Thrive or Survive: Predicting the Health of Trees following Forest Fires in Washington

Henry Cladouhos, Allie Cruikshank,
Christina Duffield, & Ella Palacios

Research Question and KPIs

Question: Can we predict tree survival and health following a fire, using data about the tree's past health and the fire severity?

Stakeholders:

- Disaster Mitigation Groups
- Commercial Logging
- Forestry researchers

KPIs:

- Accuracy of tree survival predictions post-fire when compared with actual historic outcomes

Finding the Data: Tree Health

Source: Forest Inventory and Analysis (FIA) DataMart[1]

- Run by US Forest Service, data covers the whole US
  - Specifically selected WA due to arboreal and boreal biodiversity and personal ties.

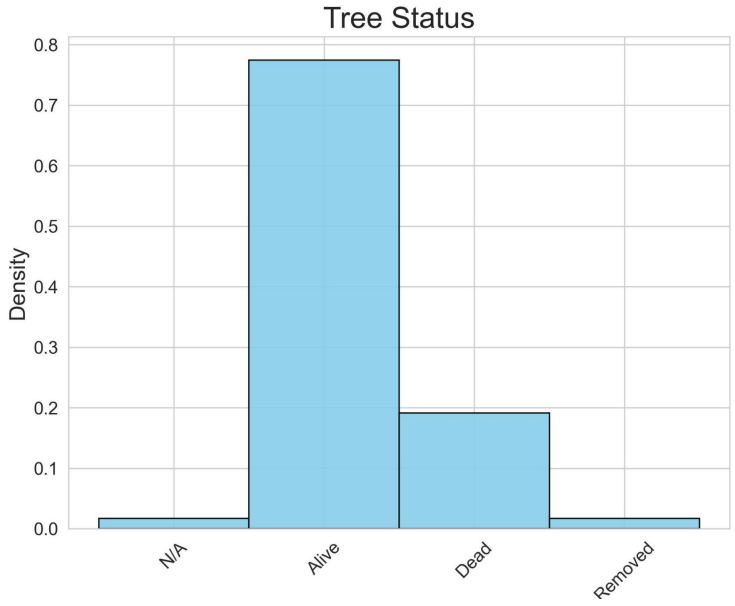
- *WA_TREE* contains tree-specific health and inventory details
  - Diameter, height, species, “cull” and “crown ratio”
  - 200 other columns, many of which are mostly empty
  - Over 200,000 individual trees!

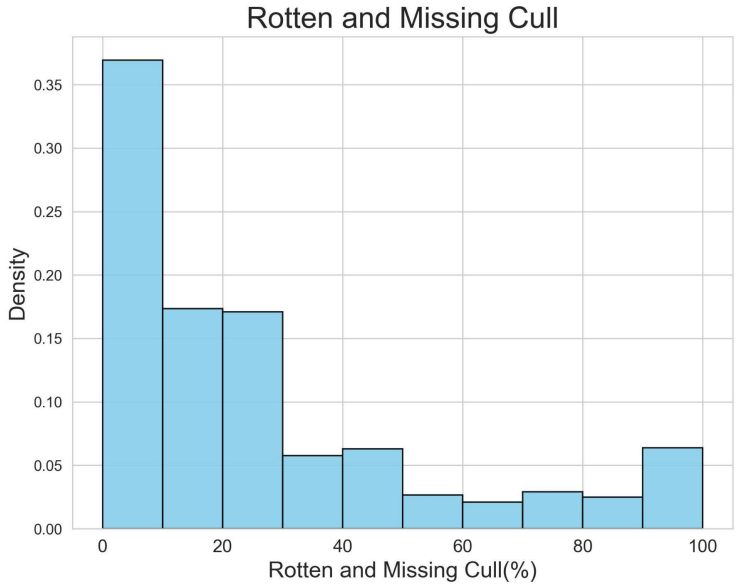
- *WA_PLOT* contains location data and sample schedule details
  - Most “recurring inventory” plots are measured on a 10 year recurring schedule since around 2002

1: https://research.fs.usda.gov/products/dataandtools/tools/fia-datamart

Finding the Data: Tree Health

Potential Target Variables: Tree Status and Rotten and Missing Cull





Finding the Data: Fire History

Source: National Interagency Fire Center[2]

- Compiles geographic fire spread data from multiple sources into a common database

- *InterAgencyFirePerimeterHistory* contains years, extents, and incident names

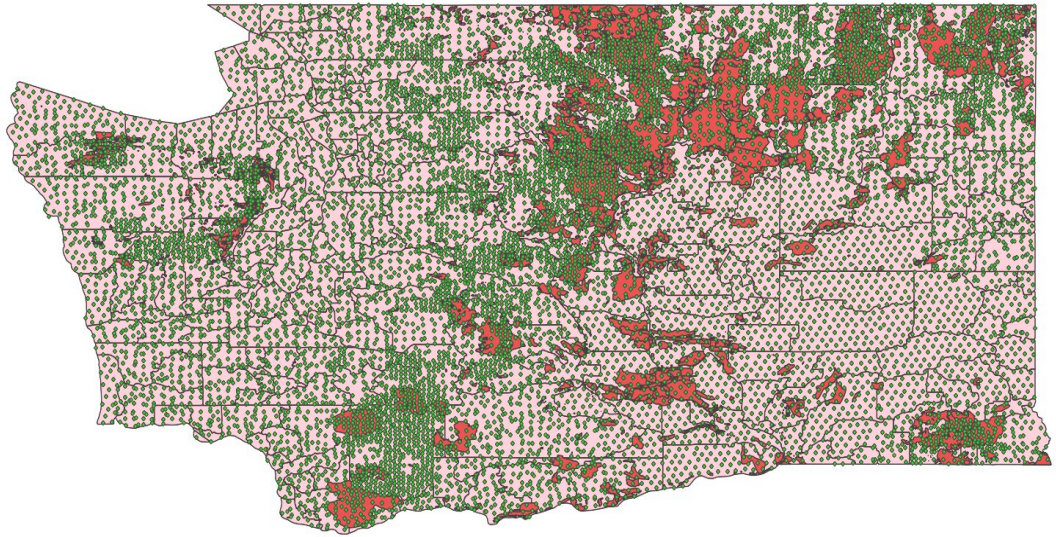
- Other datasets had more details about each incident, but we had to compromise to get more exact outlines of the fires

2: https://data-nifc.opendata.arcgis.com/

Assembling the Dataset: Geography

The puzzle pieces:

- WA_TREE links to WA_PLOT with one cross-referenced column
- WA_PLOT then contains LAT and LON to intersect with our fire data, shown here:



Assembling the Dataset: Filtering Trees

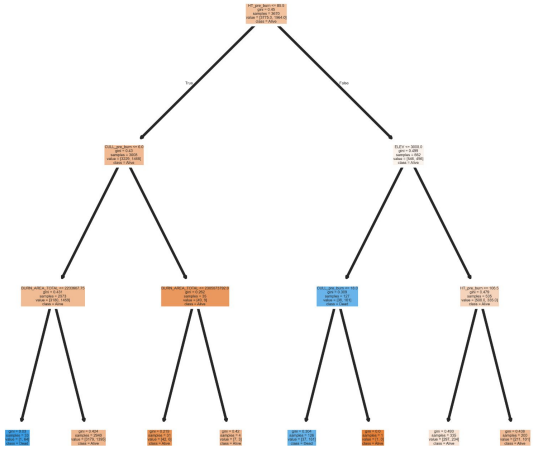
We were only interested in trees that...

- had been measured twice,
- and had at least one fire occur on the plot,
- and that fire occurred between the two measurements!

Filtering these gave us a set of about 12,000 trees across almost 350 plots, which overlapped with around 100 different fires between 2005 and 2020.

We were excited to finally get to the data analysis!

First Models



Classifying trees into “died” vs “survived”

- Baseline model: predict all trees die (72.4 % accurate)

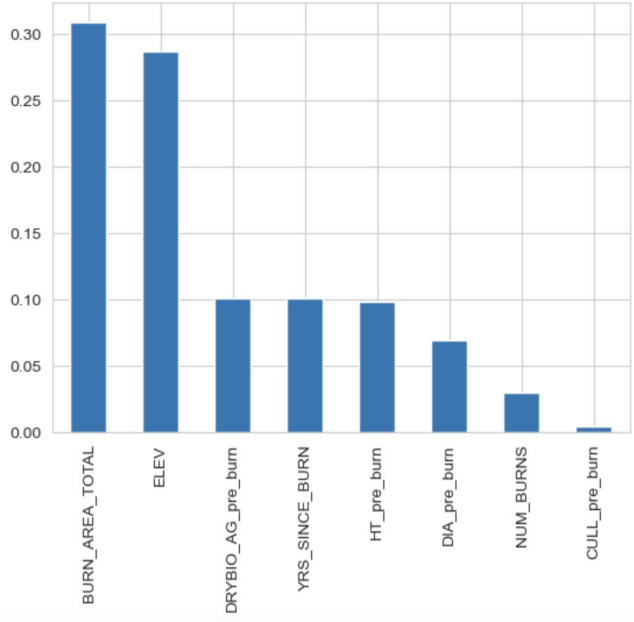
Initial results were promising:

- Support Vector Classifier
  - 75.8% accurate in cross-validation
  - 3 parameters with RBF Kernel
- Random Forests
  - 82.1 % accurate in cross-validation
  - Max depth 18, number of estimators 100
  - Single decision tree classifiers gave similar accuracy
- K Nearest Neighbors
  - 81.9% accurate in cross-validation
  - Using 8 neighbors and only 3 features... wait, only 3 features?

Data Contamination

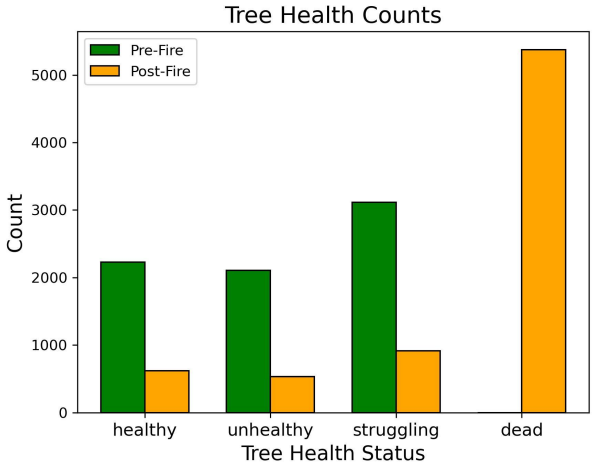
The small number of features and neighbors in KNN pointed to a data issue: nearby trees on the same plot were very likely to have similar outcomes, and inherited many of the same parameters from their source datasets (e.g. elevation, latitude, longitude, tree species, and fire information such as year, fire area, and the number of times the area burned between measurements).

We needed to split our test set out not tree by tree, but plot by plot, so that trees in the test data didn't have "sisters" in the training data, standing literally feet away.



Feature Importance from Random Forest Method. The location-based attributes, Total Burn area and Elevation, have the highest importance.

New Findings



Using our new “plotwise” split, all our models performed worse.

- SVC down to 72.6% on cross-validation
- Random Forests down to 71% accuracy
- KNN down to 71% accurate on cross-validation

Examining overall tree health among trees that survived:

- Classifying on a tree-health indicator rather than dead/alive.
    - Used first CR (crown ratio): CR ≤ 30% as unhealthy, 30% < CR < 60% as struggling, & CR ≥ 60% as healthy. Post-fire dead trees were included where the CR was reported as 0.0 or NaN.
- Used a KNN classifier with a Linear Discriminant Analysis pre-processor
    - 72.4% of trees died after fire
    - Tried various numbers of neighbors and subsets of features
    - Best testing accuracy score: 62.6% with only tree features and k = 24 neighbors, vs 44.2% from the dummy classifier!

Future Research Directions

- Analyze tree health in non-fire regions and compare with trees in fire zones.
- Try mixed-effects models, which are used for data with clusters of related statistical units. In our case, these clusters would be the trees from the same plot.
- Investigating the effect of tree species type on post-burn recovery
    - Do Hardwoods recover faster than softwoods? How does elevation affect this?
- Expand our data for more reliable results:
    - Get more details on the fire incidents (duration, suppression tactics, initial cause, etc) by collating a 4th dataset
    - Open it up to other US states
    - Consider forest health by tree species and/or type across major US Ecological Regions

Conclusion & Acknowledgements

- We learned how to combine large, messy datasets, which is useful experience for us as new data scientists!
- We learned to be cautious and intentional with our data handling; sometimes a plain train-test split won't be unbiased.
- Data contamination is difficult to avoid and can occur when combining datasets.

Thank you to Steven Gubkin, Amzi Jeffs, and Alec Clott for their meaningful contributions & guidance to this project. We appreciate you!