DS 6030
Statistical Learning
Disaster Relief Project Group 11: Part 1

Angelo Orciuoli
Christina Land
Scott Tumperi
Tyler Hobbs

Pre-Processing

The most important part of pre-processing data is separating the classes and making it a binary outcome from such classes as well. In regard to the training dataset, the classes were originally separated into five different classes in the original dataset. These such classes were Blue Tarp, Rooftop, Soil, Various Non-Tarp, and Vegetation. As the dataset is coming in the aftermath of the destruction of Haiti's earthquake in 2010, and there is a team from Rochester Institute of Technology flying an aircraft and collecting high-resolution geo-referenced imagery, it became known that those who had been displaced from their homes had temporary blue tarps outside of their homes. As such, the class for this dataset will require only focusing on the Blue Tarp class and not any of the other classes. Since the rest of the classes are not needed, the classes of Rooftop, Soil, Various Non-Tarp, and Vegetation were set to Non-Blue Tarp within the training dataset. As this was merely now a binary classification of Blue Tarp and Non-Blue Tarp, these became different groups within the data, and they were factored as well. During this process, a column type was created to make sure that it was shown if a certain value in the training dataset was the class Blue Tarp or Non-Blue Tarp.

The holdout set comprised eight different files, compared to the training set that was a single file. These holdout sets contained four files that were exclusively just the Blue Tarp data, and there were four files solely for the Non-Blue Tarp data. The first step in pre-processing this data was to load these all and concatenate them in a dataframe that included all of the holdout data. As K-nearest neighbors (KNN) will not be used, the location data was removed, and a subset of this data was created with only the labels for the colors. The prior file for the holdout data had column labels of B1, B2, and B3, and after the exploratory data analysis, it was found that these labels should be Red, Green, and Blue, respectively. These column labels were changed on the dataframe to correctly support the data.

The training set had columns labeled red, green, and blue, but the holdout data did not. To make sure we were applying the proper colors to the columns B1, B2, and B3 of the holdout set, we defined variables of blue/green and blue/red for the training data, and then B3/B2 and B3/B1 in the holdout data to make the following graphs. The values of Red, Green, and Blue are not independent of each other; therefore, we created three new variables representing color ratios (Blue/Red, Blue/Green, and Green/Red). The figures are plotted using these new features to visualize the relationship between combinations of color values and their associated classes. Figure 1 shows the distribution of the training set data, with all of the classes (blue tarp, rooftop, vegetation, etc.). Figure 2 shows the distribution of the holdout data, with just blue tarp and non-blue tarp shown. 75% of the blue tarps are above and to the right of the dotted lines in both graphs.

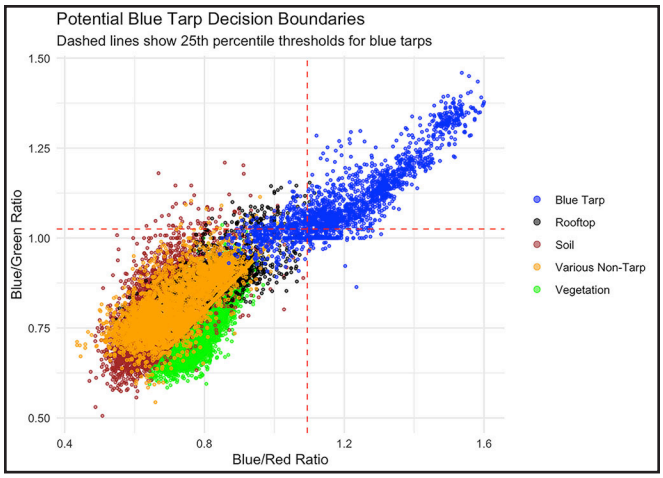

Figure 1

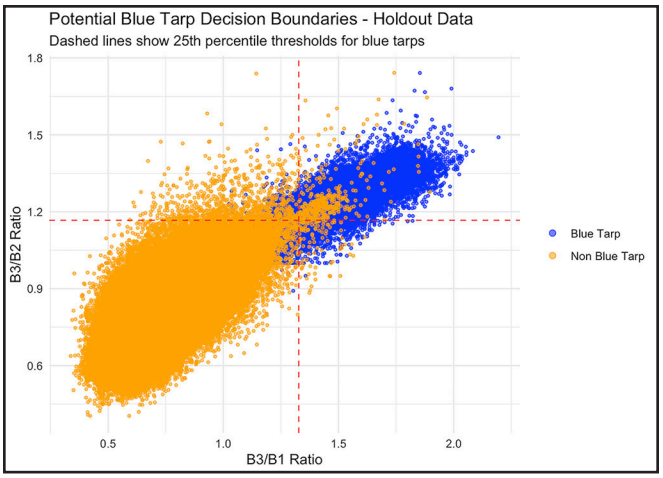

Figure 2

Exploratory Data Analysis (EDA)

Within the exploratory data analysis (EDA) of the training and holdout set, the class labels were found for the holdout set. These were stated within the pre-processing in that the labels of the holdout set of B1, B2, and B3 are the classes Red, Green, and Blue, respectively. This was found by analyzing the ggpairs plots of the training dataset and the ggpairs plot of the holdout set with only the class data within it. Also, through the analysis, there is a clear skew of Blue Tarps vs Non-Blue Tarps within the pixel data, as the values for the pixel data for the blue tarps are significantly higher than those of green or red. It was shown within the density plot a very high density within the blue tarps, so there is a high likelihood that all of the blue tarps were located in very close proximity. Then, the green tarps were still quite densely populated in a close area, but with the red tarps, they had the smallest density, so it is likely that they are further apart than the rest of the tarps within the model.

Models

The models to be investigated will be Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Penalized Logistic Regression (Elastic net penalty), and Ensemble Method (Random Forests).

Validating the Models

As for Logistic Regression, Linear Discriminant Analysis, and Quadratic Discriminant Analysis, these models will be trained using the training set. Then, they will go through a set of cross-validation. After the cross-validation, the metrics for each of these models will be analyzed, specifically the accuracy and the AUC from the ROC curve. All of these AUC values will then be compared simultaneously, and the ROC curves will also be plotted together for each initial model.

For the Penalized Logistic Regression and Ensemble Method models, they will be evaluated by using the same pre-processed training set that was used on the earlier models. First, with the Ensemble Method, there will be a tuned model on the training set, and the tuning parameter will be set to the threshold so that the optimal value for finding the false positives can be found. With this method, it will specifically follow the method of random forest tuning within the ranger library. To find this, the pre-processed training set will have a workflow defined for itself, and it will go through cross-validation.

Once this is found, a final model will be trained accordingly. The model performance will then be shown with its metrics of the AUC and the ROC curve as well. Finally, the Penalized Logistic model

will be tuned with the penalty parameter, allowing for a stronger regularization of the model. This model will follow the glmnet engine of regularized linear regression, and in this case it will be a classification model. Similar to the Ensemble Method, it will require a workflow to be defined from the pre-processed training set and to go through cross-validation. After this process, the model will be analyzed by using the metrics of accuracy and the AUC of the ROC curve. When these metrics are found, the ROC curve with the appropriate regularization for the model will be shown for the Penalized Logistic model as well.

If the threshold from tuning in the models is not appropriate, and it needs to be adjusted, a series of steps will be required to adjust it. Using the workflow and the cross-validation results from each of the tuned models, determine the appropriate threshold by using the metrics and, specifically, the F-measure in this case. Once the appropriate threshold is found where the AUC is maximized for each of the tuned models, a final model will be trained and will be evaluated for its accuracy on the new threshold. It will be evaluated by performing cross-validation once again, but this time, on the holdout set, and outputting the AUC and the ROC curve of the holdout set at the new threshold. By going through this series of steps, it will help maximize the true positives, and the main goal, helping to find those who are displaced, and under the blue tarps.