Building an Early Detection Model for Forest-Fires: A Machine Learning Approach

Abstract

Satellite imagery is not currently used in *real-time, or even near real-time*, to aid in forest fire prevention. Currently, there are hundreds of satellite images from NASA's AQUA and TERRA satellites run through an Active Fire Product out of the University of Maryland. Each day, this Active Fire Product outputs a data set that holds hundreds of 'detected fires' at given latitude/longitude coordinates. Using machine learning, we can tease out which of these 'detected fires' are forest-fires, and use this knowledge in *near real-time* to aide in forest fire prevention.

1. Introduction

In 2015, hundreds of forest-fires burned over 9 million acres of land, causing millions of dollars in property damage and immeasurable loss to those families effected. Understanding, tracking, and effectively fighting forest-fires is crucial in terms of minimizing this damage and loss. Using satellite imagery of potentially detected forest-fires can greatly aid in this process.

Each day, hundreds of satellite images from NASA's AQUA and TERRA satellites are run through an Active Fire Product algorithm at the University of Maryland. The output of this algorithm is a data set that holds hundreds of detected fires for a given day. Unfortunately, this dataset contains a large number of false-positives in terms of forest-fires. That is, not all observations are actually forest-fires. Using machine learning, we can parse this dataset down to only those fires which are forest-fires, in near-real time.

The ability to identify forest-fires from the AQUA and TERRA satellites in near real-time could provide an additional tool for more quickly identifying forest-fires and tracking them through time and space. It could also allow for more informed decisions in terms of resource allocation. Finally, it could allow for future research on modeling forest-fire severity ahead of time, which is crucial in the decision making and resource allocation process.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. Approach

In this section, I will first present a brief discussion of the data sources used (Section 2.1). Then, I will detail the process of joining these data sources to establish ground truth (Section 2.2). Lastly, I will describe the modeling techniques used (Section 2.3).

2.1. Data Sources

Data for this project includes the following:

- Fire Detection GIS Data
- Geographic Boundaries
- Forest-Fire Perimeter Boundaries

Fire Detection GIS Data is provided by the USDA Forest Services website. Each row of this detection data is a 'detected fire', as determined by running AQUA and TERRA satellite imagery through the University of Maryland's Active Fire Detection Product. Geographic boundaries are provided by the Census Bureau for selected geographic areas, of which region, state, county, and urban area boundaries are used. Finally, forest-fire perimeter boundaries are submitted individually by states, and aggregated together by the the U.S. Geological Survey. These files contain forest-fire perimeter boundaries (along with other identifying information) that are submitted by *many (but not all)* states at the end of each day.

Both the Fire Detection GIS Data and the Forest-Fire Perimeter Boundaries from 2012-2015 are used, and the Geographic Boundaries from 2013 and 2014 are used. The Census Bureau only provides boundaries for 1990, 2000, 2010, 2013, and 2014, leaving 2013 and 2014 the most appropriate boundaries to use. All data sources are provided as zipped folders containing spatial data format shapefiles (.shp). The Fire Detec-

tion GIS Data contains geographical points, whereas the Forest-Fire Perimeter and Geographic Boundaries contain geographical polygons or multi-polygons.

2.2. Data Joining

By far, the most important piece of the data joining and aggregation process is determining how to establish ground truth. That is, which of the 'detected fires' in the Fire Detection GIS Data are forest-fires?

The easiest and most simplistic method for labeling observations would be to identify only those observations that fall within a forest-fire perimeter on their given date as forest-fires. However, as the documentation for the boundaries clearly notes, these boundaries are not final nor official, and are derived from data produced by GIS specialists. Such a boundary generation process leaves room for error, and the simplistic labeling process proposed directly above would expound upon that error.

The most obvious source of error introduced in such a labeling process is that introduced by the extreme detail that is involved with lining up an individual latitude/longitude coordinate of a 'detected fire' with a forest-fire perimeter boundary. Even a 10th of a degree difference in the latitude or longitude could change the value of an observations label (e.g. forest-fire or not). The implications of this are shown in Figure 1, where it can clearly be seen that observations only a small distance outside of a forest-fire perimeter boundary would not be labeled as forest-fires, whereas those only a minute distance away (but within a fire perimeter boundary) would be labeled as forest-fires. It seems reasonable to conclude that in reality, both sets of observations should be labeled as forest-fires.

A less obvious source of error in such a labeling process is that introduced by incomplete (e.g. not reported) forest-fire perimeter boundaries. The implications of this are shown in Figure 2. Here, it can be seen that a grouping of observations on one day *would not* be labeled as forest-fires because of a lack of a surrounding forest-fire perimeter boundary, while a grouping of observations in a similar location on the next day *would* be labeled as forest-fires. This is another instance where it seems logical to accept that both sets of observations should be labeled as forest-fires.

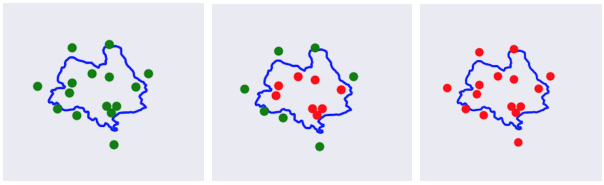

Figure 1. Trinity County, CA (northwestern part) - August 01, 2015. Left: Unlabeled observations. Center: Observations are labeled as forest-fires (red) using the simple approach of whether they fall within a forest-fire perimeter boundary (blue) for their given date. Right: Observations are labeled as forest-fires (red) if they fall within 500m of a forest-fire perimeter boundary (blue) on their given date, or three days ahead.

Given these potential sources of error when using this simplistic labeling process, the current methodology of labeling 'detected fires' as forest-fires revolves around finding those observations that fall within 500m of a forest-fire perimeter boundary on their given date, or within 3 days ahead of their given date. These observations are then labeled as forest-fires, while all others are labeled as non forest-fires. Examples of this can be seen in Figures 1 and 2.

As a final note, this process is still an active area of research. Being the crux of the project, it is crucial to make sure this labeling process is as accurate as possible.

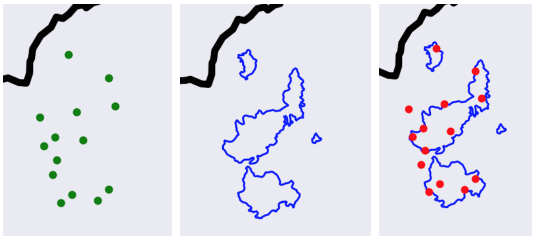

Figure 2. Trinity County, CA (northwestern part). Left: No reported forest-fire perimeter boundaries for August 01, 2015. Note that no observations are labeled as forest-fires under the simple approach of whether they fall within a forest-fire perimeter boundary for their given date. Center: Reported forest-fire perimeter boundaries for August 02, 2015. Right: Observations from August 01, 2015, labeled as forest-fires (red) if they fall within 500m of a forest-fire perimeter boundary on their given date, or three days ahead. Note that all observations are labeled as forest-fires because they fall within 500m of the forest-fire perimeter boundary on August 02, 2015 (center).

2.3. Modeling Approach

Thus far, the labeling process described in Section 2.2 has been the focal point of this project. The majority of the modeling process has focused on using a model that would both pick up the detailed interactions that arise when working with forest-fire data and allow for quick iterations. To that end, I have thus far been fitting Random Forests.

Modeling has proceeded using a training, validation, and test set. The training and validation sets are restricted to only those observations in 2012-2014, and the test set contains those observations from 2015. Cross-validation has been used to select model hyperparameters, and care has been taken to account for the time-sensitive nature of the data (e.g. the observation with the latest timestamp in the training set occurred earlier in time than the observation with the earliest timestamp in the validation set).[1]

Given a goal of identifying observations that are labeled as forest-fires (e.g. positively labeled observations), area under the precision recall curve has been used as the metric by which to judge models. Area under the ROC curve has also been examined.

3. Results

3.1. Evaluation Metrics

As mentioned in Section 2.3, area under the precision recall curve has been the metric used by which to judge models. This measures the tradeoff between correctly identifying observed forest-fires (e.g. recall) and ensuring that a high fraction of observations predicted to be forest-fires are actually forest-fires (e.g. precision). The average of the area under the precision recall curve across the test set is 0.807, while the average of the area under the roc curve across the test set is 0.758.[2]

Given the time-component of this modeling procedure, a natural question is how well the model performs on the test set over time. Table 1 shows the average area under the precision-recall and roc curves across each month of the test set. As it might be expected, the model performs noticeably worse towards the beginning of the year. At this point, fire-season is just starting, and it may be hard for the model to distinguish between forest-fires and non-forest-fires.

Table 1. Average area under the precision-recall and roc curves across months of the test set (2015). Note that those months that do not have any observed forest-fires are not shown, since these metrics cannot be calculated for these months.

Month	Precision-Recall	ROC	Obs.	Fires
March	0.49	0.01	46	1
April	0.56	0.46	831	30
May	0.81	0.65	2406	100
June	0.82	0.67	5233	1265
July	0.80	0.64	6405	1312
August	0.72	0.87	68358	47092
September	0.92	0.85	8671	2141
October	0.85	0.55	7023	182
November	0.997	0.96	308	2
December	0.50	0.50	376	1
Overall	0.807	0.758	99657	52126

3.2. Feature Importances

Another natural question to ask is what variables are most important in terms of predicting whether or not an observation is a forest fire. Figure 3 presents the feature importances of the top ten predictors, as determined by the decision tree paths of the individual trees in the Random Forest (e.g. how many observations were affected by a given split on a particular variable, aggregated across all trees).

By far, the most important features revolve around the number of observations that were labeled as forest-fires around any given observation across time and space. Specifically, the top two features are the number of observations labeled as forest-fires that are within 0.1km of a given observation, up to 4 or 6 days prior (with similar features having a similar interpretation). Other top features include factors that one might expect to be helpful in predicting whether or not an observation is a forest-fire: latitude/longitude, time of day, and temperature.

4. Future Work

Future work includes a number of steps. First and foremost, the process of labeling forest-fires needs to be solidified. While the current methodology provides a logical manner in which to label observations as forest-fires, more research could be done to determine how to most appropriately establish ground truth.

A second step would be to include additional data sets.

[1] To balance the sparsity of observed forest-fires on each day with a desire to validate in as realistic of a way as possible, cross-fold validation consisted of using some subset of observations prior to a given date as the training set, and all observations in the month after as the validation set. This 'given date' started out as January 1st, 2014, and moved forward by a single month at a time until the end of 2014.

[2] This average is the result of a weighted average for the given metric across each day of the test set (2015), where the weights were calculated as the fraction of the yearly observations that occurred on a given date. An alternative weighting scheme would be to use the fraction of the yearly fires that occurred on a given date. Using this weighting scheme, the average area under the precision-recall curve was 0.867, and the average area under the roc curve was 0.728.

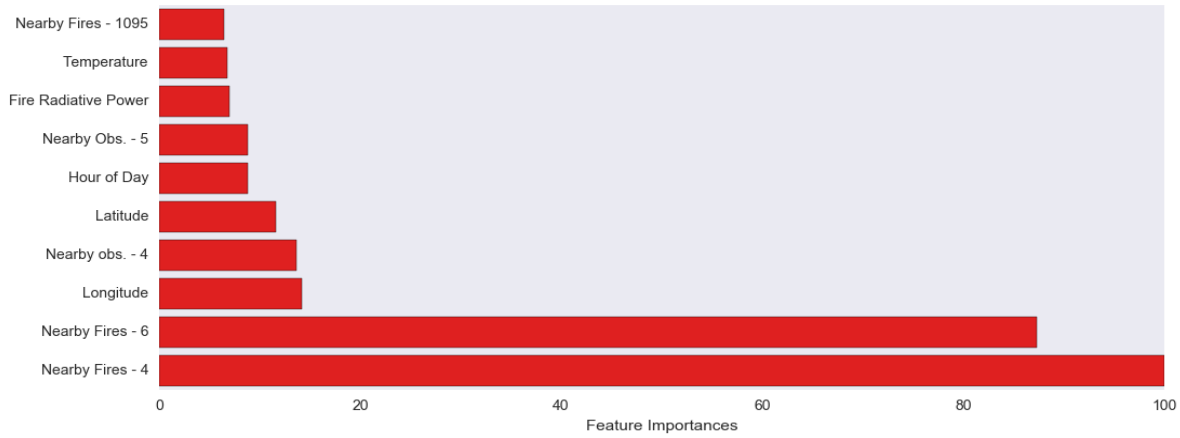

Figure 3. Feature importances across the random forest models, with the most important feature scaled to 100. 'Nearby Fires - n' are observations that are labeled as forest-fires, are within 0.1km and up to n days back in time of a given observation. 'Nearby obs. - n' are any observations within 0.1km and up to days n back in time of a given observation.

Weather data is a critical piece of the resource allocation and decision making process surrounding forest-fires, and currently this project does not make use of such data.[3] Another data set that would most likely be helpful is one that gives the type of forest coverage for a given geographic region.

Finally, after solidifying the ground truth and obtaining additional relevant data sets, fine-tuning of machine learning models could give additional gains in terms of predicting which observations are forest-fires. The results presented here suggest that with each of these steps, a machine learning model could be built that could be used in near real-time to aide in forest-fire prevention.

[3]This is not to say that attempts at including weather data to aide in modeling have not been made. At the moment, significant effort is being made at obtaining a robust data set at fine enough time-intervals to aide in modeling.