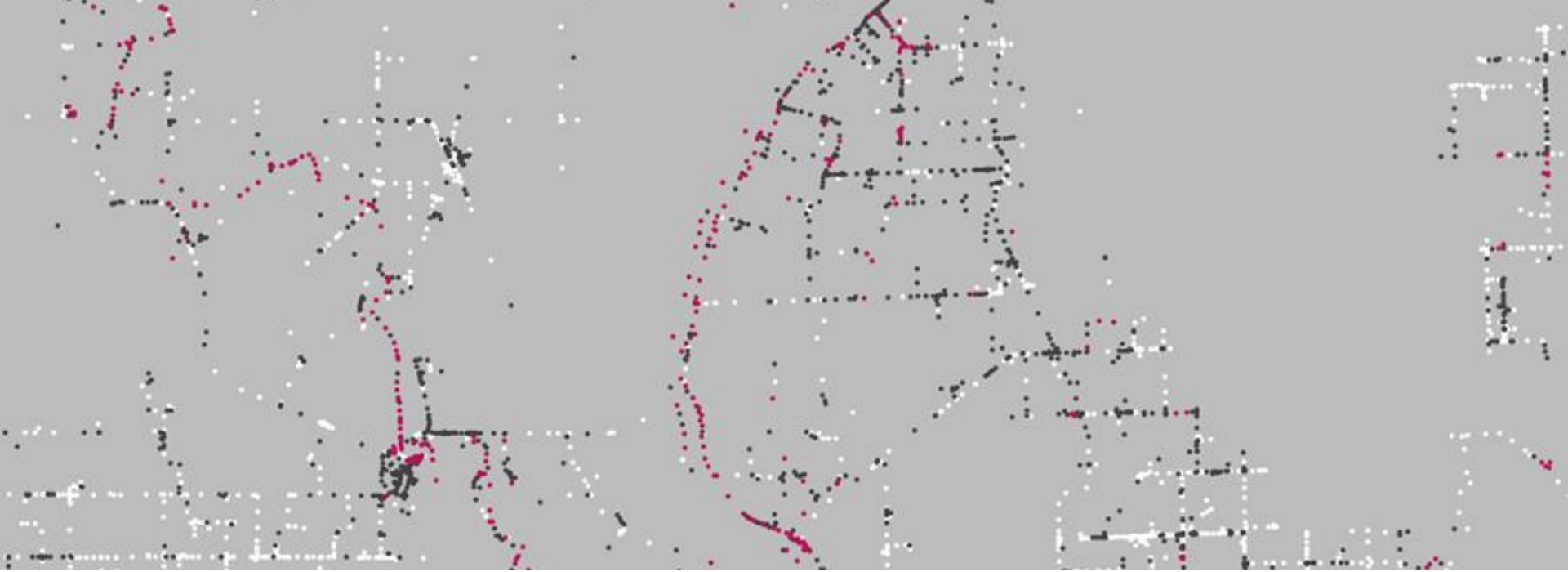


Predicting annual vegetation-driven ignition probabilities using a spatial model
a maximum entropy analysis

Feb. 13, 2020
For PG&E internal use

Convergence Data Analytics, Salo Sciences, Presence Product Group

Presentation overview

- Background material
- Stating the problem
- How we approached the problem
  - An introduction to the MaxEnt modeling approach
  - Model performance metrics
- Input data
- Model results
- Early interpretation
- Next steps

Presentation overview

- Background material
- Stating the problem
- How we approached the problem
  - An introduction to the MaxEnt modeling approach
  - Model performance metrics
- Input data
- Model results
- Early interpretation
- Next steps

Context – Why are we undertaking this initiative?

In support of the new EO Risk Paradigm, PG&E is developing a Distribution (Dx) Asset Risk Model (the Model), tuned for Wildfire Risk, which will:

- Provide situational awareness of the current wildfire risk on the Dx system
- Enable risk-informed decision making in the budget planning process
- Allow PG&E to report risk reduction to regulatory entities

Note: This project will be an input into and is proceeding in coordination with ongoing PRA risk modeling and will be validated by and is expected to be an input into EORM's process.

Phase 1 key objectives and desired outcomes (end of March 2020)

A Prototype Model has been developed for one or more Dx asset classes such that:

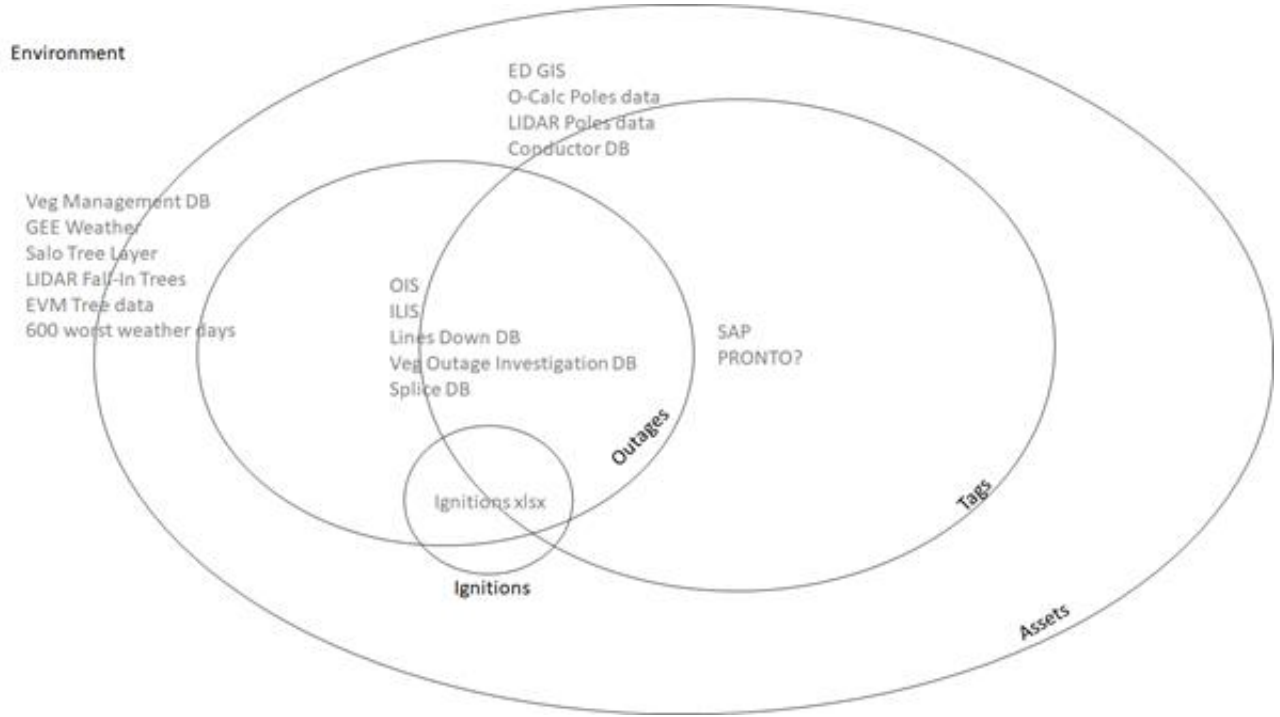
- Statistical experts within PG&E verify that the Model is developed on a solid statistical foundation
- Risk calculation methodology has been approved by EORM
- Prototype results are used to inform the Q1 Dx asset planning budget adjustments.
- The Prototype will only consider Probability of Failure and Wildfire Risk
- MAVF and other components of asset risk will be included in Phase II

Phase 1 project schedule (3 milestones)

Note that project work commenced in Nov, 2019 - not shown on this chart for clarity

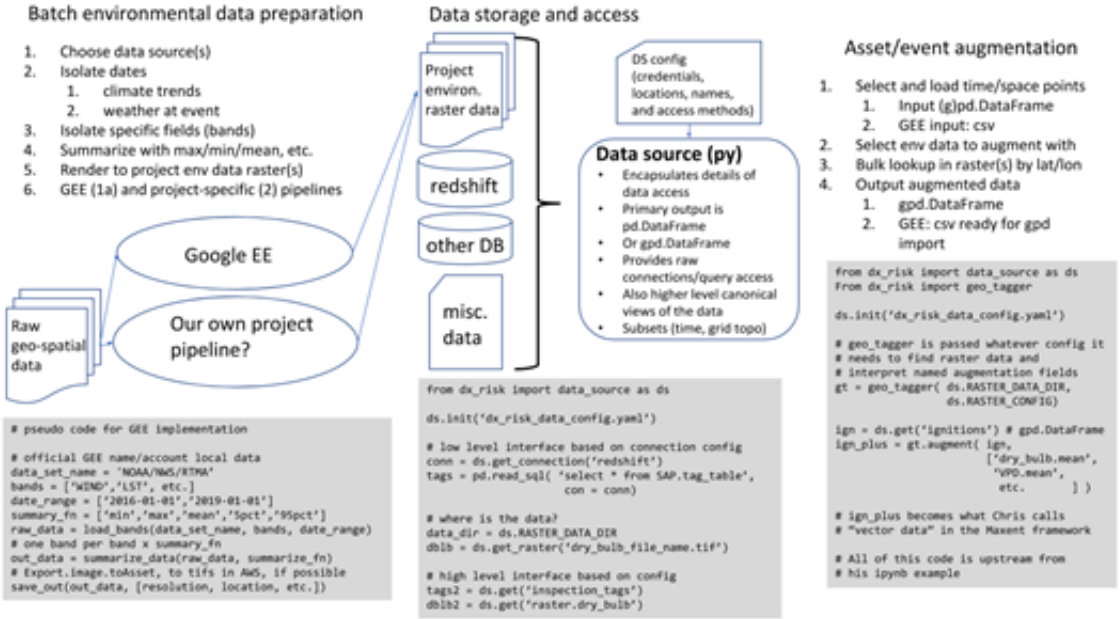
Where we are now (1/3)

- Requirements gathering and data research
  - 17+ meetings over 2 months with key stakeholders
  - Comprehensive catalog of relevant data sets, written documents defining the modeling problem(s) and related approaches and tradeoffs



Where we are now (2/3)

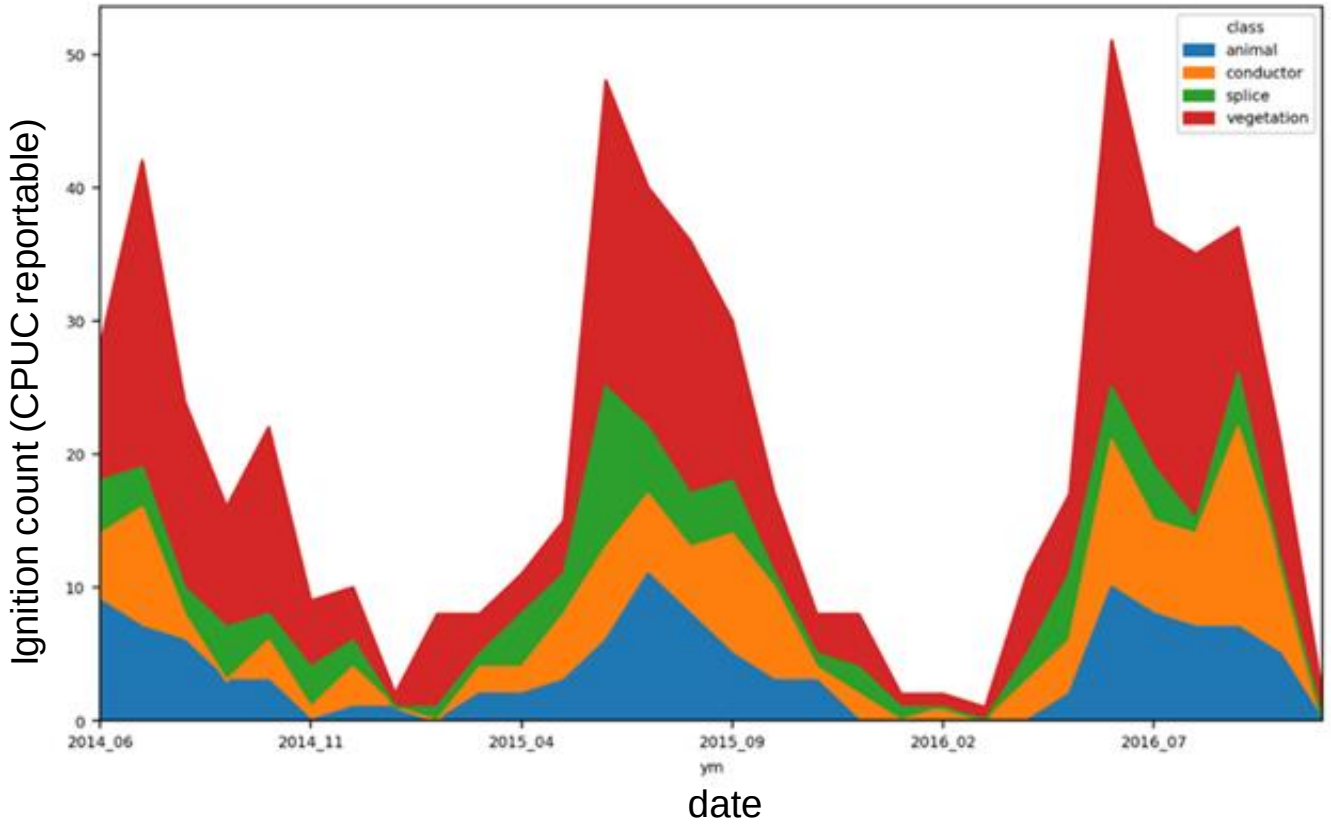
- Infrastructure - modeling in software
  - Pipeline for gathering and formatting geo-spatial data
  - Pipeline for augmenting any location (by lat/lon) with geo-spatial information
    - Ignition sites, Dx grid, etc.
  - Software system to prepare data for, configure, execute, and post-process modeling runs



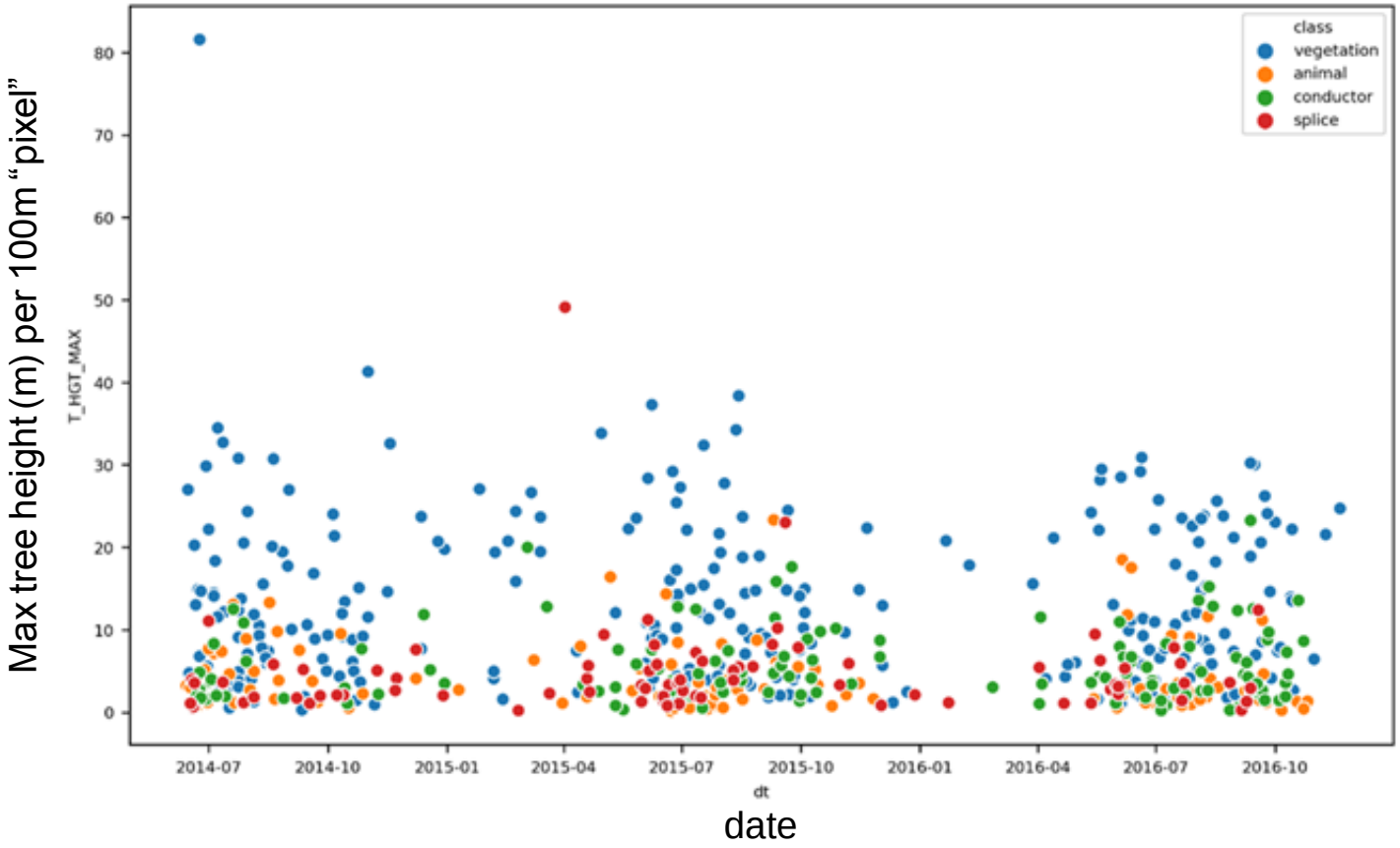
Where we are now (3/3)

- Infrastructure - cloud-based data science environment at PG&E
  - jupyter/python/geopandas/rasterio
  - AWS SageMaker environment - same platform as ARAD
  - Team members have access to: PG&E private data, collaboration tools, source code repositories, etc
- **Backstop model**

Model progress - Seasonal frequency of ignitions by ignition cause class



Model progress - Ignitions by tree height, date, class



Presentation overview

- Background material
- Stating the problem
- How we approached the problem
  - An introduction to the MaxEnt modeling approach
  - Model performance metrics
- Input data
- Model results
- Early interpretation
- Next steps

The Distribution (Dx) Asset Risk Model
Tuned for wildfire

wildfire risk = probability of ignition · consequences of fire spread

Goals

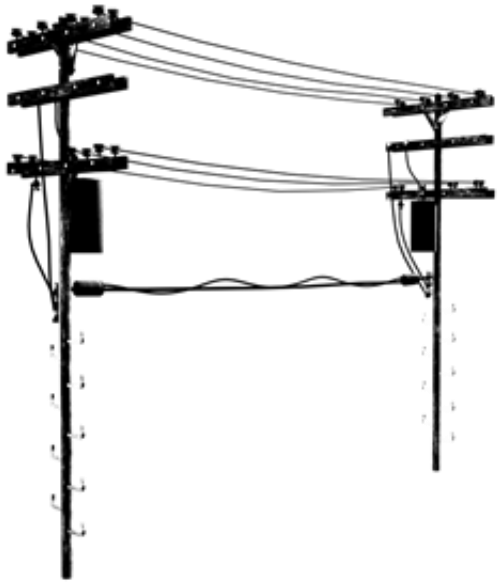
- Calculate the probability of ignition for each Dx asset
- Understand the environmental drivers of ignition probabilities
- Identify high probability ignition locations for mitigation

Constraints

- Limited to just vegetation contact ignitions
- Currently using publicly-accessible data
- Assessing over aggregate (i.e. yearly) timescales

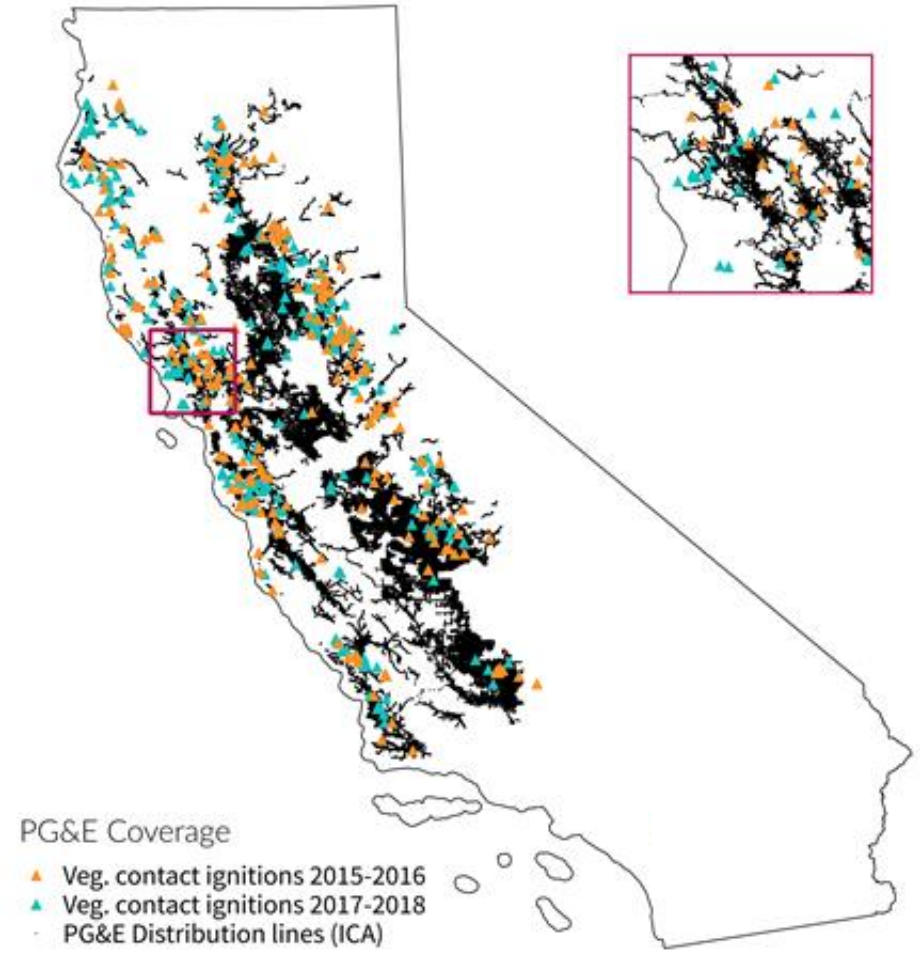
Guiding questions

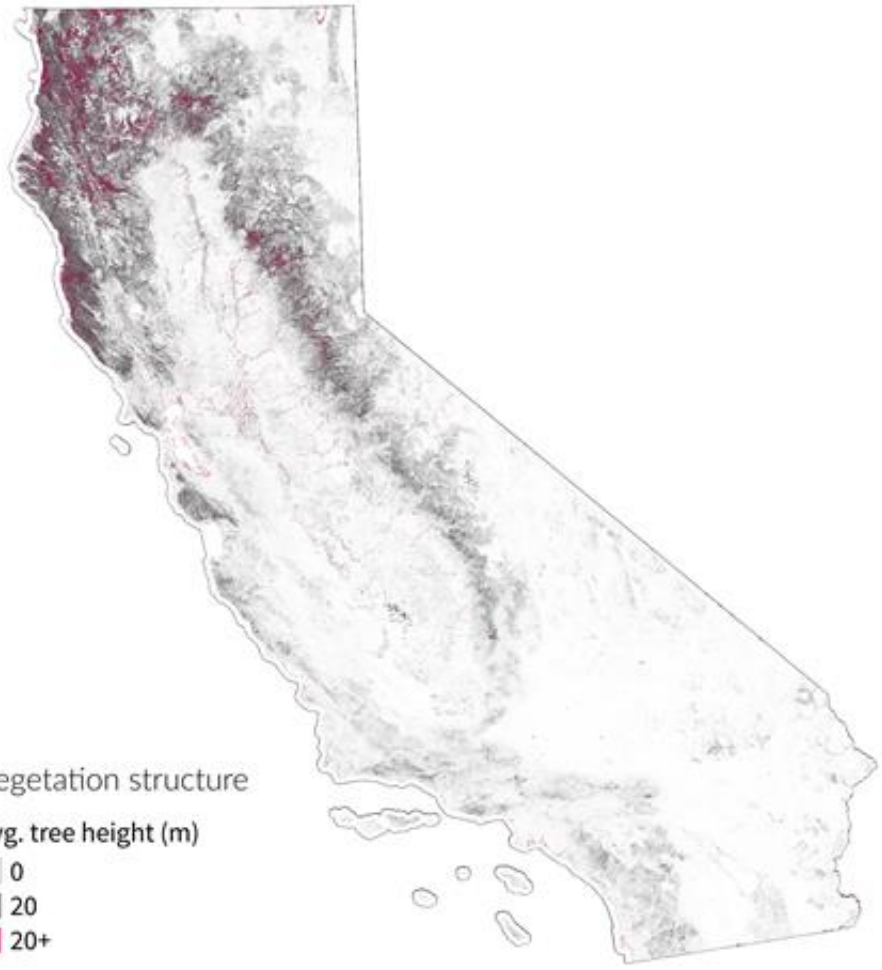
- How many distribution assets are susceptible to vegetation contact-driven ignitions?
- What environmental conditions are most likely to lead to vegetation contact ignitions?
- Which assets are the most likely to experience a vegetation contact event that leads to an ignition?

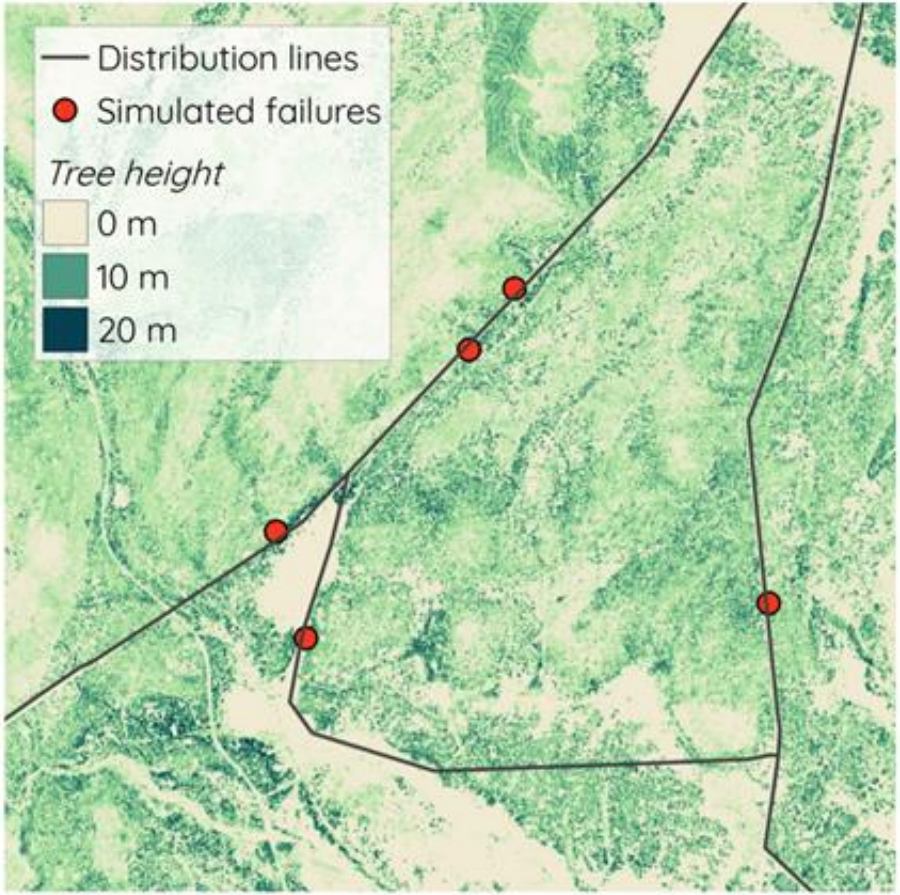


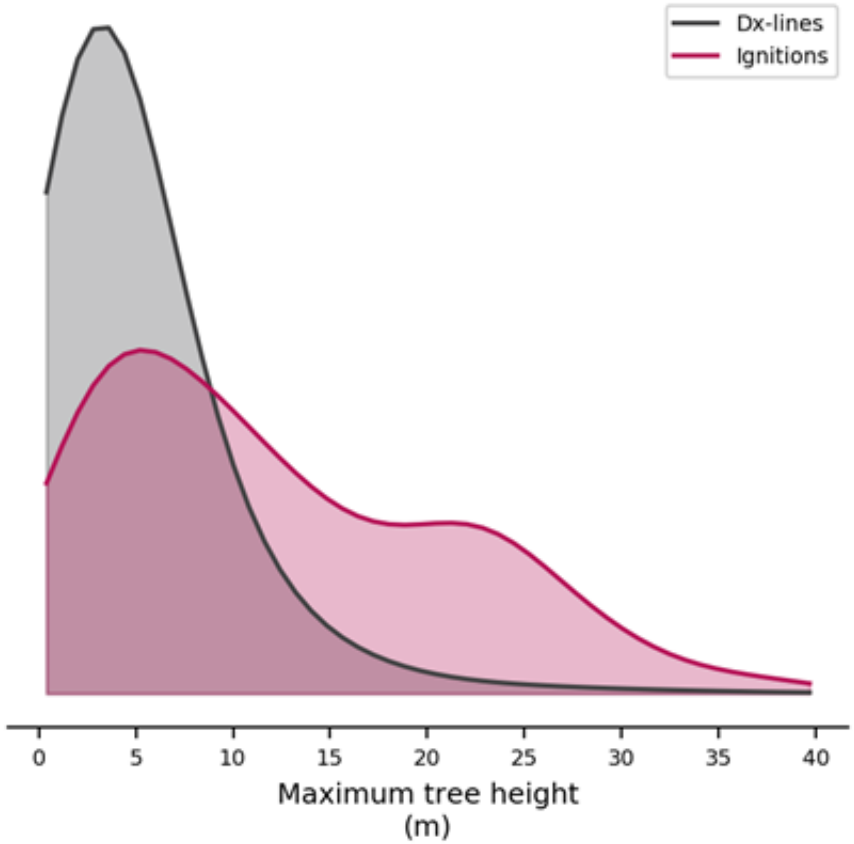
Presentation overview

- Background material
- Stating the problem
- How we approached the problem
  - An introduction to the MaxEnt modeling approach
  - Model performance metrics
- Input data
- Model results
- Early interpretation
- Next steps


PG&E Coverage
Veg. contact ignitions 2015-2016
Veg. contact ignitions 2017-2018
PG&E Distribution lines (ICA)


Vegetation structure
Avg. tree height (m)
0
20
20+



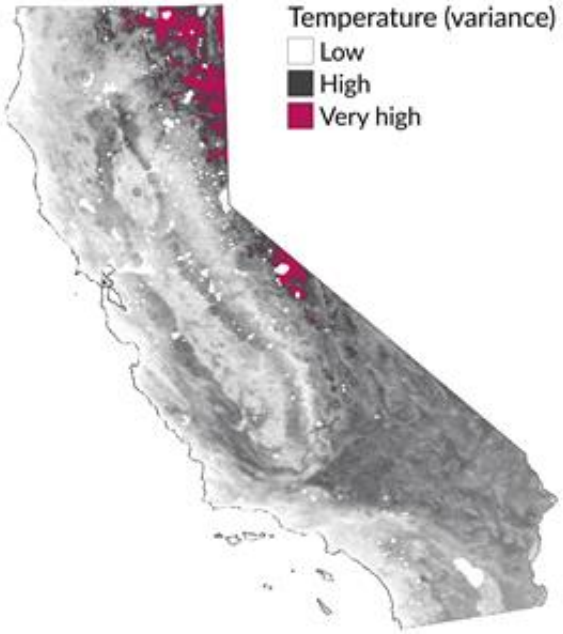


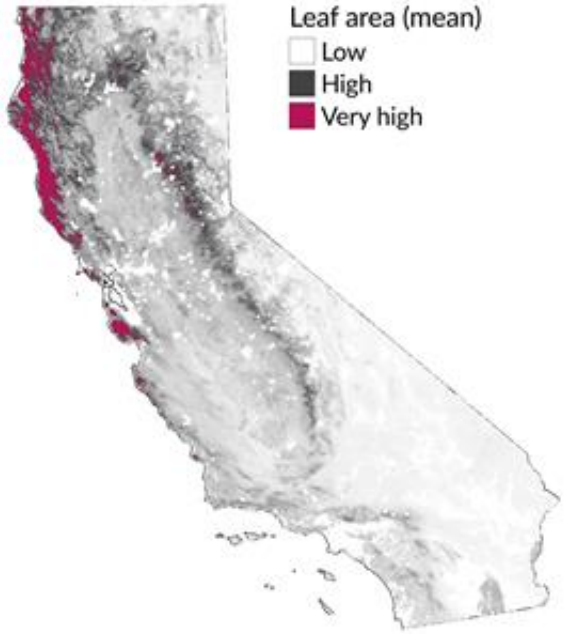
MaxEnt - a presence/background method for calculating ignition probability

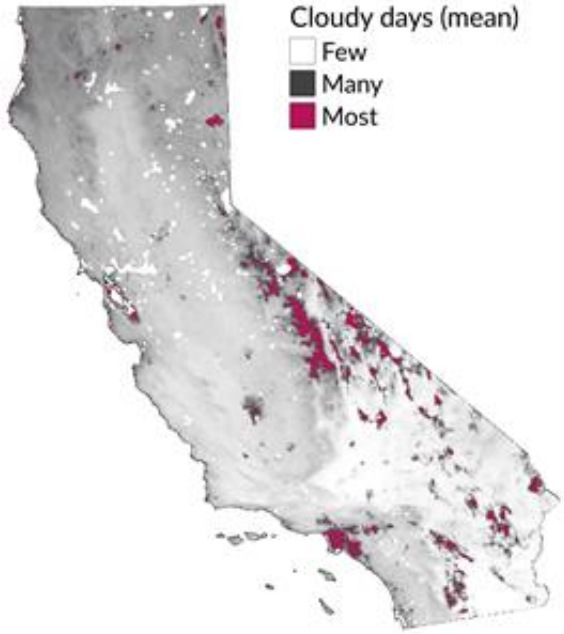
Consider the slug - an introduction to MaxEnt

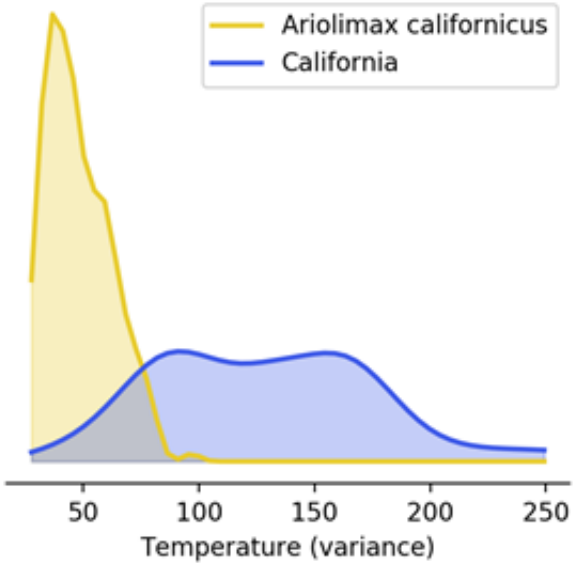
Ariolimax californicus

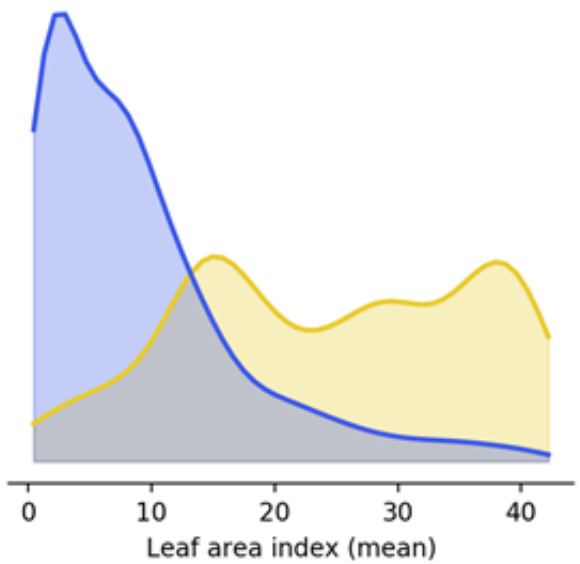
289 observation records

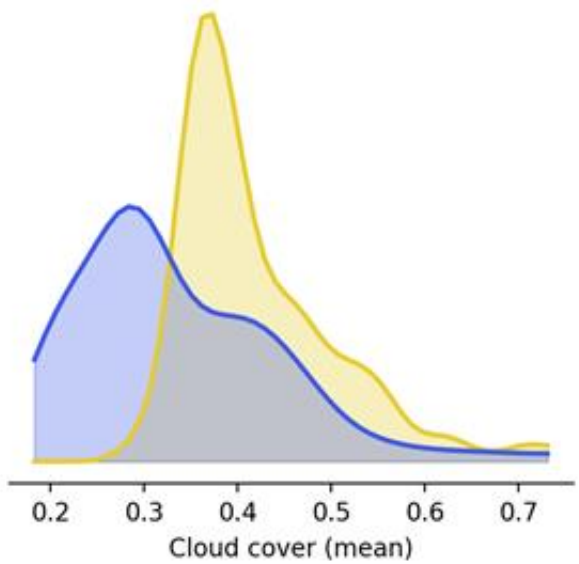

Temperature (variance)
Low
High
Very high

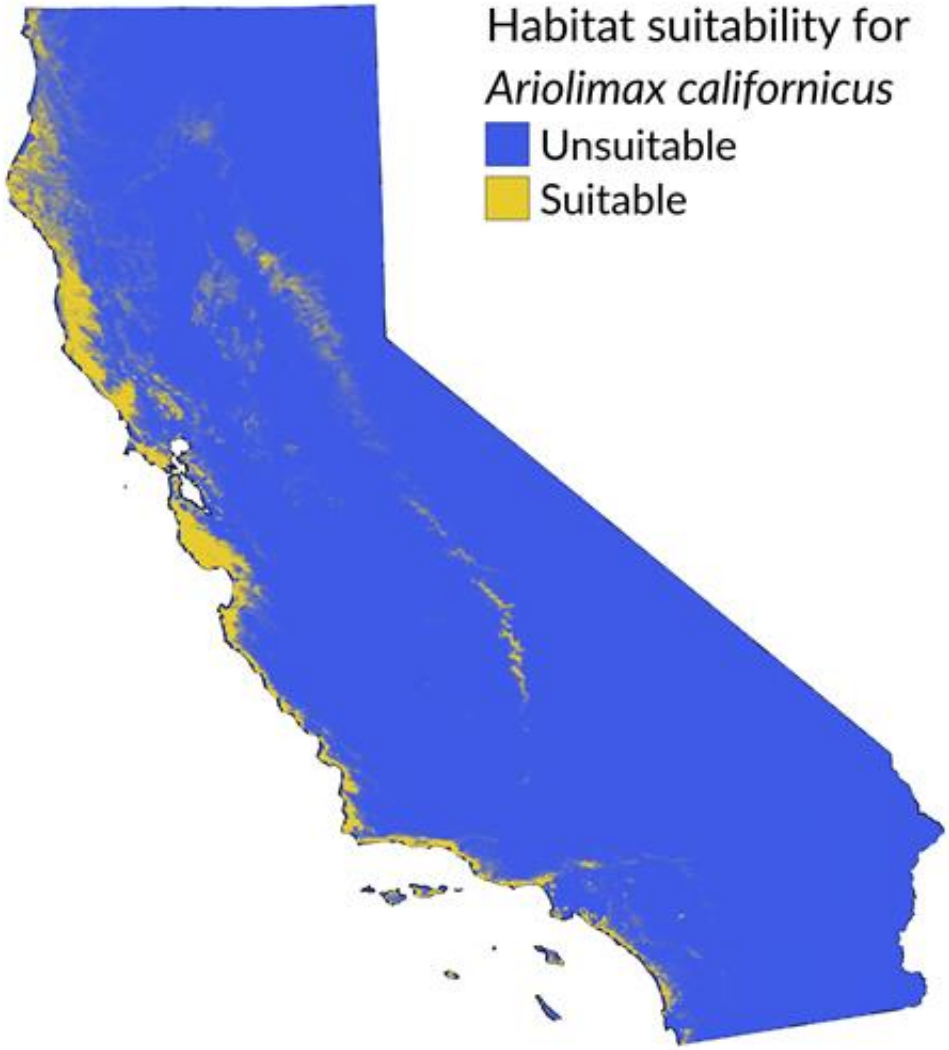

Leaf area (mean)
Low
High
Very high


Cloudy days (mean)
Few
Many
Most

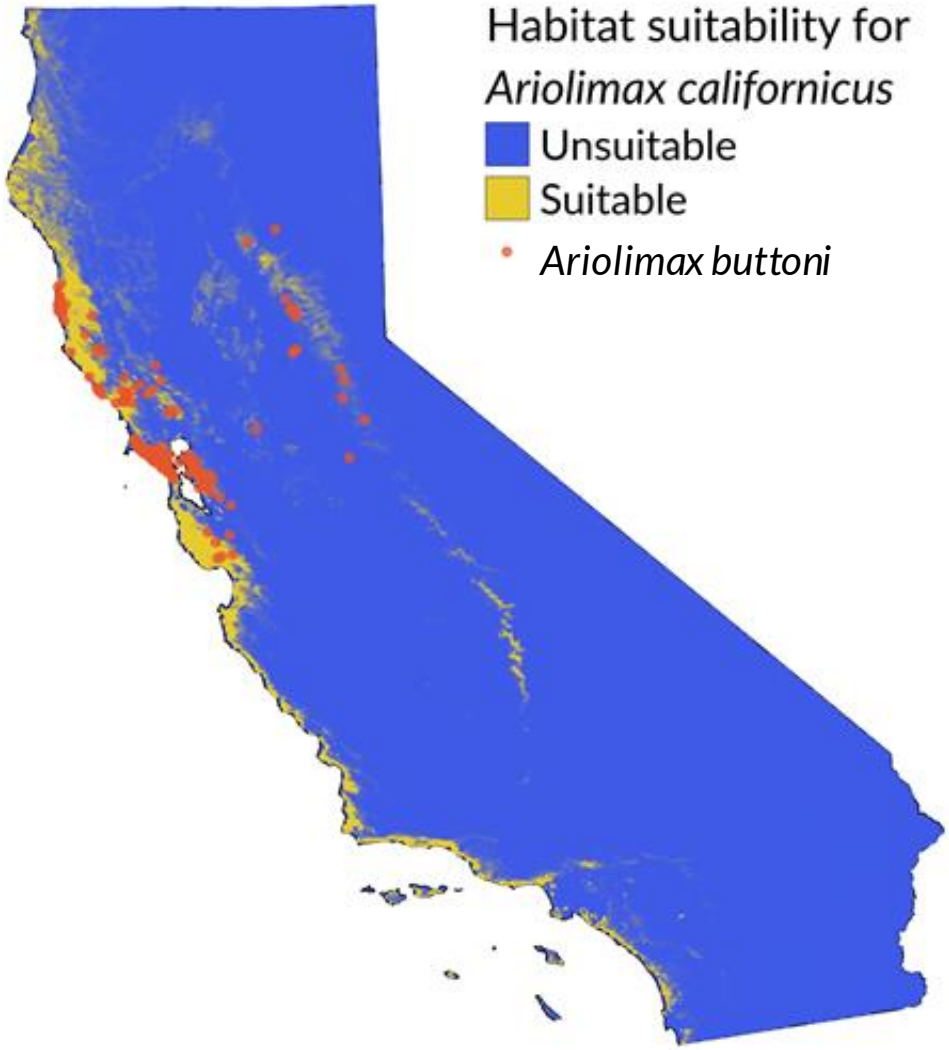

Ariolimax californicus
California
50
100
150
200
250
Temperature (variance)


0
10
20
30
40
Leaf area index (mean)


0.2
0.3
0.4
0.5
0.6
0.7
Cloud cover (mean)



Variable	Percent contribution	Permutation importance
Temperature (variance)	72.1	76
Cloud cover (mean)	11.4	7.4
Temperature (mean)	10.4	15.2
Leaf area (variance)	5.3	0.3
Leaf area (mean)	0.9	1.2



Variable	Percent contribution	Permutation importance
Temperature (variance)	72.1	76
Cloud cover (mean)	11.4	7.4
Temperature (mean)	10.4	15.2
Leaf area (variance)	5.3	0.3
Leaf area (mean)	0.9	1.2

Model performance metrics

Recall

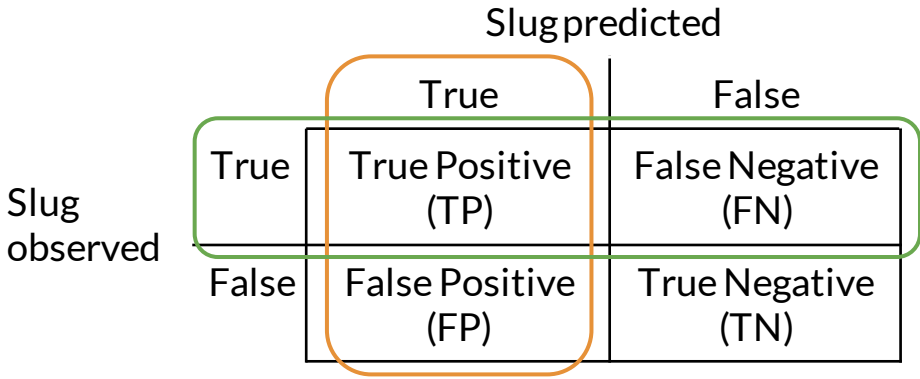
- True positive rate
- Quantifies ability of a model to find all observations within a dataset

AUC

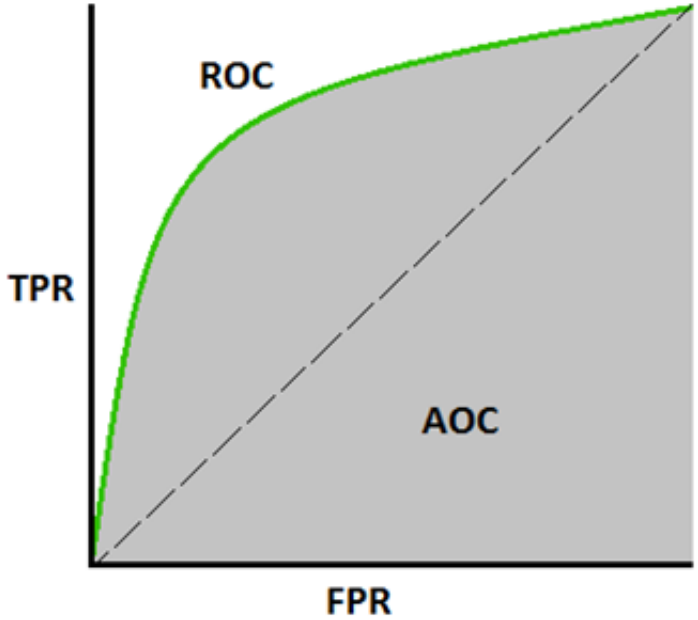
- Metric of separability
- Tracks the true positive and false positive rates

Precision

- Positive discrimination rate
- Quantifies ability of a model to discriminate between observed and



Unpacking AUC

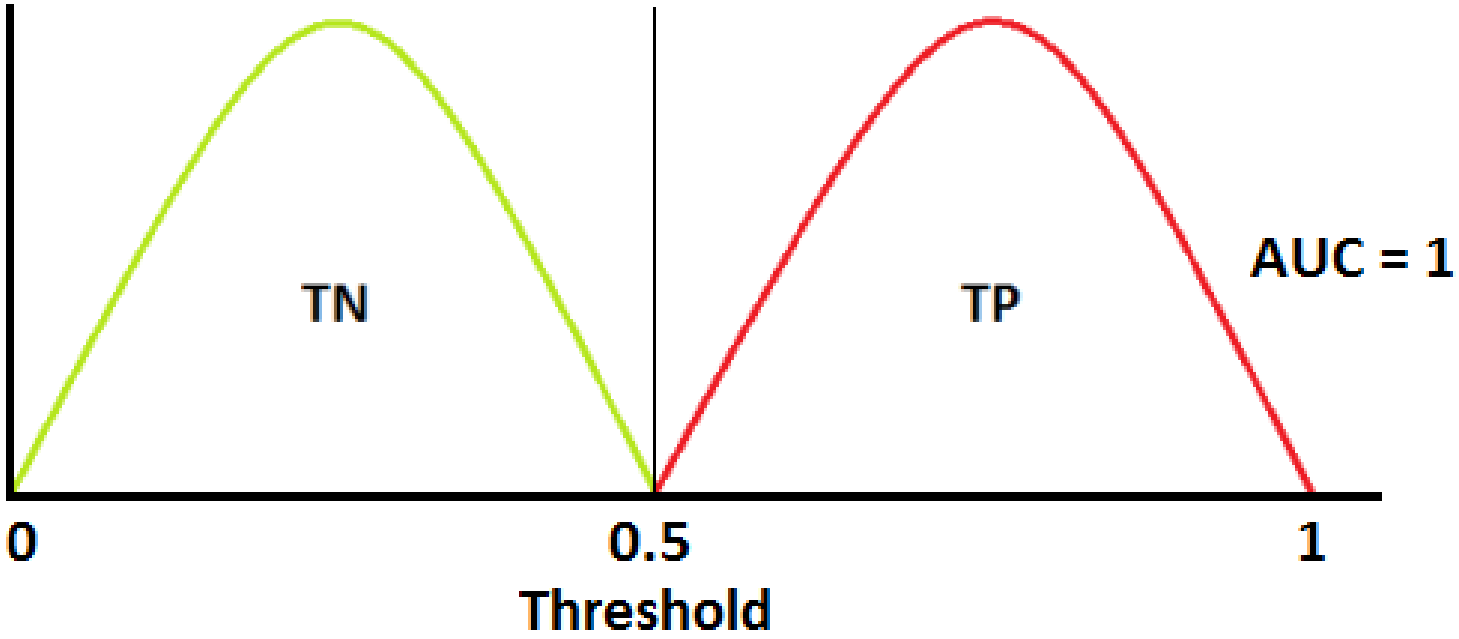


$$\text{TPR} = \frac{TP}{TP + FN}$$

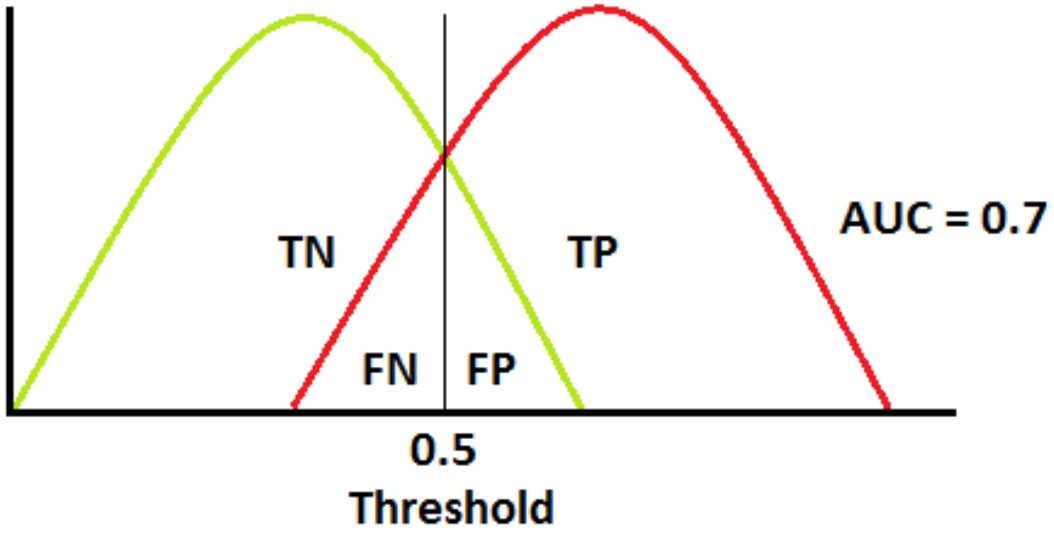
$$\text{FPR} = \frac{FP}{TN + FP}$$

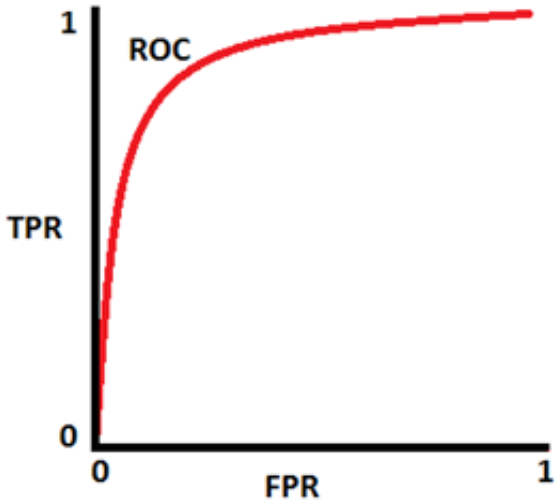
		Predicted: True	Predicted: False
Observed	True	True Positive (TP)	False Negative (FN)
Observed	False	False Positive (FP)	True Negative (TN)

https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5



https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

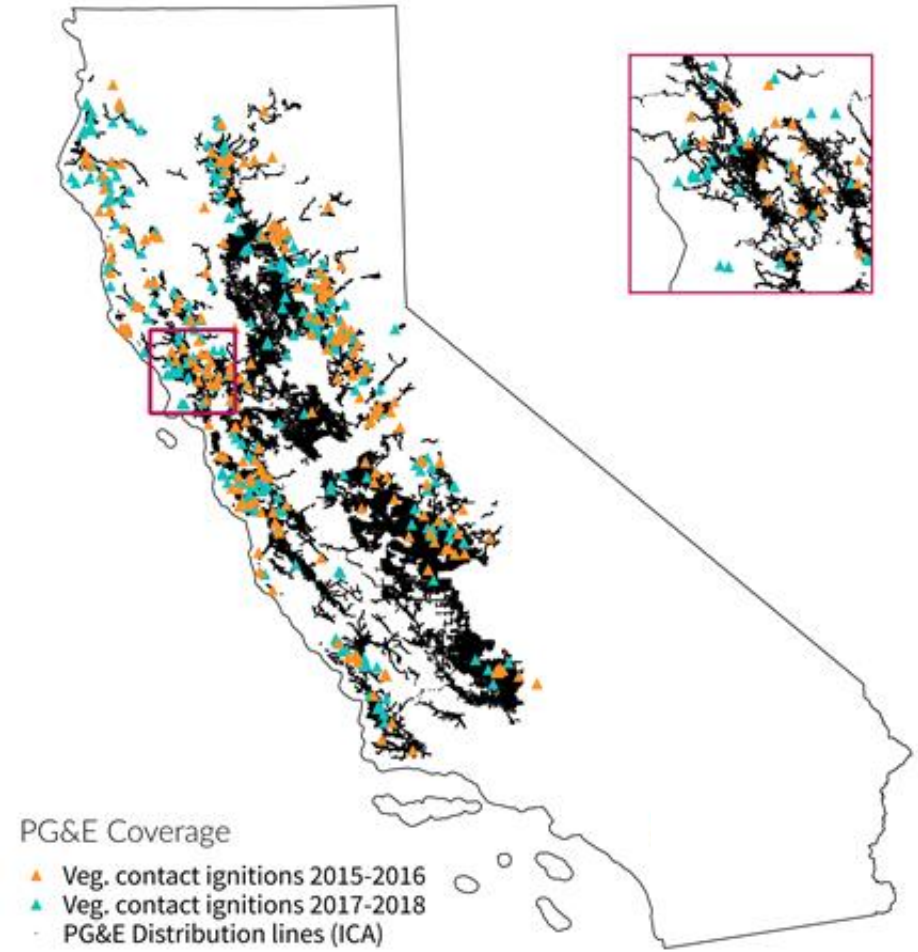




https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Presentation overview

- Background material
- Stating the problem
- How we approached the problem
  - An introduction to the MaxEnt modeling approach
  - Model performance metrics
- Input data
- Model results
- Early interpretation
- Next steps



Ignition locations

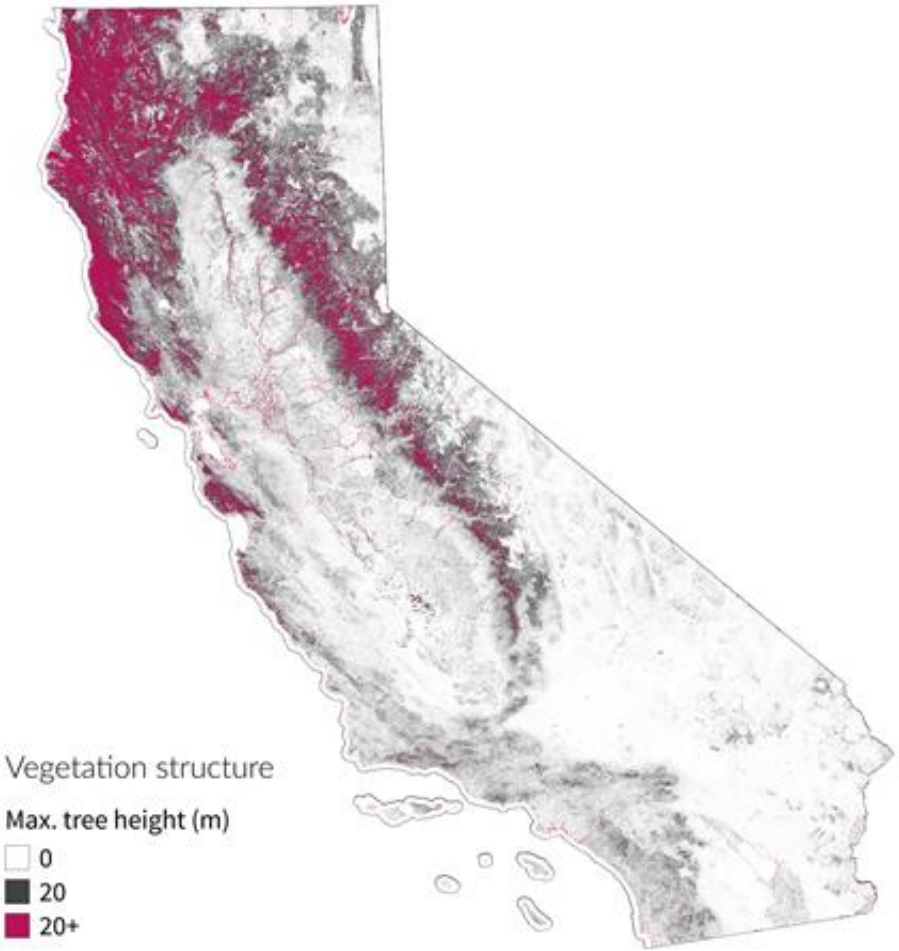
- 2015-2016 ignitions
- 210 points

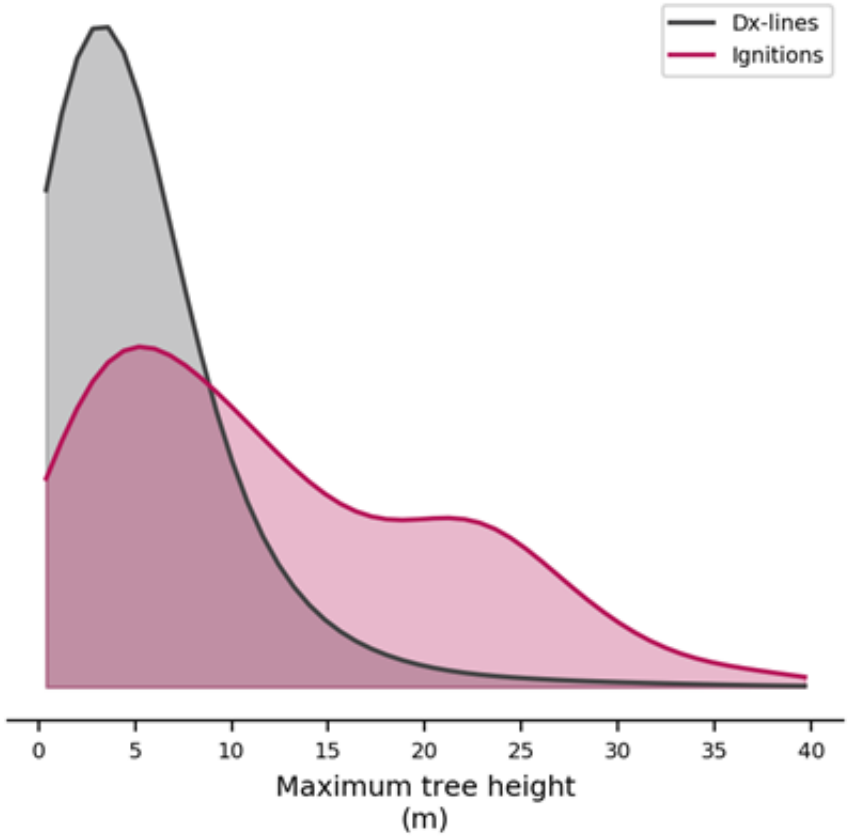
Environmental covariates

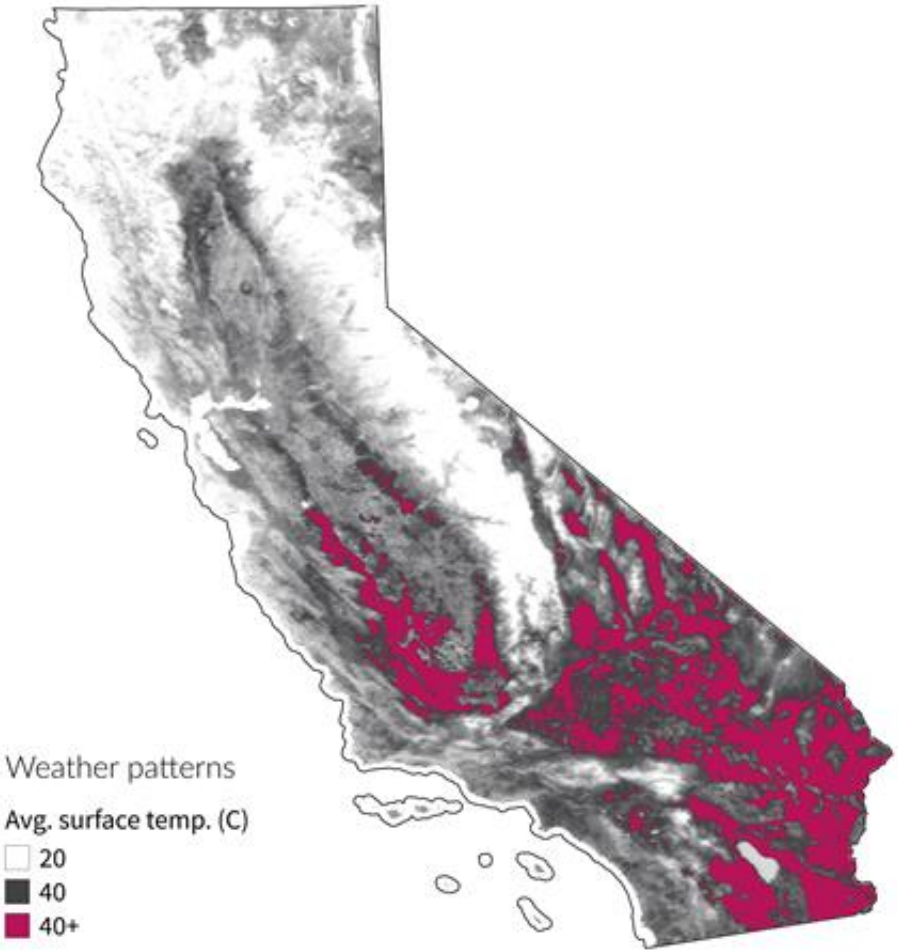
- Vegetation, wind speeds, gust speeds, temperature, topography
- 10 covariates

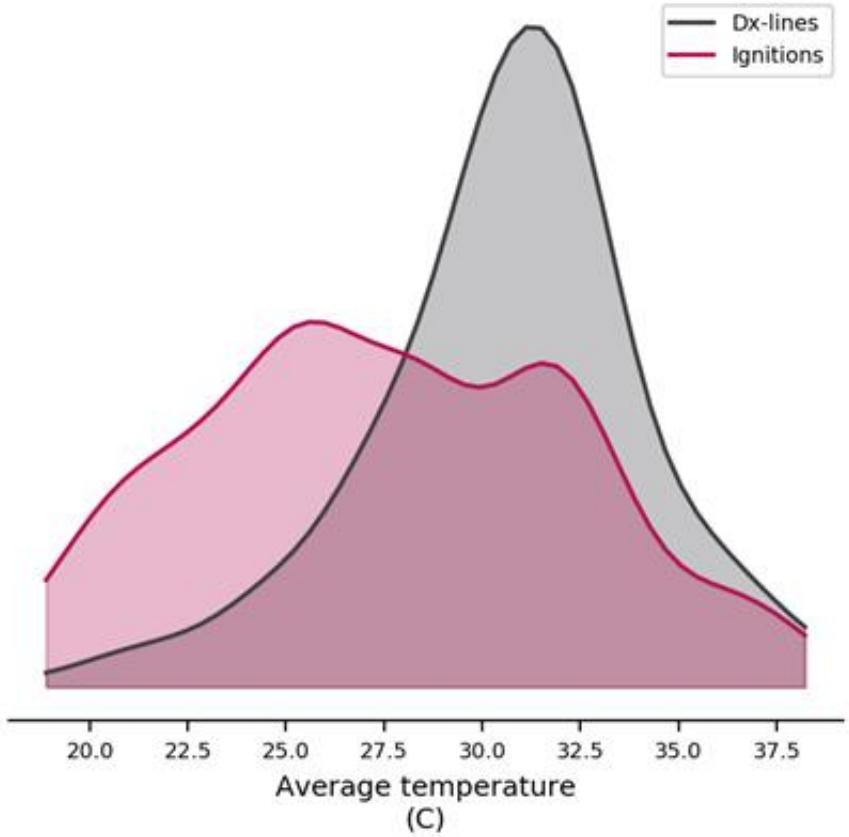
Test locations

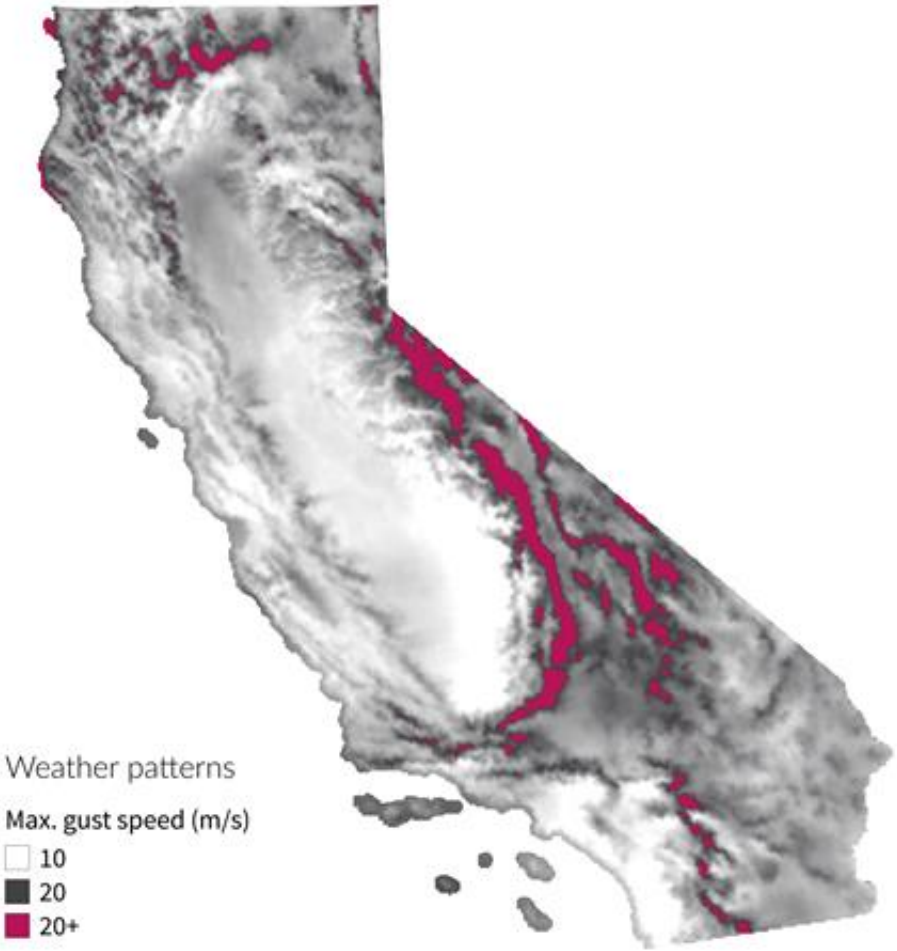
- 2017-2018 ignitions
- 266 points

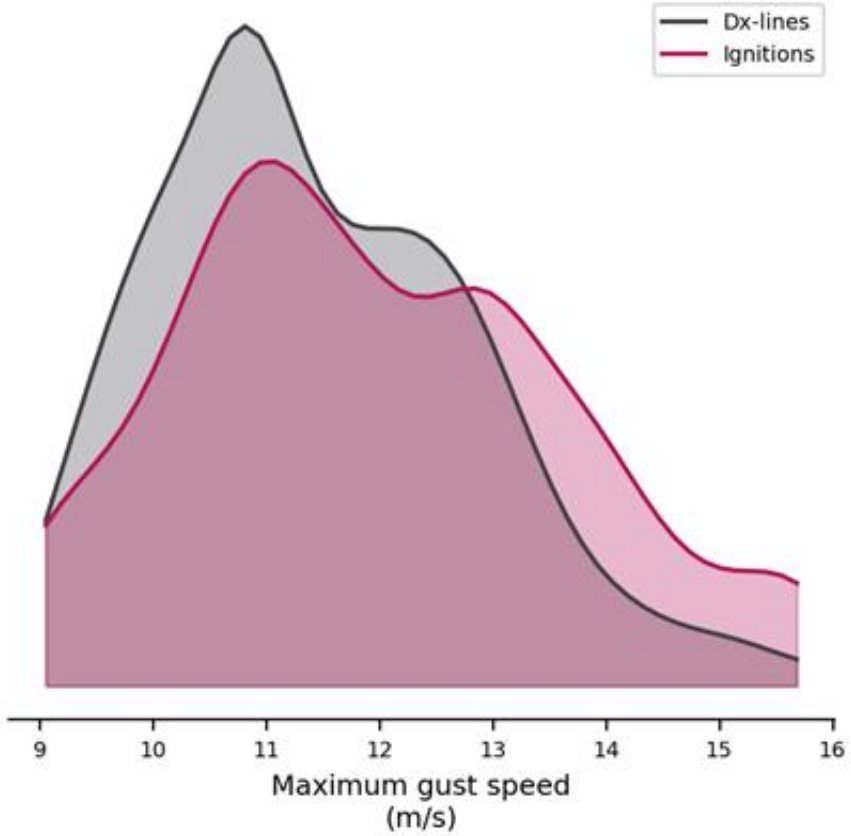

Vegetation structure
Max. tree height (m)
0
20
20+


Dx-lines
Ignitions
0
5
10
15
20
25
30
35
40
Maximum tree height
(m)


Weather patterns
Avg. surface temp. (C)
20
40
40+


Dx-lines
Ignitions
20.0
22.5
25.0
27.5
30.0
32.5
35.0
37.5
Average temperature
(C)


Weather patterns
Max. gust speed (m/s)
10
20
20+


Dx-lines
Ignitions
9
10
11
12
13
14
15
16
Maximum gust speed
(m/s)

Class	Covariate	Unit	Spatial scale	Notes
Vegetation	Mean tree height	(m)	100 m	Mean tree height of area around asset
	Tallest nearby trees	(m)	100 m	Calculated as maximum tree height in area around an asset
Wind	Mean wind speed	(m/s)	2,500 m	From RTMA
	Local wind speed maximum	(m/s)	2,500 m	Calculated as the 99th percentile of local wind speeds
Gust	Mean gust speed	(m/s)	2,500 m	From RTMA
	Local gust speed maximum	(m/s)	2,500 m	Calculated as the 99th percentile of local gust speeds
Temperature	Mean temperature	(°C)	1,000 m	From MODIS LST
	Local temperature maximum	(°C)	1,000 m	Calculated as the 99th percentile of local temperatures
Topography	Local topographic position	unitless	100 m	From the topographic position index (TPI)
	Landscape topographic position	unitless	1,000 m	Calculating TPI at fine and large scales allows distinguishing multiple landforms (i.e. difference in local and landscape topography)

Model outputs

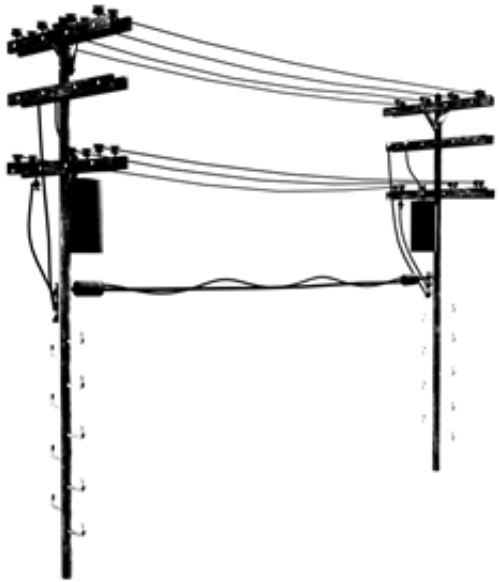
1. Relative probability scores
    - Units: arbitrary
    - Computes ignition probability for each asset using raw probability distributions
    - Evaluated using AUC scores
2. Omission rates
    - Units: %
    - Scales relative probability scores based on the total area evaluated
    - Can threshold rates to evaluate likely/unlikely in binary sense
    - Threshold set to > 5%
    - Evaluated using recall scores
3. Occurrence probability scores
    - Units: %
    - Scales relative probability scores to probability of ignition scores via logistic transformation of raw scores
    - Done via scaling parameter, τ, (the probability of ignition at 'average' ignition locations)
    - τ calculated as (number of total ignitions) / (number of Dx assets evaluated)

Presentation overview

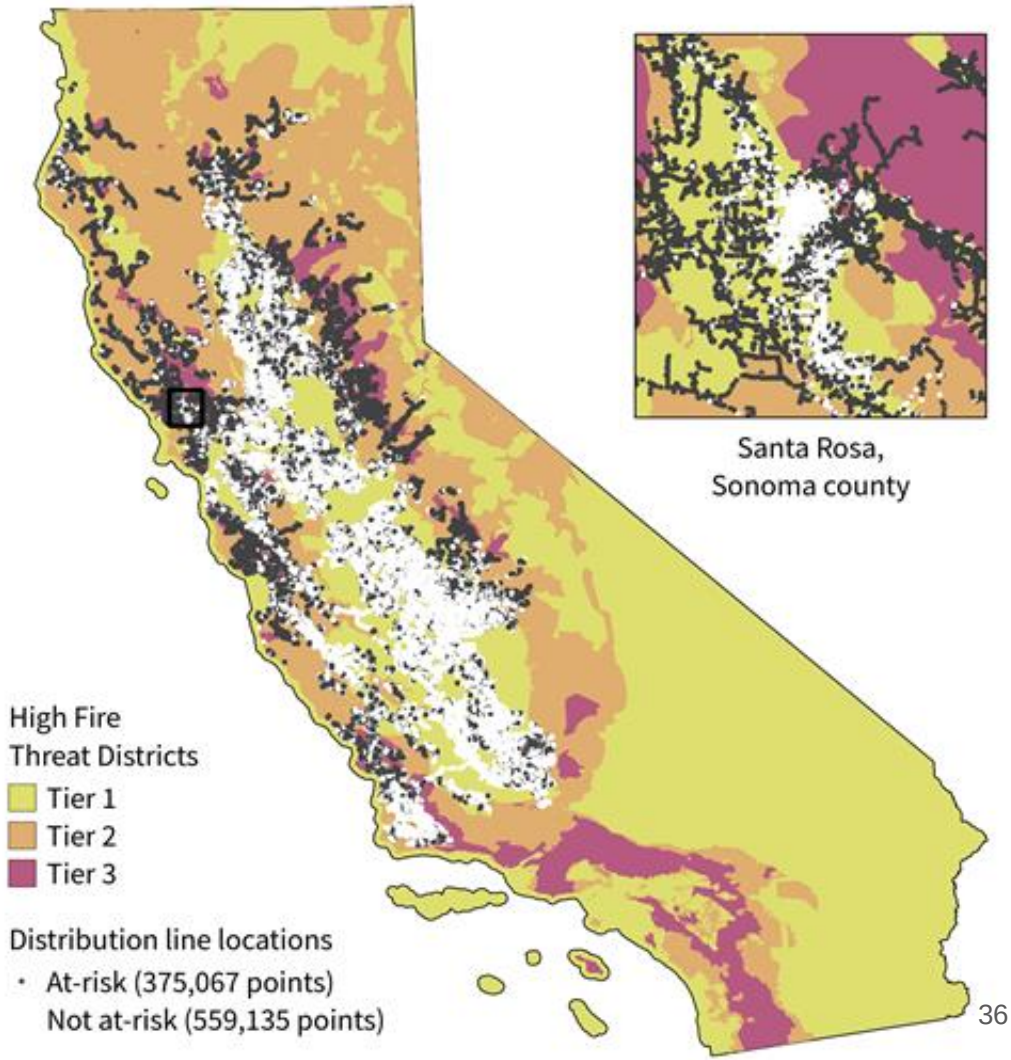
- Background material
- Stating the problem
- How we approached the problem
  - An introduction to the MaxEnt modeling approach
  - Model performance metrics
- Input data
- Model results
- Early interpretation
- Next steps

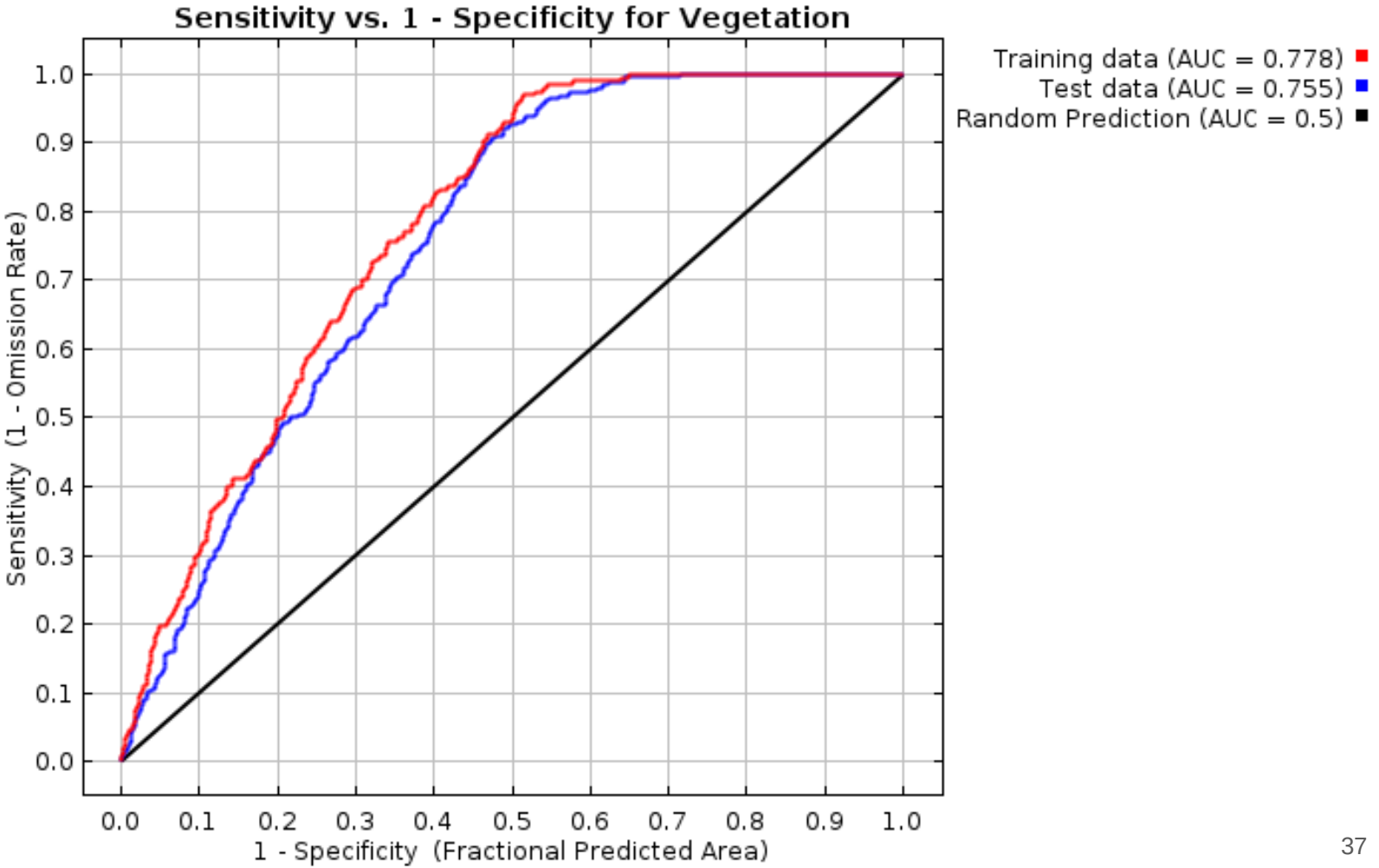
Guiding questions

- How many distribution assets are susceptible to vegetation contact-driven ignitions?
- What environmental conditions are most likely to lead to vegetation contact ignitions?
- Which assets are the most likely to experience a vegetation contact event that leads to an ignition?

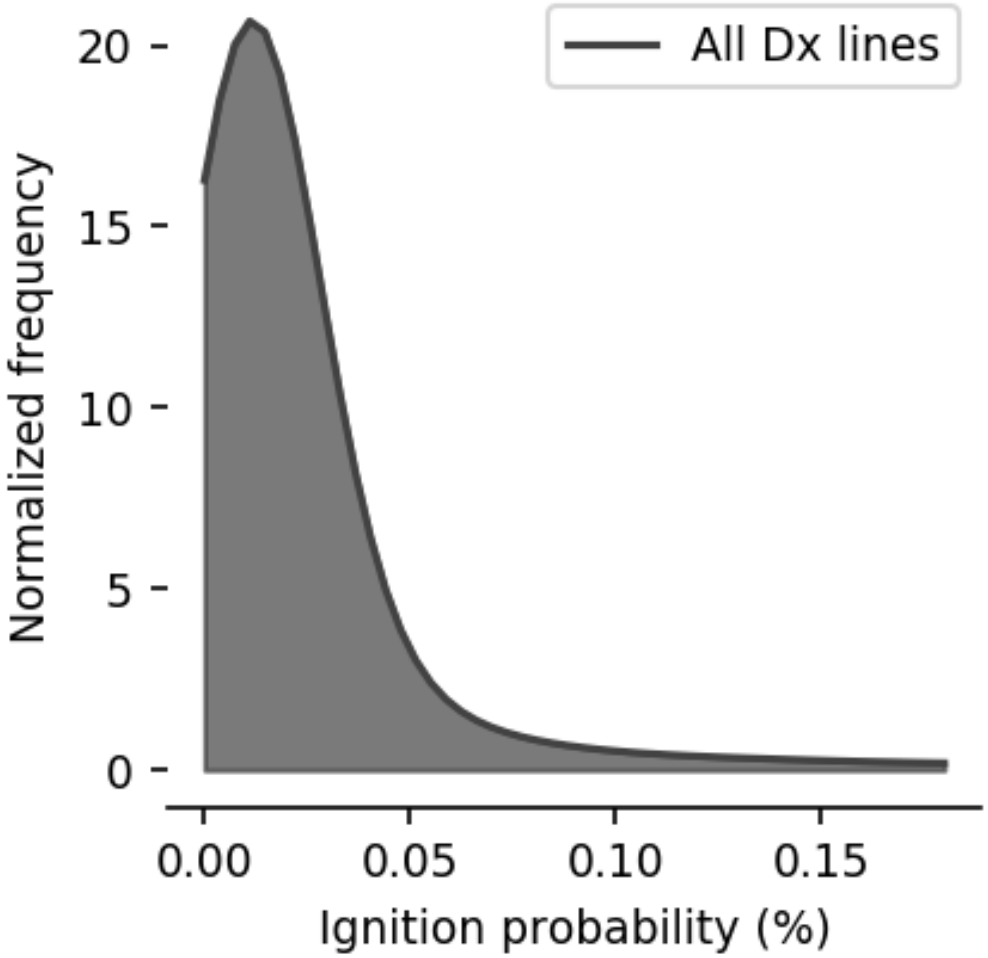


	Training 2015-2016	Testing 2017-2018
AUC (probability of distinguishing not/at-risk assets)	0.765	0.755
Recall (Fraction of ignitions found within the at-risk territory)	0.799	0.781
Precision	0.702	0.689

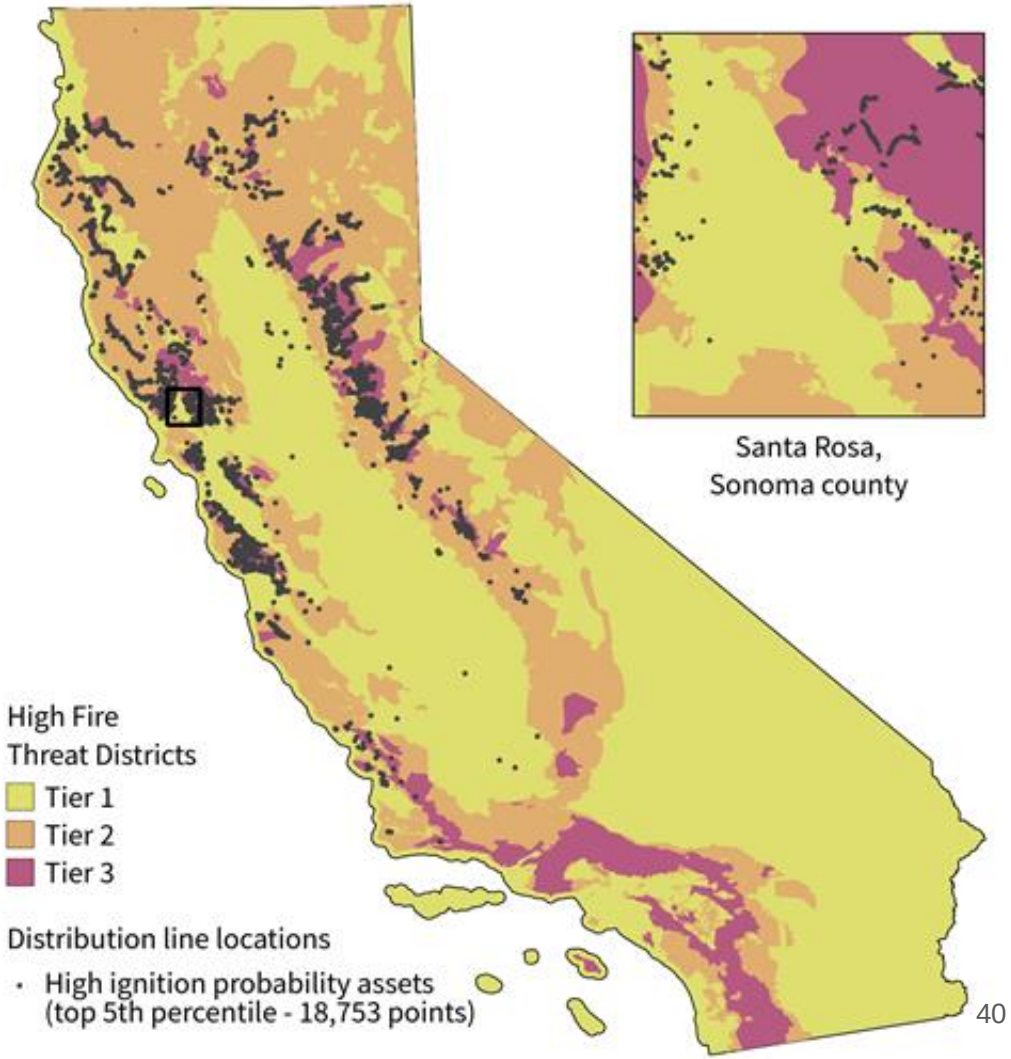


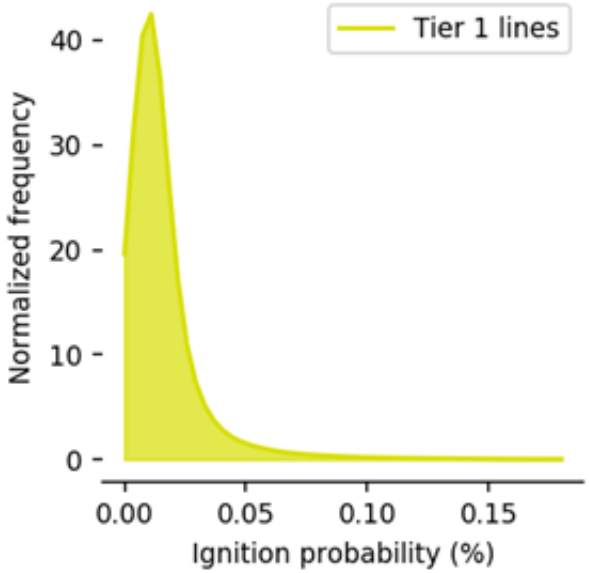

Sensitivity vs. 1 - Specificity for Vegetation
Training data (AUC = 0.778)
Test data (AUC = 0.755)
Random Prediction (AUC = 0.5)
Sensitivity (1 - Omission Rate)
1 - Specificity (Fractional Predicted Area)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

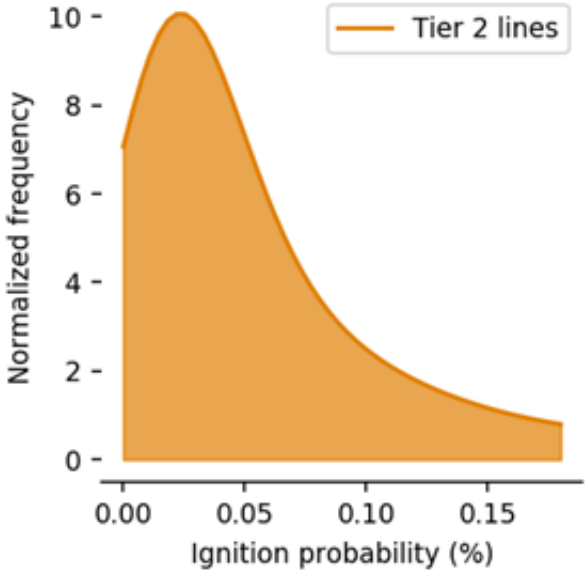
Variable	Percent contribution	Permutation importance
Max. tree height	60.6	63.5
Avg. tree height	23.2	0
Local topographic position	3.6	0
Average gust speed	3.3	16
Landscape topographic position	3.3	0
Max. temperature	1.8	2.4
Max. gust speed	1.6	0
Avg. temperature	1.4	9.6
Avg. wind speed	0.9	3.6
Max wind speed	0.3	4.9

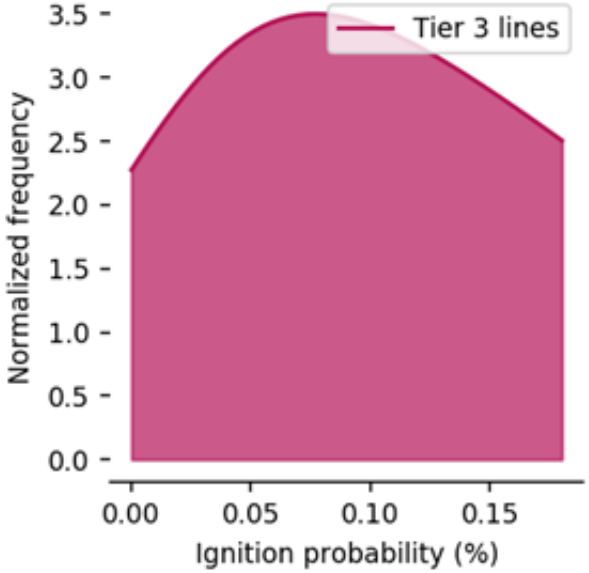

All Dx lines
Normalized frequency
20
15
10
5
0
0.00
0.05
0.10
0.15
Ignition probability (%)

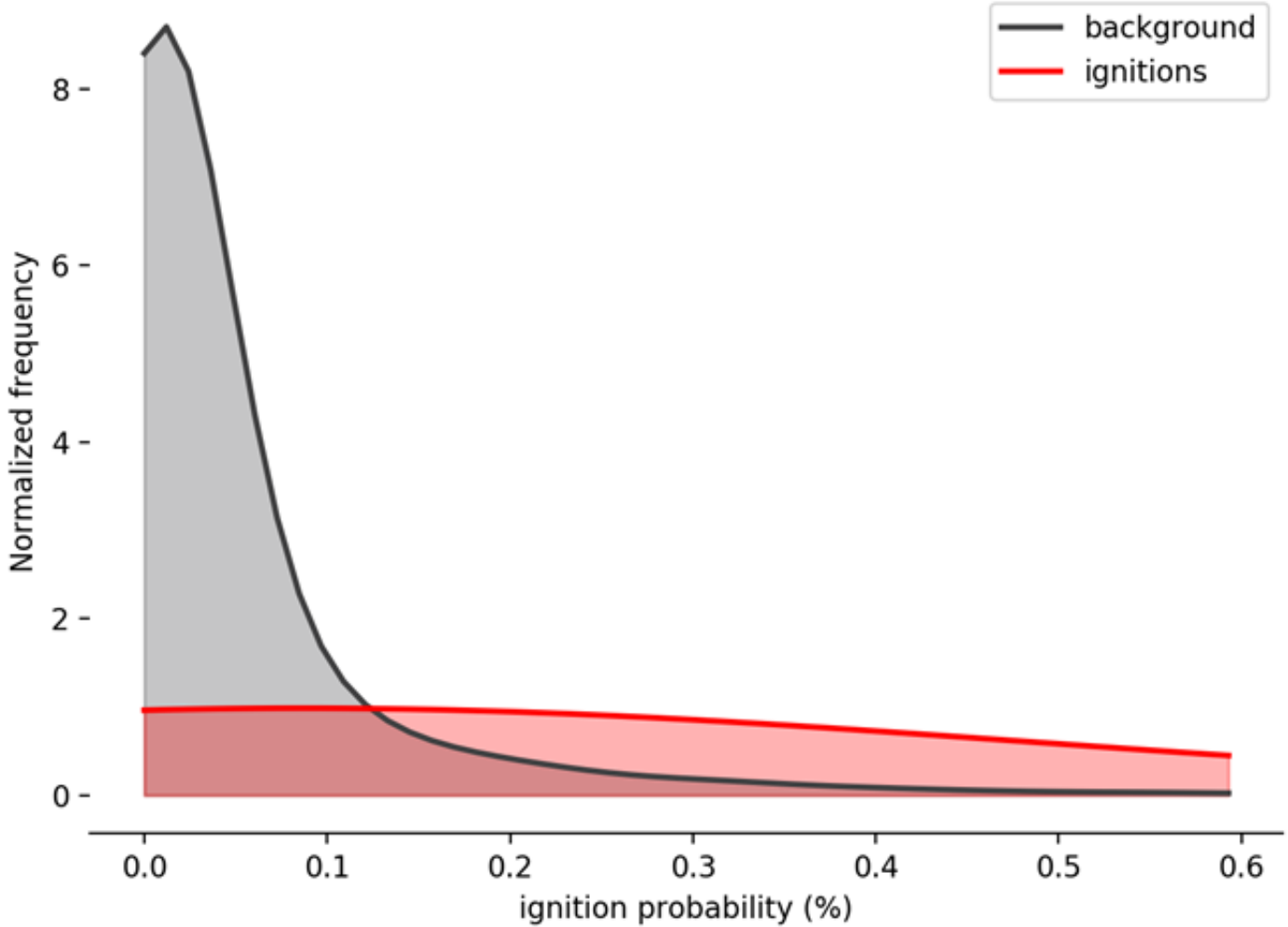
	Training 2015-2016	**Testing** 2017-2018
Predicted ignition count	229.1	200.0
Observed ignition count	210	266




Tier 1 lines
Normalized frequency
40
30
20
10
0
0.00
0.05
0.10
0.15
Ignition probability (%)


Tier 2 lines
Normalized frequency
10
8
6
4
2
0
0.00
0.05
0.10
0.15
Ignition probability (%)

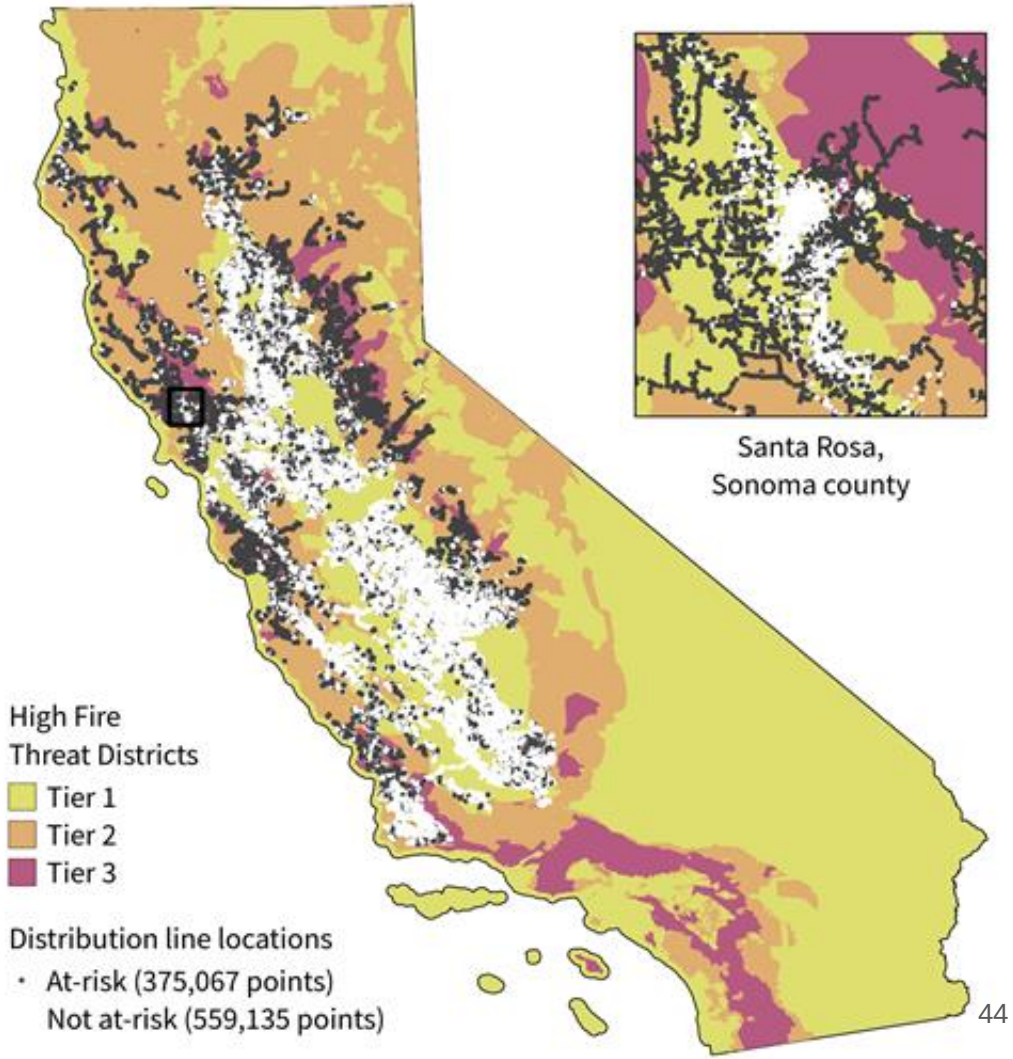

Tier 3 lines
Normalized frequency
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.00
0.05
0.10
0.15
Ignition probability (%)

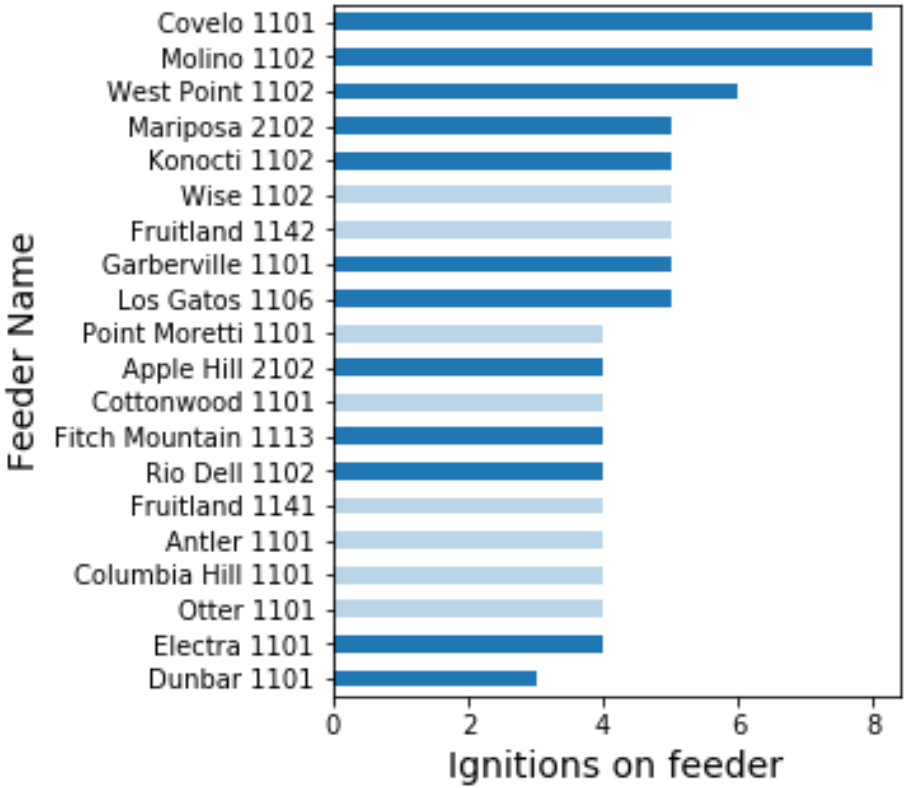

background
ignitions
Normalized frequency
8
6
4
2
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
ignition probability (%)

Presentation overview

- Background material
- Stating the problem
- How we approached the problem
  - An introduction to the MaxEnt modeling approach
  - Model performance metrics
- Input data
- Model results
- Early interpretation
- Next steps

	Training 2015-2016	Testing 2017-2018
AUC (probability of distinguishing not/at-risk assets)	0.765	0.755
Recall (Fraction of ignitions found within the at-risk territory)	0.799	0.781
Precision	0.702	0.689





Data sources and how used

Ignitions data (bars) [limited to veg]

Our predictions (filtered 100 highest risk feeders)

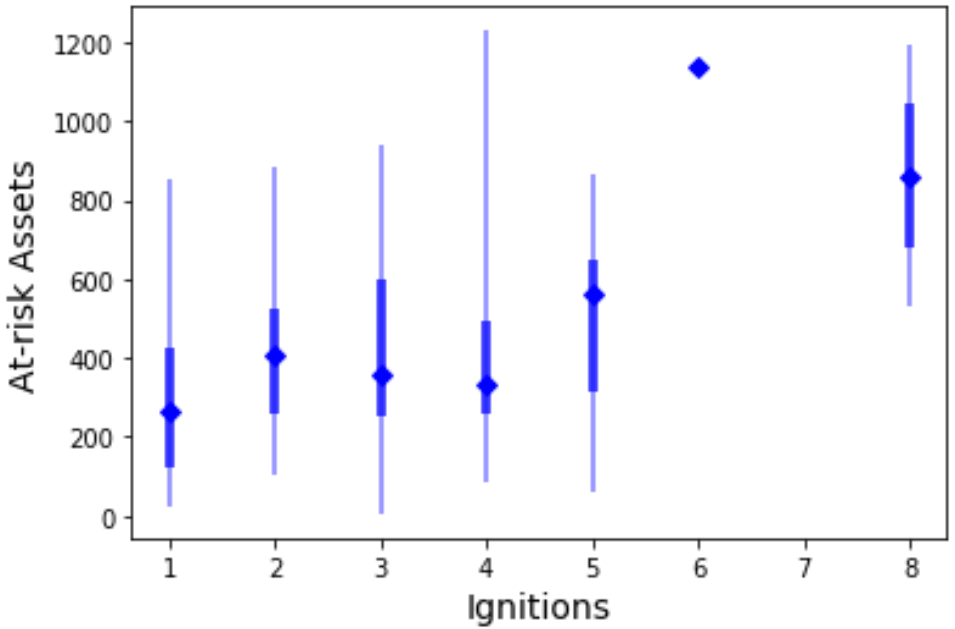
ICA data (filtered for inclusion)

Description

Number of ignitions per feeder. Dark blue indicates feeders among the 100 feeders with the highest risk score.

Comments and Caveats

Limited to feeders included in both ignitions and ICA datasets.



Data sources and how used

Our predictions (vertical axis)

Ignition data (counts along x axis)

Description

Bar chart showing risk distribution by feeder (y-axis) grouped by the number of ignitions that actually occurred on that feeder (x-axis)

Comments and Caveats

Only one feeder had 6 ignitions, and none had 7.

Presentation overview

- Background material
- Stating the problem
- How we approached the problem
  - An introduction to the MaxEnt modeling approach
  - Model performance metrics
- Input data
- Model results
- Early interpretation
- Next steps

Next steps

- Characterize and resolve spatial uncertainty
- Include additional environmental covariates
- Implement temporally-explicit modeling
- Characterize chain-of-events hierarchically

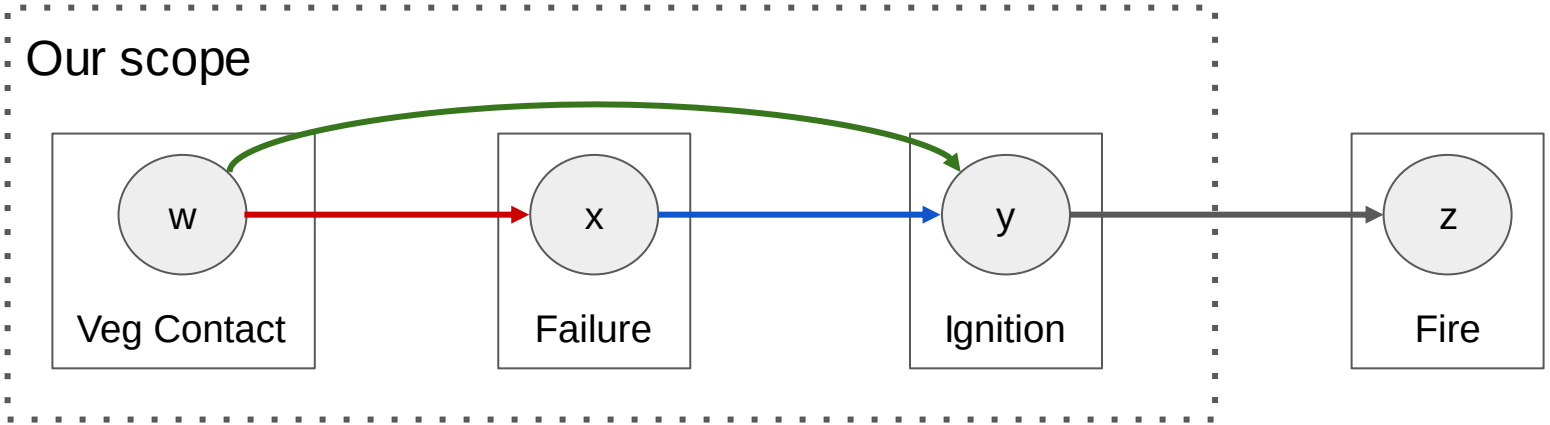
Key modeling considerations

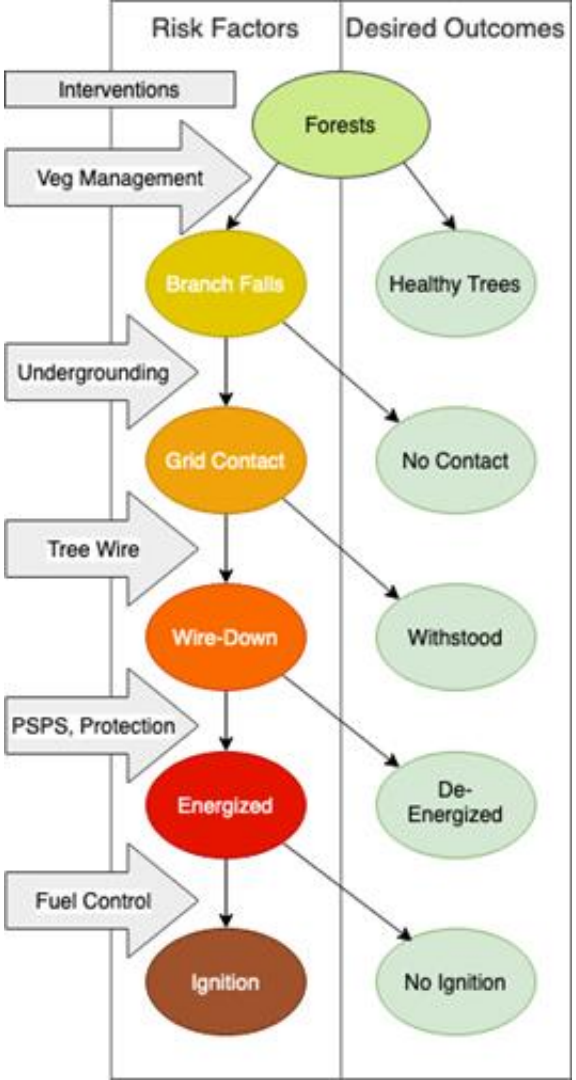
1. Ignitions are **rare events**
2. **Multiple points of failure** precede an ignition
3. The drivers of failures include both **endogenous** and **exogenous** processes
4. Failures can result from **instantaneous** and **cumulative** processes
5. Multiple forms of **uncertainty** in available data
    a. Relational topology unclear (e.g., hard to link outages to wire-downs to ignitions)
    b. Spatial uncertainty high (recorded positions are often imprecise)
6. Physical models are robust and easy to interpret, but only describe a few processes
7. Statistical models can identify novel failure patterns, but are easily **biased in predicting rare events**
8. Needs to be capture the benefits of **management/intervention activities**
9. Needs to **improve over time** as new data comes in from the field

Chain-of-events

Estimated risk reduction due to decisions related to:

1. Veg management
2. Grid hardening
3. Protection




Risk Factors
Desired Outcomes
Interventions
Forests
Veg Management
Branch Falls
Healthy Trees
Undergrounding
Grid Contact
No Contact
Tree Wire
Wire-Down
Withstood
PSPS, Protection
Energized
De-Energized
Fuel Control
Ignition
No Ignition

Central formulation	$Pr(y = 1\|z) = f_1(z) \cdot Pr(y = 1) / f(z)$
"Raw output" estimates	$f_1(z) / f(z)$
"Raw output"	$f_1(z) / f(z)$