Thrive or Survive: Predicting the Health of Trees following Forest Fires in Washington
Henry Cladouhos, Allie Cruikshank, Christina Duffield, & Ella Palacios

Motivation: The motivation behind this project is to build a model that can accurately predict the health of trees following a forest fire, given previous health data. We picked the state of Washington to begin developing the model due to its diverse boreal & arboreal ecosystems at varying elevations that fall victim to yearly forest fires. Preserving Washington forests is a passion of ours.

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

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

Datasets: For tree health data, we used two datasets from the Forest Inventory and Analysis (FIA) Datamart which is run by the US Forest Service. *WA_TREE* was used for tree-specific health and inventory details such as diameter, height, species, etc. The location of these trees are contained in the dataset *WA_PLOT*. For fire history, we used the dataset *InterAgencyFirePerimeterHistory* from the National Interagency Fire Center (NIFC) which contains years, extents, and incident names of fires throughout Washington. We used the location details in *WA_PLOT* to link *WA_TREE* to the fire dataset.

Methods: To create our final dataset we identified which plots in *WA_PLOT* were in regions of fires from the NIFC shapefile. In these plots we took trees from *WA_TREE* which had been measured twice and had at least one fire occur between the two measurements. Then these trees

Once we had our final dataset we used a variety of classification models to predict post-fire tree survival. These models included K Nearest Neighbors, Support Vector Classifiers, Logistic Regression, and Random Forests. We aimed to improve upon our baseline model, which simply predicted that all trees in the plots with fires on them died.

We chose a selection of features out of the 200 features from *WA_TREE,* many of which were mostly NaNs, to include in our models. At first we made our train/test sets and cross validation sets in the standard way, but due to evidence of data contamination we created additional versions of these sets in which trees from the same plot stayed together, this way plot conditional features could also be applied. Our models were tested on both versions of the train/test/CV sets.

Results: As described in the methods, we tested our models on two versions of the train/test splits. For the standard train/test split with stratified k-fold cross validation (stratifying with target variable *Alive/Dead*), the support vector classifier with the kernel Radial Base Function had an accuracy of 75.8% in cross-validation, random forests with a maximum depth of 18 and 100 estimators had an accuracy of 82.1%, and K Nearest Neighbors using 8 neighbors and 3 features had an accuracy of 81.9% in cross-validation. All three models beat the baseline model which predicted all trees as dead post-fire and had an accuracy of 72% upon cross-validation.

The new and improved train/test split allowed for all trees on a plot to stick together throughout the train/test split and the k-fold cross validation. With this version, the same baseline had an accuracy of 72%. The support vector classifier with the optimal kernel polynomial had an accuracy of 72.6% in cross-validation, random forests had an accuracy of 71%, and K Nearest Neighbors had an accuracy of 71%. In addition, we also used a logistic model with principal components analysis for feature selection and received an accuracy of 72.5%.

Conclusion: While our results initially looked promising, the models trained on the updated train/test/CV splits, which kept trees from the same plot in the same set, were not. This initial inflated accuracy had to do with data contamination from the inheritance of plot and fire features. Our initial models relied heavily on elevation and fire size to predict survival of the burnt trees, two features which were the same for all trees in the same plot. In this way, the models used elevation and fire size as a stand-in for geographic location and simply predicted that trees near each other would have similar outcomes. While these models used other features to reach their accuracies, when we split the data while keeping trees from the same plot together—effectively removing the possibility of this model ‘shortcut’—our accuracies sank to near the baseline.

Although there were pitfalls, there were meaningful accomplishments along the way. We sorted through a large, messy dataset and turned it into something useful which is an essential skill to have in data science. Also, in order to combat data contamination, we created a modified train-test split and k-fold cross validation function that combined trees on the same plots.

Further Directions: In order to achieve more reliable and robust results, it would be valuable to expand the dataset in a few key areas. Obtaining more detailed information on fire incidents such as intensity, duration, suppression tactics, and/or initial causes would provide a more comprehensive understanding of the factors that affect tree health post-burn. Including data from other states beyond Washington would allow for more generalization of our model. In addition to expanding our dataset, another direction would be to compare/analyze tree health in non-fire regions and compare with trees in fire zones. Along the same lines, investigating the effect of tree species on post-burn recovery would add more insight to the study. Finally, from a methodological standpoint, employing mixed-effect models would be helpful as these are used for data with clusters of related statistical units. In our case, these clusters would be the trees from the same plot.