate vegetation event and ignition event probability at any point in the network under different weather conditions and as such can be used for estimating expected ignition risk over arbitrary spatio-temporal windows. In addition, we are able to simulate events from the framework and thereby investigate the range of probable vegetation and ignition event frequencies under different weather, landscape and network conditions. We find that the major factors influencing vegetation events are weather variables including wind speed and rain followed by landscape variables such as vegetation and soil type, while network engineering variables appear to have a small influence. The major factors influencing ignition events that follow vegetation events are landscape features such as vegetation and soil type features followed by weather features such as the McArthur's Forest Fire Danger Index (FFDI) and rainfall. As with the vegetation event model, network engineering features have a smaller impact, with bay length being the most influential engineering feature.

2. Materials and Methods

2.1. Study Area

Endeavour Energy (Endeavour) operates a power distribution network that spans 24,800 square kilometers across Sydney's Greater West, Illawarra and the South Coast, Blue Mountains and Southern Highlands (Figure 1). This particular area in New South Wales, which is dominated by sclerophyll forests, generally is considered fire-prone [10]. For example, the area was subject to the "black summer fires" of 2019–2020, a fire event that resulted in 7.9 Mha of burnt land across south-eastern Australia and directly resulted in 33 deaths [16]. The network covers a wide variety of land-use types including urban, rural and native bushland. Much of the network is located at the interface of urban areas and native bushland and as such the network carries a significant risk of high-impact fire events.

2.2. The Event Modelling Scale

The vegetation and ignition events were modelled at the bay/day scale. Bay refers the unit of an electricity network that includes a conductor or set of conductors that are between two poles. A day refers to a given 24 h period. Therefore, the modelling was concerned with estimating the probability of vegetation and consequent ignition events occurring within any particular bay on the network during a given day.

2.3. Event Variables

There are two primary response variables, namely the vegetation event, which is simply defined as vegetation contacting powerlines in some way, and the ignition event, which is defined as a fire that started on the network as a result of a vegetation event.

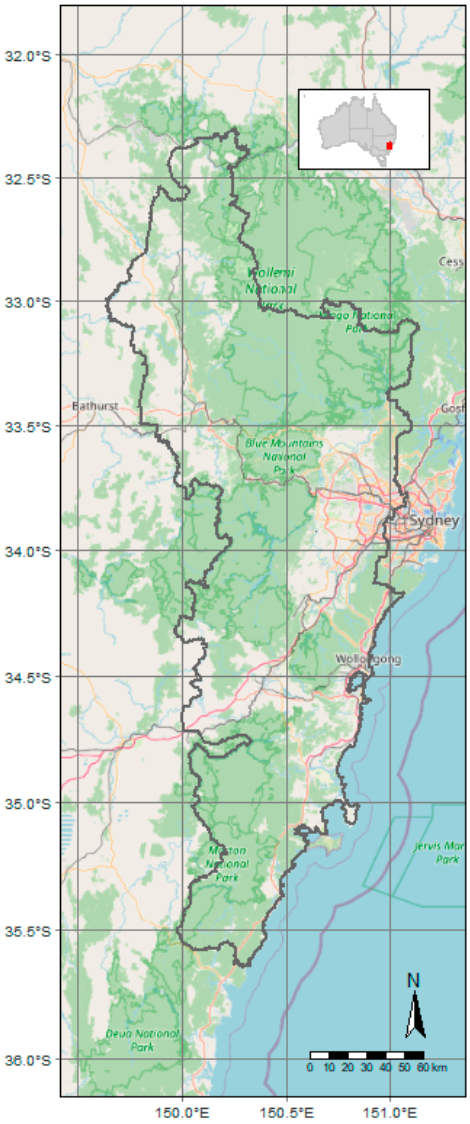

Figure 1. The Endeavour Energy network coverage area.

Vegetation events can be roughly categorized as "fall-ins", "blow-ins" and "grow-ins". In this paper we focus on "fall-ins" or "blow-ins", these are events that involve vegetation falling onto or blowing onto powerlines and can be distinguished from "grow-in" events which involve vegetation growing into powerlines. Vegetation grow-in events are typically much rarer because routine maintenance of powerline easements tends to greatly reduce their frequency of occurrence. For the remainder of the paper the term vegetation event refers to a vegetation–powerline interaction that can be characterized as either a "fall-in" or "blow-in" event.

Ignition events are characterized as events that involve some level of ignition of combustible material either on or immediately adjacent to the network. We do not distinguish between events that completely remain on the network (e.g., Pole top fires) or those that spread to surrounding vegetation. In this study we only consider ignition events that can reasonably be attributed to vegetation events as described above.

Vegetation events that were recorded between the dates of 1 July 2012 and 4 April 2021 were extracted from Endeavour's Incident Reporting Database. All incidents were selected regardless of whether the event resulted in a customer interruption.

Ignition events that were recorded between 1 July 2012 and 4 April 2021 were extracted from Endeavour's Fire Investigations Database. The ignitions were filtered by only including ignitions that were caused by network-vegetation interactions. The ignition

events were then matched to the identified vegetation fall-in and blow-in events based on spatio-temporal proximity thresholds.

2.4. Input Features

The model input features can be divided into three main categories, these being landscape, weather and network features (Table 1). The landscape features were extracted from: a whole of network LiDAR analysis undertaken by helicopter; a Digital Terrain Model (DTM) for the network; and from Australian Government soil type and vegetation community maps. The network features were obtained from Endeavour's Geographic Information System (GIS) database and the weather features were derived from half-hourly weather data obtained from the Australian Bureau of Meteorology (BOM). The same set of input features was used for both the vegetation and ignition event models.

Table 1. Input features used for modelling.

Variable Type	Name	Data Source	Data Description
Daily Weather	max wind speed	BOM	maximum daily wind speed (km/hr)
	Rain	BOM	sum of precipitation since 9 a.m local time (mm)
	dew point	BOM	mean dew point temperature (°C)
	temperature	BOM	maximum air temperature (°C)
	FFDI	BOM/Endeavour	refer to explanation in text
Landscape	soil type	Soil and landscape grid of Australia	The soil type at the location of the canopy peak point
	soil depth	Soil and landscape grid of Australia	The average soil depth across the given bay (m)
	fall-in tree exposure	Endeavour	Count of vegetation canopy objects classed as fall-in hazards linked to this bay by closest distance
	maximum vegetation height	Endeavour	The maximum vegetation height of vegetation within a given bay (m)
	average vegetation height	Endeavour	The average vegetation height within a given bay (m)
	vegetation type	Australian Dept of Agriculture and Water Resources	The most common vegetation type within the given bay
	canopy count in easement	Endeavour	Count of all canopies within a given bay's easement
	terrain aspect	Endeavour	The aspect of the terrain for a given bay (degrees from north)
	terrain wind exposure	Endeavour	Exposure of the terrain at a given bay to the prevailing wind direction for that bay
	terrain gust exposure	Endeavour	Exposure of the terrain at a given bay to non-prevailing wind gusts
Network	voltage	Endeavour	The highest voltage conductor present within a given bay
	bay length	Endeavour	The length of a given bay (m)

The raw data for the weather features were sourced from BOM half hourly weather data from 20 weather stations within and near the Endeavour franchise area. The data were summarized at a daily level to match the temporal resolution of the response variables. For each bay, the nearest weather station was identified using a spatial join, and the weather data from the matching stations were assigned to bays.

In addition to basic daily weather variable summaries, the McArthur's Forest Fire Danger Index [17] (FFDI) was calculated. The FFDI was calculated using daily aggregations of half-hourly weather data obtained from the BOM (maximum temperature, sum of

half hourly precipitation and average dew point). The Keetch–Byram Drought Index (KBDI) [18,19] was used to provide an estimate of soil dryness and the dew point was used to calculate relative humidity, this information was combined to estimate daily FFDI feature.

2.5. Data Preparation for Modelling

2.5.1. Dealing with Class Imbalance

The modelling data have highly imbalanced labels for the two response variables: vegetation event and ignition event (Table 2). This is demonstrated by comparing the number of events recorded in Table 2 with the number of day/bay combinations where no vegetation event or vegetation ignition related event took place, which is typically around 168 million for any given year. Vegetation events are typically not rare events within the context of the entire network; however, they are highly outnumbered by the negative class, i.e., no vegetation event for a given bay/day.

Table 2. Counts of vegetation and ignition events during the modelling period.

Fire Season	Vegetation Events		Ignition Events	
	Events	Non-Events	Events	Non-Events
2012–13	2127	167,657,001	12	2115
2013–14	2844	167,656,284	20	2824
2014–15	3137	167,655,991	11	3126
2015–16	2935	168,116,795	9	2926
2016–17	2789	167,656,339	13	2776
2017–18	2069	166,275,253	15	2054
2018–19	3796	166,655,332	22	3774
2019–20	4732	168,114,998	38	4964
2020–21	2814	127,583,940	10	2804

In order to overcome this class imbalance, we used the rare event sampling and prior correction strategy recommended by [20], that is:

1. Include all the positive (minority) class in the sample;
2. Down-sample the much larger negative class using a random sample from the entire negative class population; and
3. Adjust for the introduced sample bias using the prior correction method.

The prior correction method works by introducing bias to the unconditional model estimate of the event probability through down-sampling of the majority class. This down-sampling roughly equilibrates the positive and negative class numbers and allows us to obtain unbiased estimates for the model parameters. Model predictions obtained from the trained model are then adjusted by a constant to take into account the effect of the down-sampling on the average prediction probability. The adjusted probabilities are derived as follows:

$$\pi_{adj} = log\left(\frac{\pi}{1-\pi}\right) - log\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\overline{y}}{1-\overline{y}}\right)\right]$$

where:

π_{adj} = The adjusted probability
π = The model predicted probability
τ = The positive class proportion before downsampling
$\overline{y}$ = The positive class proportion after downsampling

For the current study, this method served two purposes, firstly, it reduced the size of the model training set to one that is more reasonable for model fitting, secondly it removed any biases that may result from the high level of class imbalance in the data.

2.5.2. Defining Training and Test Sets

The full data were split into a training set for model training and validation and a separate test dataset. This allows us to train a model using just the training dataset, and independently verify if the model generalises well to unseen data by assessing its performance against a separate test dataset.

For the training/test split, we used an out-of-time test set approach with the training/validation set using data from 1 July 2012 through to 3 April 2020, and the test set including the final year of data (4 April 2020 to 4 April 2021). The out-of-time test set approach allows us to test the ability of the models to forecast future event distributions.

For the vegetation event model, the training set was obtained by:

1. Including all vegetation incidents within the training data period.
2. Randomly sampling a negative class sample of the same size as the positive class from all of the bay/day combinations that did not result in a vegetation incident.

For the ignition event model, the training set was obtained by:

1. Conditioning on vegetation events to form the full ignition training set.
2. Sub-sampling from the negative class (i.e., vegetation event with no ignition) to reduce the class imbalance such that the negative class is five times the size of the positive class (i.e., vegetation event with ignition). This was carried out because ignition events are rare, and having a negative class that is some small multiple of the positive class (typically less than or equal to 5 is acceptable [20]) allows us to include more information in the modelling.

2.6. Modelling Framework

To understand ignition risk across the network we require an estimate at the bay/day level of the probability of a bushfire ignition event that is caused by a vegetation event. Such an ignition event can be represented as the product of the probability of a vegetation event and an ignition event that is conditional on a vegetation event occurring. We make the assumption that vegetation and ignition events that co-occur within a small spatio-temporal window (bay/day) are causally related and that they occur in order of a vegetation event followed by an ignition event. This assumption is further supported by the fact that we only include ignition events that have a vegetation event listed as the cause. Consequently, for a given network bay on a given day we have the following two models:

1. Vegetation Event Model: $\Pr(V|N, L, W)$
2. Ignition Event Model: $\Pr(I|V, N, L, W)$

where V represents a vegetation event, I represents an ignition event and N, L and W represent network, landscape and weather features.

The dependence between the vegetation and ignition event probability models is implicitly included in this framework because they are conditioned on the same set of features. An additional benefit of the conditional dependence framework is that it allows us to reduce the amount of class imbalance by fitting a model for a more common event (vegetation event) and then conditioning on that event to fit a model for a genuinely rare event (ignition event).

2.7. Vegetation and Igntion Event Models

All modelling and data preparations were undertaken using R statistical software, with data and modelling pipelines developed using the targets [21] package.

For both the vegetation and ignition event models, we fit a Gradient-boosted Machine (GBM) model [22] with a logistic loss function using the gbm package in R. GBM models have demonstrated good performance in similar studies and are useful for quickly determining the potential performance of statistical regression models.

The GBM model encodes the relationship between the input variables and the probability of vegetation events based on the available historical data and as such allows the modelling framework to obtain predictions of the probability of vegetation and ignition

events under different landscape, weather and network conditions. As GBM is a tree-based regression method, the relationships are encoded in a complex tree structure rather than estimating single weights for each input feature. The method allows us to extract the relative importance (relative proportion of explained variance attributable to each feature) of the input features and to investigate the directionality of the feature effects using marginal effect plots [22]. It is important to note that while the observed relationships are useful in predicting vegetation events, we cannot necessarily assume that they are causal.

We trained a GBM model with an interaction depth set at 3 and the number of trees to include in the model determined using 5-fold cross validation on the training set. In total, 17 variables representing network infrastructure, landscape and weather were used as input features.

2.8. Model Performance Assessment

For testing the models, we defined a test set that includes the final year of data, which is 4 April 2020 through to 4 April 2021.

Model performance for the vegetation model was tested using the following methods:

1. Area Under the Curve (AUC) performance on the test set;
2. Baseline vs. full model expected count comparison on the test set.

Model performance for the ignition model was undertaken only using AUC as the number of ignitions in the data was too small to undertake the baseline vs. full model expected count comparison.

2.8.1. Model Assessment Using AUC

AUC is a standard classifier performance assessment method and gives an indication of how well the classifier orders different bay/day combinations with respect to the likelihood of event occurrence. A value of 0.5 indicates the classifier is no better than a random guess, whereas a value of 1 indicates a perfect classifier. The AUC value can be interpreted as the probability that a randomly selected positive class example ranks above a randomly selected negative class example.

2.8.2. Baseline vs. Full Model Expected Count Comparison

The baseline vs. full model expected count comparison allows us to test if the rare event correction results in a correct adjustment of the probabilities, and to see if the model has a significant amount of signal in comparison to a baseline model that applies a constant probability of vegetation incident across bays/days.

For each bay/day combination in the test set, 20 simulations were run to create 20 simulated years of data for the network. These simulations were undertaken using:

1. Baseline model. This simulation used an identical average probability of a vegetation event over every bay/day combination.
2. Full model. This simulation used the full model to predict the probability of a vegetation event based on the combination of bay landscape and network features and the weather features for the day.

The simulated number of incidents for both methods were summed over a pre-defined geographic aggregation at the sub-depot level (this divides the Endeavour franchise into 46 spatial regions) and divided by 20 to obtain an expected number of incidents at each sub-depot for the simulated test set year.

In addition, the actual number of incidents was summed over each sub-depot to obtain the actual number of incidents for each sub-depot for the test set year.

Correlations of actual incidents to predicted incidents using both the baseline and full models were then compared. This allowed assessment of the model signal and the rare event adjustment technique.

2.9. Event Prediction

Using the ignition event modelling framework set out above, we are able to extract various measures of ignition risk over any given spatio-temporal window. There are two main types of ignition risk measure, these being:

1. Expected ignition count;
2. Simulated ignition count distributions.

2.9.1. Expected Ignition Count

The expected ignition probability for any given bay/day combination provides a point estimate of ignition probability at a particular spatio-temporal location at the bay/day scale. When summed over any given spatio-temporal window we can derive a point estimate of the expected count of ignition events within that spatio-temporal region. These measures are useful for ranking risk and including in vegetation management decisioning tools. One drawback of this approach is that expected counts are point estimates and do not allow a full understanding of the ignition risk distribution, which may be quite complex.

2.9.2. Simulated Ignition Count Distributions

The simulated ignition probability distributions provide a full risk distribution over any defined spatio-temporal window. For example, they allow us to derive event and risk distributions for the entire Endeavour Energy franchise area for a given year. These simulations are useful for investigating rare events such as vegetation-related ignition events because they allow us to understand the possible range of events that may occur within a defined spatio-temporal window along with their probability of occurrence. This provides a much richer understanding of the ignition exposure risks in any given area, and how much they may be expected to vary.

3. Results

3.1. Model Fitting

3.1.1. Vegetation Event Model

The resultant GBM model for vegetation events can be described in terms of the relative importance of the model input features with respect to prediction of vegetation events [22] (Table 3).

Weather variables consisting of wind speed and rainfall are the most influential variable category for predicting a vegetation event with a combined relative importance of 56%. Landscape variables such as soil type and vegetation type are the second most important category with a combined relative importance of 39%, followed by network variables whose relative importance stands at only 6%. The low importance of network variables is possibly due to the limited number and scope of network variables included in the analysis.

The GBM also allows the characterization of the association between input variables and the probability of a vegetation event using marginal effect plots (Figure 2).

The marginal effect plots show the influence of any given input variable in the model on the outcome variable while averaging over the influence all other variables and interactions in the model.

The most influential variable in terms of predicting vegetation events is wind speed. The marginal effect plot shows that increasing wind speeds are associated with a greater probability of vegetation events occurring (Figure 2a). While this is an unsurprising result, it is important to note that the model can be used to estimate the increase in vegetation event probability that is caused by a specific increase in wind speed. The other top influential variables are: fall in tree exposure (Figure 2b), maximum vegetation height in the bay (Figure 2c), and rainfall (Figure 2d).

Table 3. Relative importance of input variables in the vegetation and ignition event models.

Variable Type	Name	Relative Importance (%)	
		Vegetation Event Model	Ignition Event Model
Daily Weather	wind speed	39	0
	rain	8	4
	dew point	5	0
	temperature	2	2
	FFDI	2	18
	weather subtotal	56	24
Landscape	soil type	11	35
	soil depth	1	6
	fall-in tree exposure	10	4
	maximum vegetation height	10	3
	average vegetation height	3	6
	vegetation type	1	2
	canopy count in easement	2	3
	terrain aspect	1	2
	terrain wind exposure	0	0
	terrain gust exposure	0	0
	landscape subtotal	39	61
Network	voltage	3	1
	bay length	3	12
	network subtotal	6	13

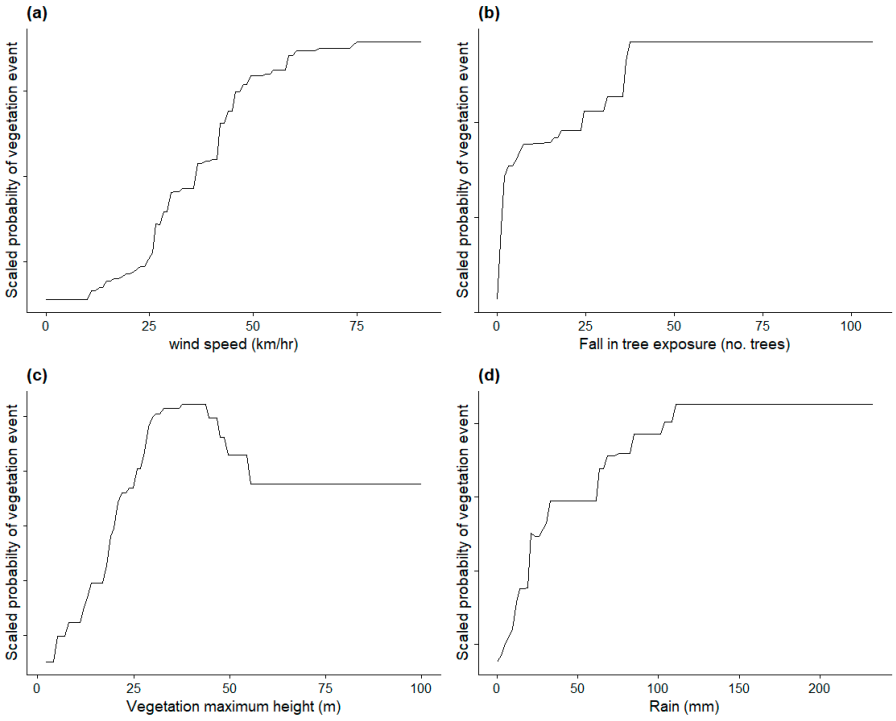

Figure 2. Marginal effect plots for (**a**) wind speed; (**b**) fall-in tree exposure; (**c**) vegetation maximum height; and (**d**) daily rainfall for the vegetation event model. These are the top 4 influential variables (excluding categorical variables) in the vegetation event model. The black line shows the relationship between the probability of a vegetation incident and the input variable averaging over all other input variables in the model. The y axis is the probability scaled by the logistic function and is calculated prior to the rare event correction. The sub-plots are not directly comparable; however, their relative influence can be obtained from Table 3.

3.1.2. Ignition Event Model

For the ignition model, landscape variables such as soil type and average vegetation height were the most influential with a combined relative importance of 45%, followed by weather variables such as FFDI and rainfall with a combined relative importance of 32% (Table 3). Network variables such as span length and voltage had the lowest relative importance (23%); however, this is considerably higher than the relative importance of network variables displayed for the vegetation event model.

The relationship between the top four continuous input variables in terms of relative importance in predicting ignition events is demonstrated in the marginal effect plots (Figure 3).

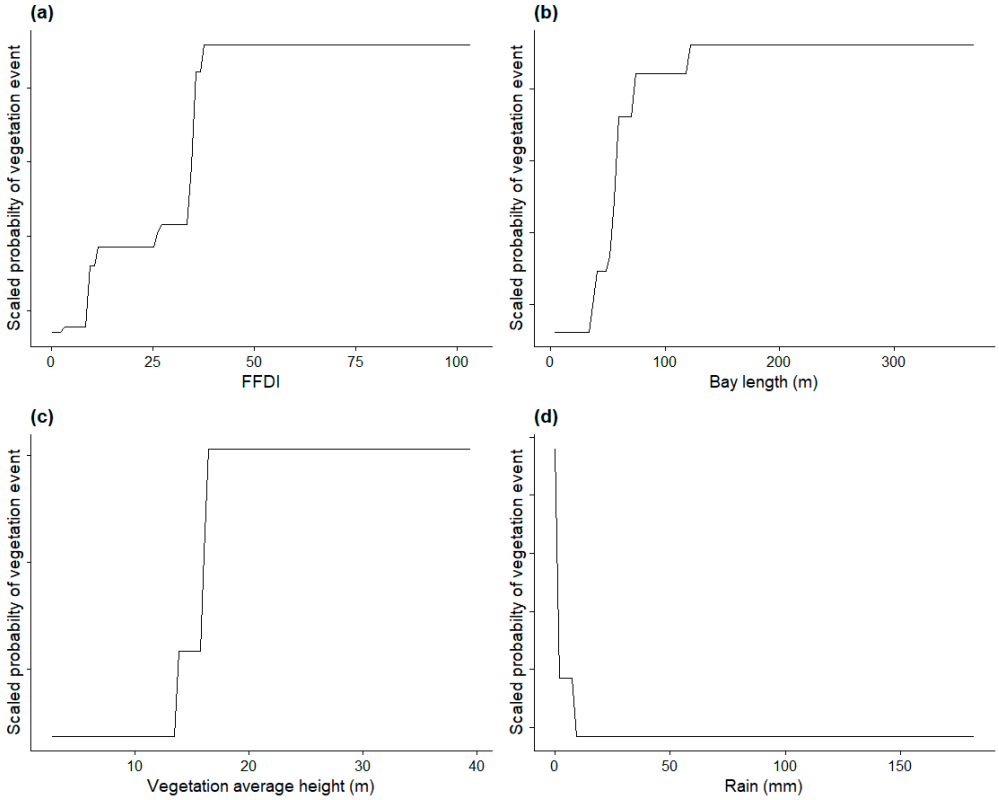

Figure 3. Marginal effect plots for (**a**) FFDI; (**b**) bay length; (**c**) vegetation average height; and (**d**) daily rainfall for the ignition event model. These are the top 4 influential variables (excluding categorical variables) in the ignition event model. The black line shows the relationship between the probability of a vegetation incident and the input variable averaging over all other input variables in the model. The y axis is the probability scaled by the logistic function and is calculated prior to the rare event correction. The sub-plots are not directly comparable; however, their relative influence can be obtained from Table 3.

3.2. Model Performance Evaluation

3.2.1. Vegetation Event Model

The vegetation event model demonstrated good predictive performance on the test set with an AUC of 0.79.

The results of the baseline vs. full model expected count assessment indicate that the model provides significant explanatory power (correlation coefficient between observed and predicted of 0.97) when compared with the baseline model (correlation coefficient of 0.82), and that the rare event correction provides a reasonable estimate of vegetation event probability at the bay/day level (Figure 4).

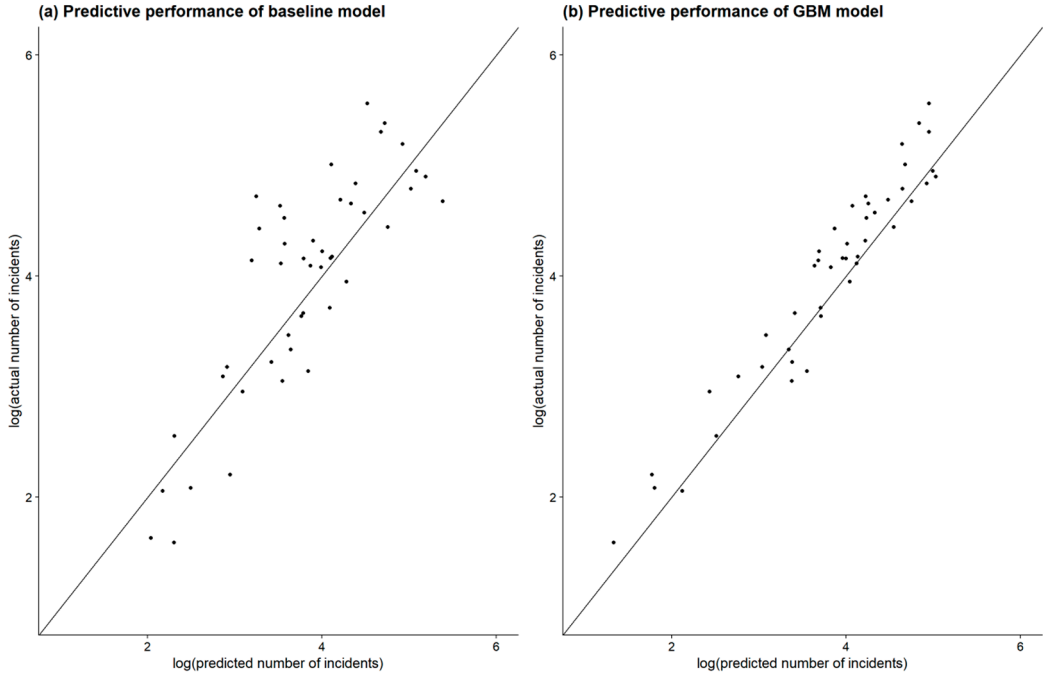

Figure 4. Comparison of the geographically aggregated predictive performance of the baseline vegetation model (**a**) with the implemented GBM vegetation model (**b**) for the test set (4 April 2020 to 4 April 2021). The correlation of predictions against observed incident counts is considerably better for the gbm model (0.97) compared with the baseline model (0.82). The solid line represents the line y = x, which corresponds to the case of perfect correlation between prediction and result.

3.2.2. Ignition Event Model

The AUC recorded for the ignition model was 0.66, which suggests that the model has a reasonable level of performance. This AUC result is similar to the top results achieved by similar models [10], and is a reasonable result given the very small size of the positive class in the test set.

3.3. Ignition Risk Estimation

The expected ignition risk was calculated for each bay/day combination for the 2016–17 fire season (Figure 5). This analysis illustrates the ability of the framework to calculate expected ignition count over arbitrarily defined spatio-temporal windows. The expected ignition risk predictions can be developed for any historical time period (provided weather data is available) and for any future time period (provided reasonable weather forecasts/simulations are available).

The expected ignition risk allows the following analyses:

1. Quantification of ignition risk over arbitrarily defined spatio-temporal regions;
2. Ranking of spatio-temporal groupings with respect to ignition risk;
3. Economic assessments of ignition risk mitigation strategies.

The expected ignition probability at the bay/day level provides a useful modular summary of bushfire risk that can be used for a wide variety of risk assessment applications; however, as it is a point estimate of a complex event distribution it is an overly simplistic measure of the real picture. Consequently, we also developed a bushfire risk simulation framework, the results of which are discussed in the following section.

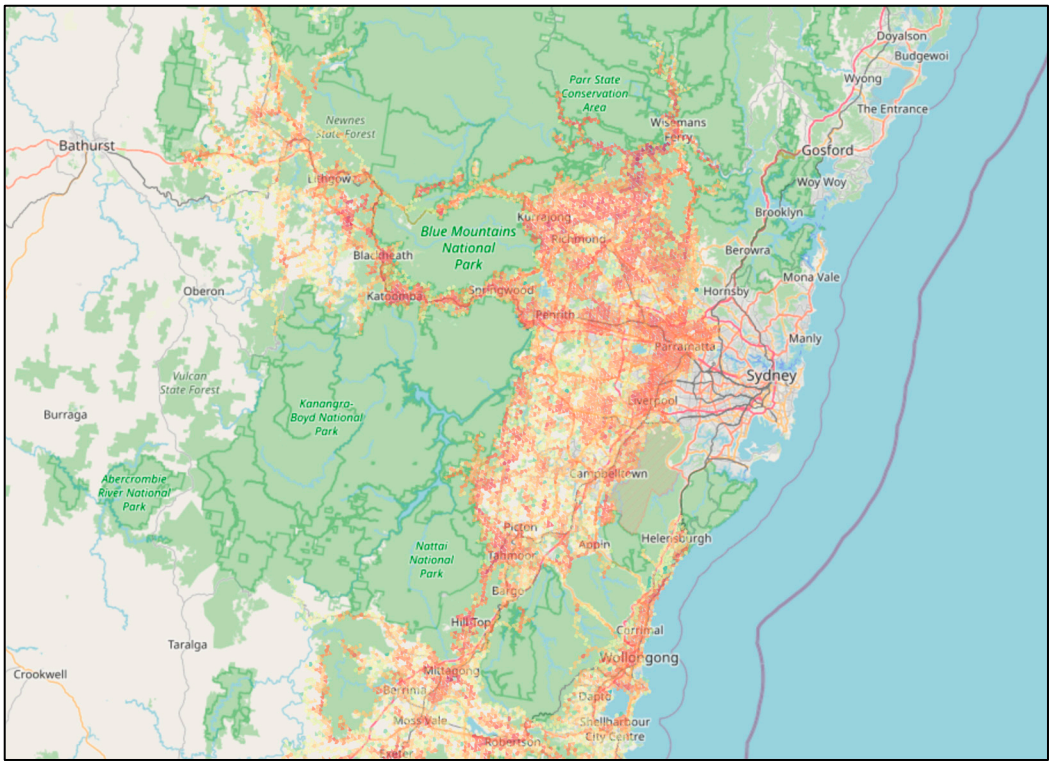

Figure 5. Ignition risk calculated over a 1 km hex grid across the network for the 2016–17 fire season. Darker red areas indicate a higher probability of an ignition event occurring. Only a subset of the network covers areas that are prone to bushfire, and as such, the above map should not be viewed as a bushfire risk map.

3.4. Ignition Risk Simulation

In order to investigate the utility of the ignition risk simulation framework 1000 monte-carlo simulations were run for three fire seasons over all bays in the franchise (Figure 6). The three fire seasons are representative of a mild fire weather season (2020–21), a typical fire weather season (2016–17) and an extreme fire weather season (2019–20).

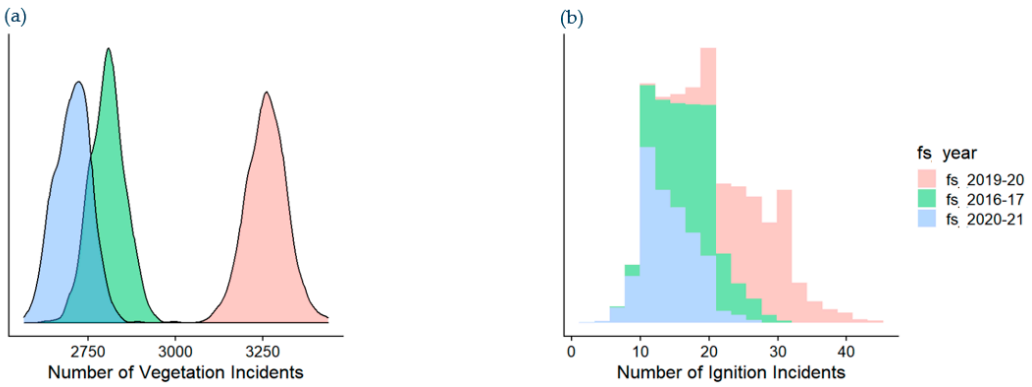

Figure 6. (**a**) Probability density plot of simulation results obtained from 1000 simulations for vegetation events across the entire network for the 2016–17; 2019–20; and 2020–21 fire seasons. (**b**) Stacked bar graph of simulation results obtained from 1000 simulations for vegetation-caused ignition events across the entire network for the 2016–17; 2019–20; and 2020–21 fire seasons.

These simulation results show the possible range of vegetation and ignition events across the network over a given fire season, along with their probability of occurrence. The results are for only 1000 simulations and as such the distribution of rare ignition events is not particularly smooth, a higher number of simulations will result in an improved

distribution. Given sufficient simulation runs, the simulation framework can be used to directly estimate the probability of different outcomes, for example: what is the probability of observing greater than 30 ignition events over the network during an extreme fire season?

4. Discussion

4.1. Modelling Framework

The modelling approach presented in this paper is similar to the optimal approach that was determined by [15], who analysed a set of different algorithms and class imbalance adjustment techniques for a very similar domain problem. This paper extends the work of [15] by modelling the ignition process as a two-step conditional process and through the introduction of a rare event adjustment technique that allows us to recover event probability estimates that are well calibrated to the actual event probabilities on the network. A similar two-step conditional process technique was presented by [12], who adopted the technique due to the fact that the events that may cause ignitions are considerably more frequent than ignitions themselves and they considered that such a process would result in improved ignition rate estimates. Our motivations for adopting the two-step conditional process were similar to those of [12]; however, this study extends the work of [12] by introducing machine learning and statistical techniques that are more appropriate for modelling the complex interactions in weather, landscape and engineering systems that lead up to ignition events.

4.2. Feature Importance

The major factors associated with vegetation events are weather features including wind speed and rain followed by landscape features such as vegetation height and soil type, while network engineering variables appear to have only a small influence. Similar results for the importance of wind speed and vegetation were found in [12,13,15]; however, none of these studies reported an influence of soil type. It could be that soil type influences vegetation events through determining vegetation community composition and the behaviour of trees within those communities during extreme weather events; however, more research is required to further investigate this relationship. The small influence of network engineering variables on the occurrence of vegetation events that was found in [12,15] was also found in this study.

The ignition event model results show that the major factors influencing ignition events are landscape features such as soil type and soil depth, followed by features such as the McArthur's Forest Fire Danger Index (FFDI) and rainfall. As with the vegetation event model, network engineering features have a smaller impact, with bay length being the most influential engineering variable. Interestingly, [15] discovered a strong relationship between ignition probability and feeder length; however, we are not certain enough on the relationship between our paper's definition of bays and the definition of feeder length in [15] to comment further on this correspondence. Further research is required on the relationship between bay length and ignition probability to determine whether the relationship is causal and what are the potential drivers of the relationship. The influence of soil type and soil depth are greater in the ignition model than in the vegetation event model and requires further research to understand the potential causal mechanisms.

While this study utilized sets of features similar to those used in other comparable studies [12,14,15], it is unique in its use of vegetation features derived from LiDAR data. These features contributed significantly to the models' ability to explain both vegetation events and subsequent ignition events with LiDAR derived features accounting for 25% and 16% of the feature importance for vegetation event and ignition event models, respectively.

4.3. Future Research Opportunities

The results show that while modelling rare events in a complex spatio-temporal context is difficult, it is possible to obtain predictive models that perform reasonably well in terms of both rank ordering and event probability estimation. The rare event issues were overcome using a combination of nested conditional modelling and data down-sampling in

combination with a rare-event correction. Importantly, the approach allows the prediction of vegetation and ignition events at a fine spatio-temporal scale that can be aggregated over arbitrarily defined spatio-temporal windows, thereby providing a flexible tool for risk management and planning.

While the methods presented in this paper are reasonably successful in meeting our goals of predictive success and management utility there are several avenues for future research that provide an opportunity for improvement.

4.3.1. Application of Specialised Geospatial Modelling Techniques

The machine learning modelling methods employed were used to quickly establish whether there is adequate signal within the landscape, network and weather data inputs to justify an empirical modelling approach and to determine whether the rare-event correction methods can be successfully used for bushfire modelling. The results indicate that there is significant signal in the model inputs and that the rare-event correction method works well. Consequently, there is a good argument for implementing and testing more specialised geospatial modelling approaches to improve model performance and interpretability. Geospatial modelling methods are able to make use of latent unobserved landscape variables that may significantly improve the models' predictive performance. In addition to exploring geospatial methods, we consider that the exploration of Bayesian methods may prove to be particularly useful as these methods allow us to more easily include parameter estimate uncertainty within the simulation framework.

4.3.2. Application to Future Climate Scenarios

Future climate shifts are expected to result in a significant increase in the frequency of high FFDI days over the Endeavour Energy network. The design of the modelling framework presented in this paper allows us to use climate model simulations as weather inputs and consequently will allow the exploration of ignition risk over a range of climate change scenarios.

4.3.3. Extension of Model Input Features

The range of input variables used in the ignition risk modelling for this project is limited to those that were readily available at the time of modelling. There is considerable scope for increasing the range of input variables, which could potentially result in a significant increase in predictive power.

Author Contributions: Conceptualization, R.W. (Roy Wilson), D.M. and R.W. (Rohan Wickramasuriya); methodology, R.W. (Roy Wilson); software, R.W. (Roy Wilson); validation, R.W. (Roy Wilson), D.M. and R.W. (Rohan Wickramasuriya).; formal analysis, R.W. (Roy Wilson); investigation, R.W. (Roy Wilson) and R.W. (Rohan Wickramasuriya).; resources, D.M.; data curation, R.W. (Roy Wilson), D.M. and R.W. (Rohan Wickramasuriya); writing—original draft preparation, R.W. (Roy Wilson) and R.W. (Rohan Wickramasuriya); writing—review and editing, R.W. (Roy Wilson), D.M. and R.W. (Rohan Wickramasuriya); visualization, R.W. (Roy Wilson); supervision, D.M.; project administration, D.M.; funding acquisition, D.M. All authors have read and agreed to the published version of the manuscript.

Funding: This project was funded by Endeavour Energy.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: Data were sourced from Endeavour Energy's commercial operations and are not publicly available.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Miller, C.; Plucinski, M.; Sullivan, A.; Stephenson, A.; Huston, C.; Charman, K.; Prakash, M.; Dunstall, S. Electrically Caused Wildfires in Victoria, Australia Are over-Represented When Fire Danger Is Elevated. *Landsc. Urban Plan.* **2017**, *167*, 267–274. [CrossRef]
2. Colleen Bryant Understanding Bushfire: Trends in Deliberate Vegetation Fires in Australia. Available online: https://www.aic.gov.au/publications/tbp/tbp27 (accessed on 6 January 2023).
3. Kandanaarachchi, S.; Anantharama, N.; Muñoz, M.A. Early Detection of Vegetation Ignition Due to Powerline Faults. *IEEE Trans. Power Deliv.* **2021**, *36*, 1324–1334. [CrossRef]
4. Collins, K.M.; Penman, T.D.; Price, O.F. Some Wildfire Ignition Causes Pose More Risk of Destroying Houses than Others. *PLoS ONE* **2016**, *11*, e0162083. [CrossRef]
5. Collins, K.M.; Price, O.F.; Penman, T.D.; Collins, K.M.; Price, O.F.; Penman, T.D. Spatial Patterns of Wildfire Ignitions in South-Eastern Australia. *Int. J. Wildland Fire* **2015**, *24*, 1098–1108. [CrossRef]
6. Vazquez, D.A.Z.; Qiu, F.; Fan, N.; Sharp, K. Wildfire Mitigation Plans in Power Systems: A Literature Review. *IEEE Trans. Power Syst.* **2022**, *37*, 3540–3551. [CrossRef]
7. Costafreda, S.; Comas, C.; Vega-Garcia, C. Human-Caused Fire Occurrence Modelling in Perspective: A Review. *Int. J. Wildland Fire* **2017**, *26*, 983. [CrossRef]
8. Jurdao, S.; Chuvieco, E.; Arevalillo, J. Modelling Fire Ignition Probability from Satellite Estimates of Live Fuel Moisture Content. *Fire Ecol.* **2012**, *7*, 77–97. [CrossRef]
9. Massada, A.; Syphard, A.; Stewart, S.; Radeloff, V. Wildfire Ignition-Distribution Modelling: A Comparative Study in the Huron-Manistee National Forest, Michigan, USA. *Int. J. Wildland Fire* **2012**, *22*, 174–183. [CrossRef]
10. Penman, T.; Bradstock, R.; Price, O. Modelling the Determinants of Ignition in the Sydney Basin, Australia: Implications for Future Management. *Int. J. Wildland Fire* **2013**, *22*, 469–478. [CrossRef]
11. Dorph, A.; Marshall, E.; Parkins, K.A.; Penman, T.D. Modelling Ignition Probability for Human-and Lightning-Caused Wildfires in Victoria, Australia. *Nat. Hazards Earth Syst. Sci. Discuss.* **2022**, *22*, 3487–3499. [CrossRef]
12. Dunstall, S.; Towns, G.; Huston, C.; Stephenson, A. *PBSP Risk Reduction Model: Overview and Technical Details*; CSIRO Data61: Sydney, Australia, 2016.
13. Mitchell, J.W. Power Line Failures and Catastrophic Wildfires under Extreme Weather Conditions. *Eng. Fail. Anal.* **2013**, *35*, 726–735. [CrossRef]
14. Malik, A.; Rao, M.R.; Puppala, N.; Koouri, P.; Thota, V.A.K.; Liu, Q.; Chiao, S.; Gao, J. Data-Driven Wildfire Risk Prediction in Northern California. *Atmosphere* **2021**, *12*, 109. [CrossRef]
15. Yao, M.; Bharadwaj, M.; Zhang, Z.; Jin, B.; Callaway, D.S. Predicting Electricity Infrastructure Induced Wildfire Risk in California. *Environ. Res. Lett.* **2022**, *17*, 094035. [CrossRef]
16. Nolan, R.H.; Bowman, D.M.J.S.; Clarke, H.; Haynes, K.; Ooi, M.K.J.; Price, O.F.; Williamson, G.J.; Whittaker, J.; Bedward, M.; Boer, M.M.; et al. What Do the Australian Black Summer Fires Signify for the Global Fire Crisis? *Fire* **2021**, *4*, 97. [CrossRef]
17. Dowdy, A.; Mills, G.; Mills, G.; Mills, G. *Australian Fire Weather as Represented by the McArthur Forest Fire Danger Index and the Canadian Forest Fire Weather Index*; Centre for Australian Weather and Climate Research: Melbourne, Australia, 2009.
18. Keetch, J.J.; Byram, G.M. *A Drought Index for Forest Fire Control*; US Department of Agriculture, Forest Service, Southeastern Forest Experiment: Asheville, NC, USA, 1968; Volume 38.
19. Alexander, M.E. Computer Calculation of the Keetch-Byram Drought Index-Programmers Beware. *Fire Manag. Notes* **1990**, *51*, 23–25.
20. King, G.; Zeng, L. Explaining Rare Events in International Relations. *Int. Organ.* **2000**, *55*, 693–715. [CrossRef]
21. Landau, W. The Targets R Package: A Dynamic Make-like Function-Oriented Pipeline Toolkit for Reproducibility and High-Performance Computing. *JOSS* **2021**, *6*, 2959. [CrossRef]
22. Friedman, J.H. Stochastic Gradient Boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378. [CrossRef]