Disaster Relief Project, Part 1

Group 13: Sarah Christen, Katherine Kelleher, Margaret Lindsay, and Brian Nolton

I. Introduction

After an earthquake devastated Haiti in 2010, coordinated disaster relief efforts needed efficient and reliable means of locating displaced people. Many of these displaced individuals used blue tarps for temporary shelter. As these blue tarps stood out in aerial images, the Rochester Institute of Technology collected high resolution imagery from across affected areas of the island. While these images provided an effective method for identifying the location of displaced people that could be shared with rescue workers, the volume of the images and the urgent need to provide resources to affected individuals posed a challenge.

The goal of this project is to address this challenge in a similar manner to how it was addressed by disaster response organizations and their partners: using statistical models to efficiently and accurately identify the location of blue tarps in aerial imagery. Using the extracted RGB values from each pixel in these images, we built models to determine whether the pixel contained a blue tarp. In the first portion of this project, we built models using Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Logistic Regression.

II. Data

A training dataset consisting of 63,241 observations was used to develop the three models. The training dataset had four columns: Class, Red, Green, and Blue. The Class variable contained five levels in the training set: Blue Tarp, Rooftop, Soil, Various Non-Tarp, and Vegetation. However, as the aim of this project is to identify a pixel as a blue tarp or not a blue tarp, we updated the Class variable in the training set to a binary variable: “Blue_Tarp” if the Class was “Blue Tarp” and “Non_Tarp” for all other values. Class therefore functions as the binary response variable, which indicates whether the pixel represents a blue tarp or not. The Red, Green, and Blue columns represent the hue the camera captured. Each color has a numerical scale from 0-255, together they make the hue of the pixel ina photograph.

The holdout dataset came with many files, mostly .txt files but a few .jpg files as well. Since we did not have the means to convert the .jpg files to an RGB data format these files were left alone. The .txt files were organized into what appear to be four regions (or flight paths from when the data was taken), region 57, 67, 71, and 78. Regions 67, 71, and 78 have separate files for Blue Tarps and Non-Blue Tarps making it easy to distinguish the response variable, while region 57 only has a Non-Blue Tarp data set. Region 67 had two Blue Tarp datasets. After inspecting both files it appeared that one of them was the original dataset and one of them was a pre-processed copy of the dataset stripped down to the RGB columns. Since they were duplicates, we chose to use the original dataset because importing the data would be easier since this file follows the same form as the other datasets allowing us to define a function to process all the datasets the same way.

The three predictor variables were the red, green, and blue pixel values from the aerial images. While the training data column headers specified the red, blue, and green columns, the holdout data files did not and instead labelled the columns B1, B2, and B3. Intuition would tell us the order should be red, green, and blue, but we decided to verify this. We compared the box plots for the training and holdout datasets specifically for each color split out by known Blue Tarps and Non-Blue Tarps in Figure 1. We noticed the same trend in the colors for each set of data. For the Blue Tarp data, Red had the lowest median, Blue had the highest media, and Green was somewhere in between for both the training and holdout data. For the Non-Blue Tarp data, the exact opposite was true. Red had the highest median, Blue had the lowest, and Green was still somewhere in between. Because of this, we were able to deduce the colors for the holdout dataset columns were indeed in the order of Red, Green, and Blue.

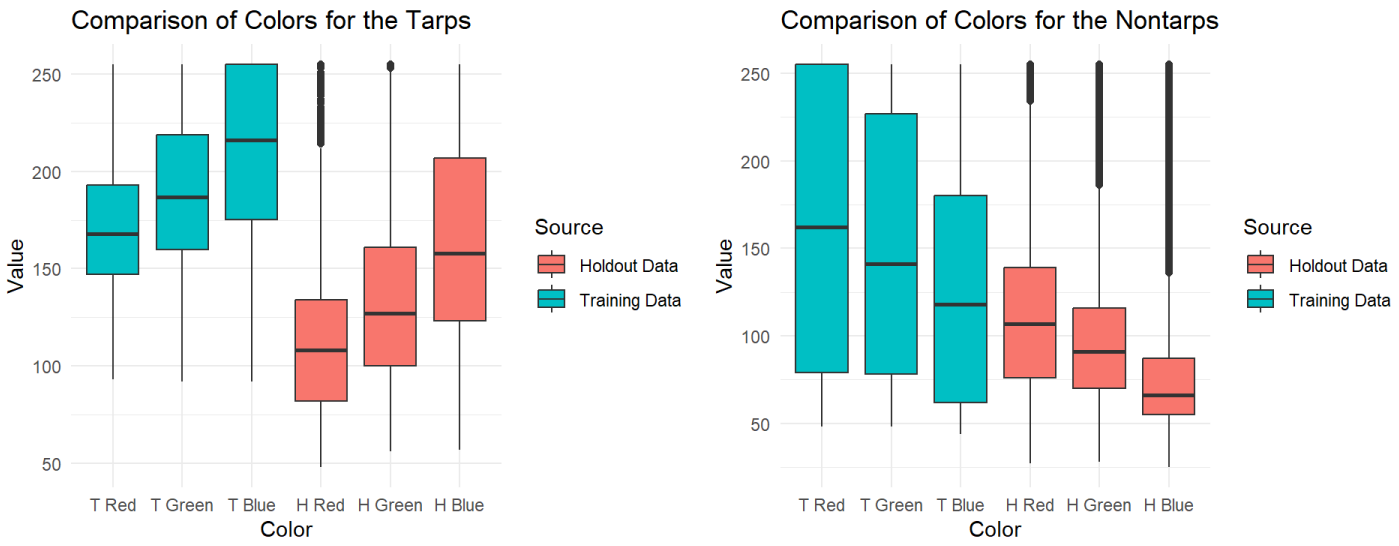


Figure 1. Boxplots of Red, Green, Blue Variables in Training and Holdout Datasets For Tarp and Non-Tarp Pixels

We also noticed that despite the data having the same trend, and the data containing the same Class (e.g., a blue tarp), the ranges in color hue were not as close as we would have thought. This difference could be more explainable in the Non-Blue Tarp data as there are many different images that could be captured. Variance in the hues in the training data occurred, across images classified as Non-Blue Tarp and Blue Tarp. All of these classes (including the vague “Various Non-Tarp”) were clumped into the one Non-Tarp class for the sake of this project. But that does not explain the different hues across all three color values, including the hue value for the colors green and blue in the training data observations classified as Blue Tarp, when compared to the Non-Tarp class.

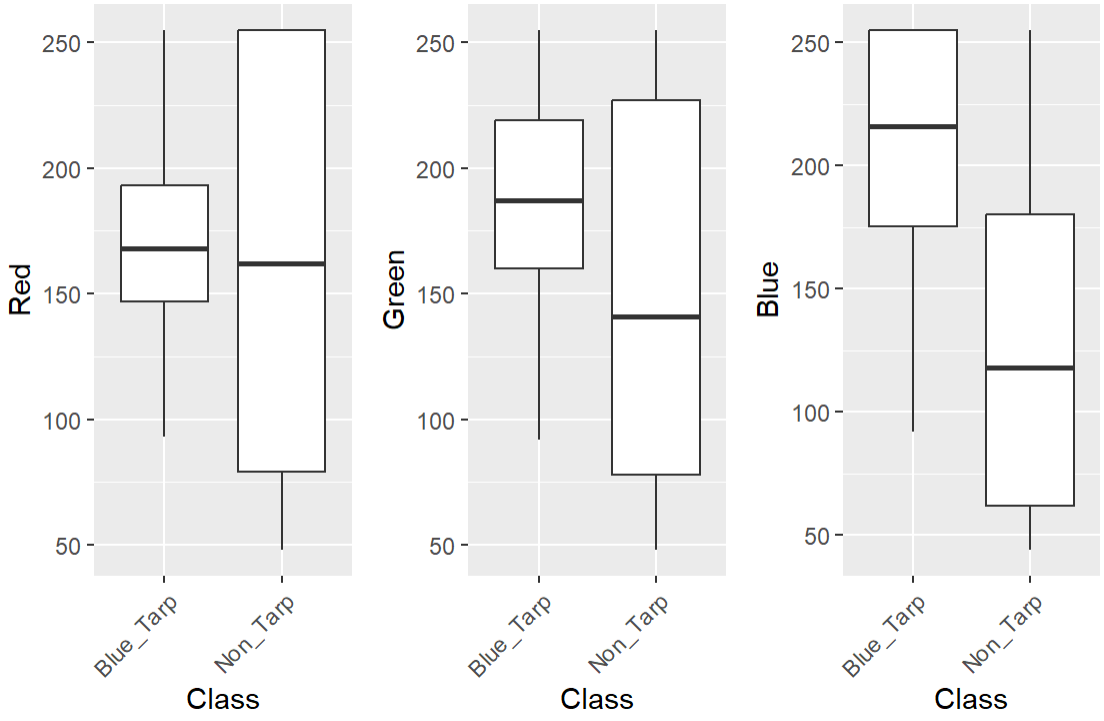


Figure 2. Boxplots of Categorical Variables for Training Data

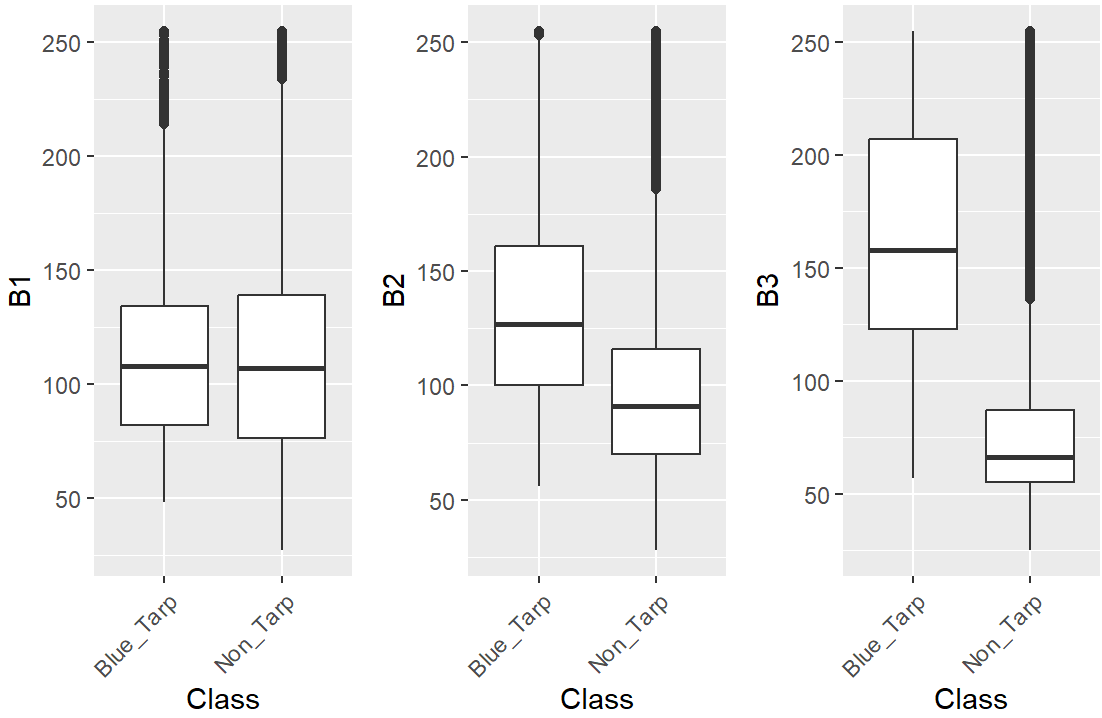


Figure 3. Boxplots of Categorical Variables for Holdout Data

Unsurprisingly, the box plots for the training set show the predictor variable with the largest difference between the median values for the blue tarp pixels as compared to non-blue tarp pixels is the Blue variable. The Green variable has the second largest difference between the median values for the two levels of the binary response variable. The Red variable does not have much a difference between the two levels of the binary response variable. The box plots for the holdout set show a similar result: one predictor has a large difference between the two classes, one variable has a smaller difference, and one variable has nearly no difference in the median values for the two classes. Thus, we assigned the colors blue, green, and red to these predictor variables, respectively, as we assumed the training and test sets would reflect similar trends.

Both the training and holdout datasets are imbalanced. As shown in Figure 4 below, there are significantly more Non_Tarp data points than Blue_Tarp data points. This imbalance is consistent with the nature of data collected. Blue tarps would only represent a small fraction of the aerial images captured.

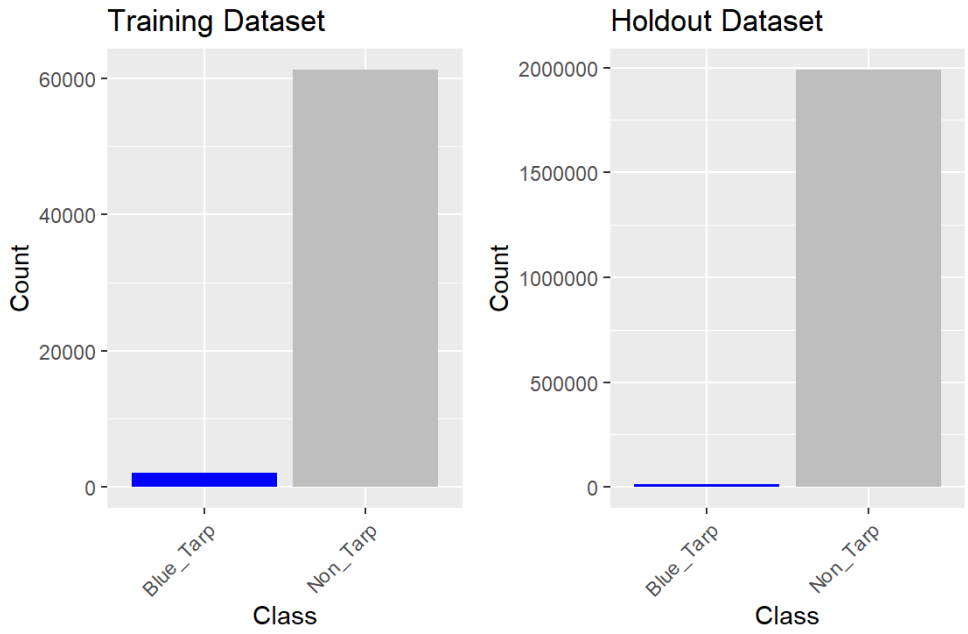


Figure 4. Plots of Class Variable Distribution of Training and Holdout Datasets

III. Description of Methodology

To build our statistical models, our team utilized the software R and tidy packages, tidymodels and tidyverse.

We used three different model approaches for this classification problem: logistic regression, linear discriminant analysis (LDA), and quadratic discriminant analysis (QDA). Each model was built and trained on the training dataset with the Class variable as the response and the Red, Green, and Blue variables as the predictors. These models were then tested using the holdout dataset.

In order to evaluate the performance of each model, we utilized ROC curves, confusion matrices, and model metrics. The model metrics we evaluated were accuracy, true positive rate (TPR)/sensitivity, false positive rate (FPR), precision, ROC AUC, and F-measure.

Due to the nature of this project, we want to ensure that the displaced people in Haiti with blue tarp temporary shelters are not improperly missed by our models and resources, such as rescue workers' time, are not wasted by going to areas that do not have people in need. We put an emphasis on the F-measure and sensitivity when evaluating the model metrics since decreasing the false negative rate and maintaining precision is vital.

We also needed to check if our models were overfitting the training dataset. In order to do this we performed a 10-fold cross validation and compared the model metrics between the training and holdout datasets. We ensured there were no major discrepancies between the cross validation results and model metrics using the holdout dataset when compared to the results and metrics of the model using the training dataset.

Since we are working with an imbalanced dataset we explored adjusting the threshold of each model in order to improve model performance. Lowering the threshold makes it less difficult for an observation to be classified as a blue tarp. Therefore, we would expect lowering the threshold to result in a lower false negative rate and a higher true positive rate. The goal of this project is to save human lives by correctly classifying blue tarps, so our models should reduce the false negative rate to minimize the number of displaced people that the model fails to recognize. Adjusting the threshold helps account for the imbalance and make the models more sensitive to the minority class, reducing the false negative rate. The F-measure is useful for imbalanced datasets since it takes into account both precision and recall. We chose thresholds where the F-measure was maximized.

Although accuracy is often used to assess model performance, it can be misleading in imbalanced classification problems because models that solely predict the majority class can still achieve a high level of accuracy. However, such a model would be of no use in predicting the minority class. If a model classified all observations in the holdout set as "Non_Tarp", the accuracy would be high, 98%. However, the model would not identify any pixels as containing blue tarps, which would not aid relief workers in identifying where to find displaced people.

IV. Results

In order to identify a statistical model that efficiently and accurately recognizes the location of blue tarps in aerial imagery, three models were built and fit. An additional null model was fit to serve as a reference comparison.

To fit the models, a random seed of 1 was set in R. Initial workflows were built for the models and model performance was assessed using a variety of metrics, including accuracy, kappa, F-measure, and the area under the ROC curve (AUC).

Table 1 and Figure 5 below summarize the performance of the classification models before any threshold selection. The ROC AUC values are high (close to 1) and similar for the logistic regression, linear discriminant analysis (LDA), and quadratic discriminant analysis (QDA) models, particularly in comparison to the null model. The ROC AUC of the logistic regression model is slightly larger than the other two models. Based on only this metric, we would select the logistic regression model as the best model.

Table 1. Metrics for the Classification Models (No Threshold Selection)

model	dataset	accuracy	f_meas	sens	spec	kap	roc_auc
Reference	Test	0.993	NA	0.000	1.000	0.000	0.500
Logistic Regression	Test	0.990	0.583	0.988	0.990	0.579	0.999
LDA	Test	0.982	0.399	0.839	0.983	0.392	0.992
QDA	Test	0.996	0.714	0.695	0.998	0.712	0.992
Reference	Train	0.968	NA	0.000	1.000	0.000	0.500
Logistic Regression	Train	0.995	0.923	0.885	0.999	0.921	0.999
LDA	Train	0.984	0.762	0.801	0.990	0.753	0.989
QDA	Train	0.995	0.909	0.840	1.000	0.906	0.998

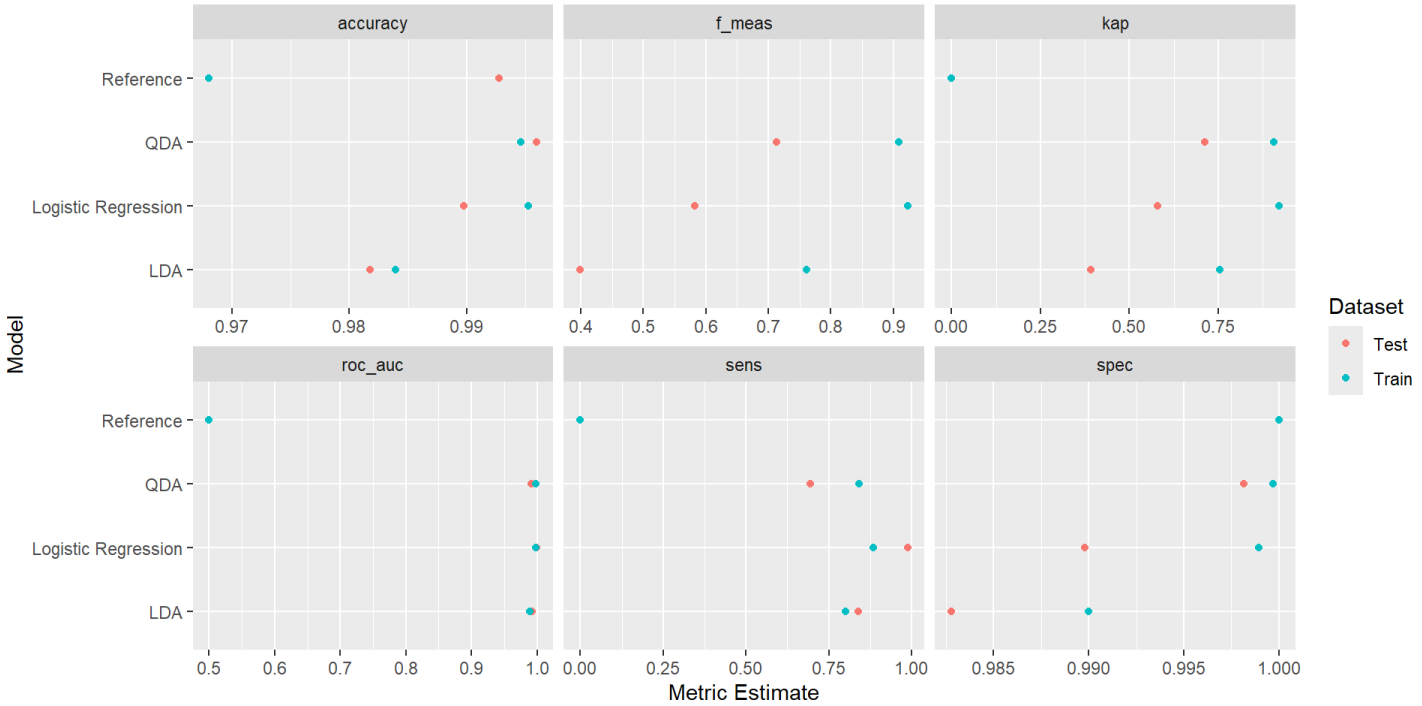


Figure 5. Metrics of the Classification Models Using the Training and the Holdout Datasets

However, looking at accuracy, the QDA model performs the best on the training and test data. Of note, the kappa value for all three models is significantly lower for the test data than the training data, which indicates possible overfitting, as the observed accuracy as compared to random chance (reference model) is higher on the test data.

Finally, the F-measure value is the highest for the QDA model. Since F1 is a measure of both precision and sensitivity/recall, the F-measure value for QDA, 0.714, indicates a useful model. The QDA model demonstrated high precision and sensitivity in identifying displaced individuals under blue tarps. Precision refers to the proportion of true blue tarp pixels accurately classified as displaced persons, while sensitivity reflects the proportion of true displaced person pixels correctly identified, including those misclassified as non-tarp pixels.

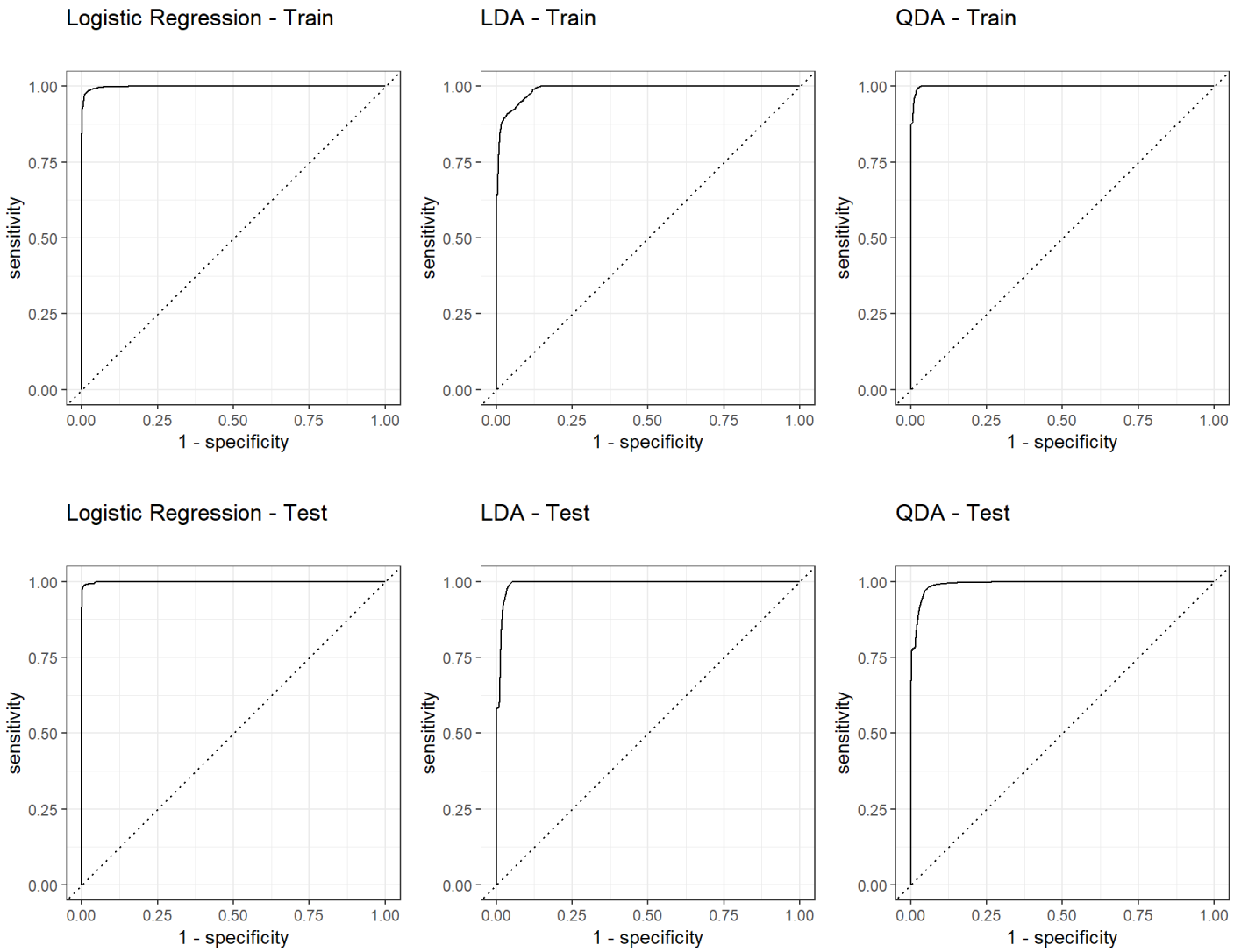


Figure 6. ROC Curve on Train and Test Data for Each Classification Model

Figure 6 shows a comparison of the model ROC AUC values for the model fit with the training data. In this case, the logistic model has the highest AUC value. Figure 7 and 8 below shows another comparison of the model ROC AUC values of the model on the train and test data, clearly showing that the logistic regression model performs slightly better than the other two classification models and the reference model for this metric.

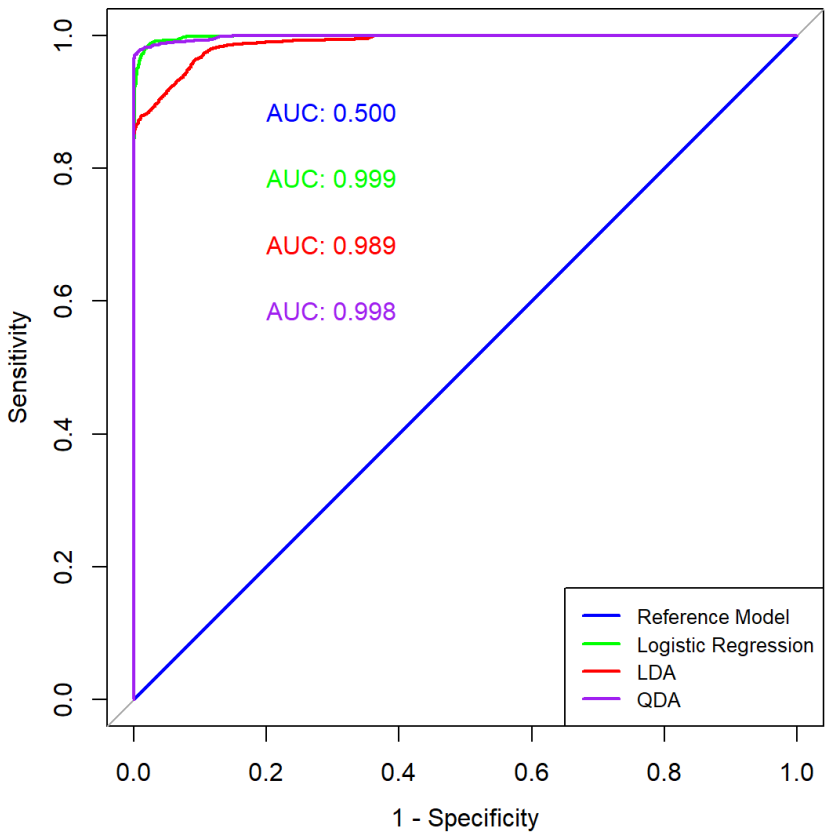


Figure 7. ROC Curve Comparison on Train Data

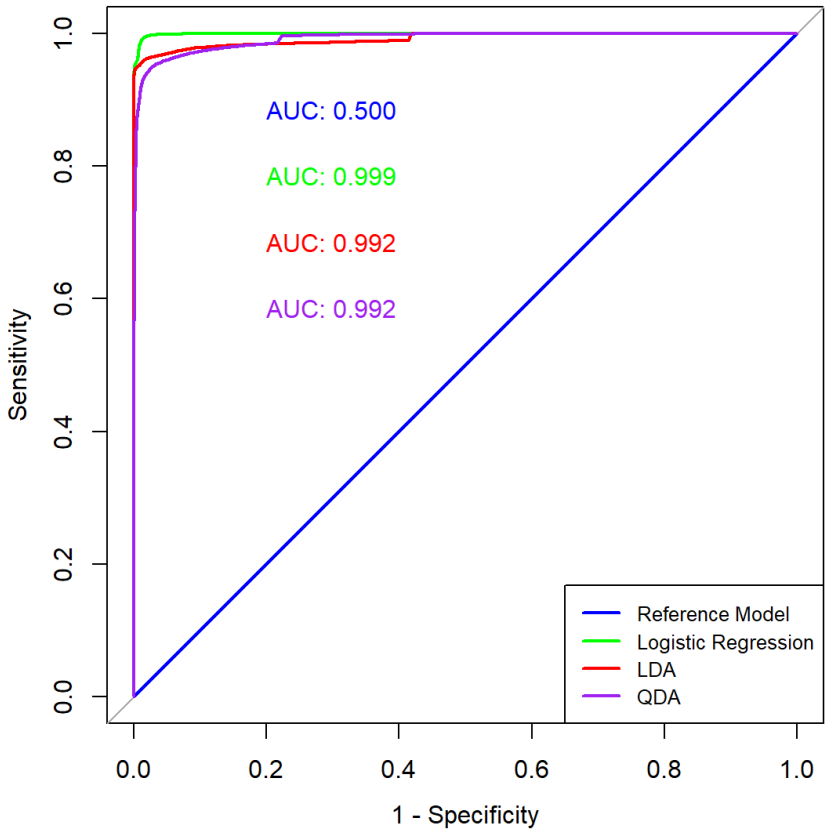


Figure 8. ROC Curve Comparison on Test Data

To check that our models were not overfitting the training dataset, we performed a 10-fold cross validation on the dataset. The results of the cross validation are below in Table 2 and Figure 9. Figure 9 above shows a comparison of the ROC AUC curves for each model between the cross validation results in black and the training data results in red. We ensured there were no major discrepancies between the cross validation results and model metrics using the holdout dataset when compared to the results and metrics of the model using the training dataset. The cross validation metrics are very similar to the training set metrics. All three ROC curves from the training and cross validation datasets look very similar. This indicates that the model is not overfitting the data.

Table 2. Metrics for the Cross Validation Classification Models (No Threshold Selection)

model	threshold	accuracy	f_meas	kap	roc_auc
Logistic Regression	No Threshold	0.999	0.904	0.904	0.999
LDA	No Threshold	0.985	0.454	0.447	0.994
QDA	No Threshold	0.998	0.881	0.881	0.998

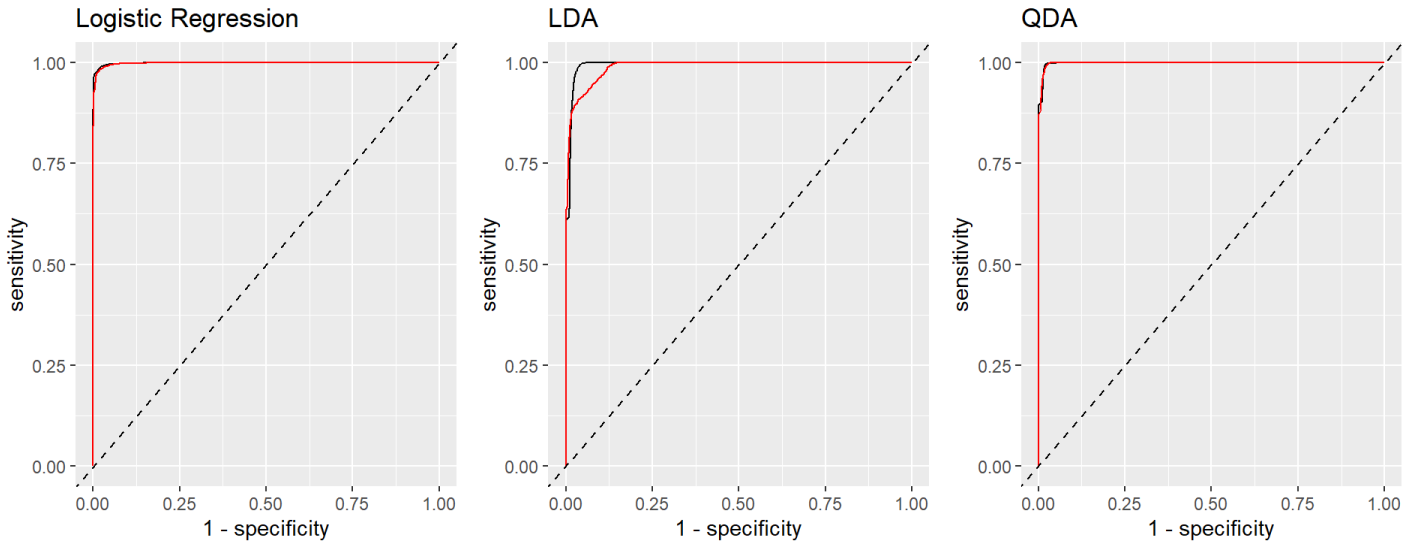
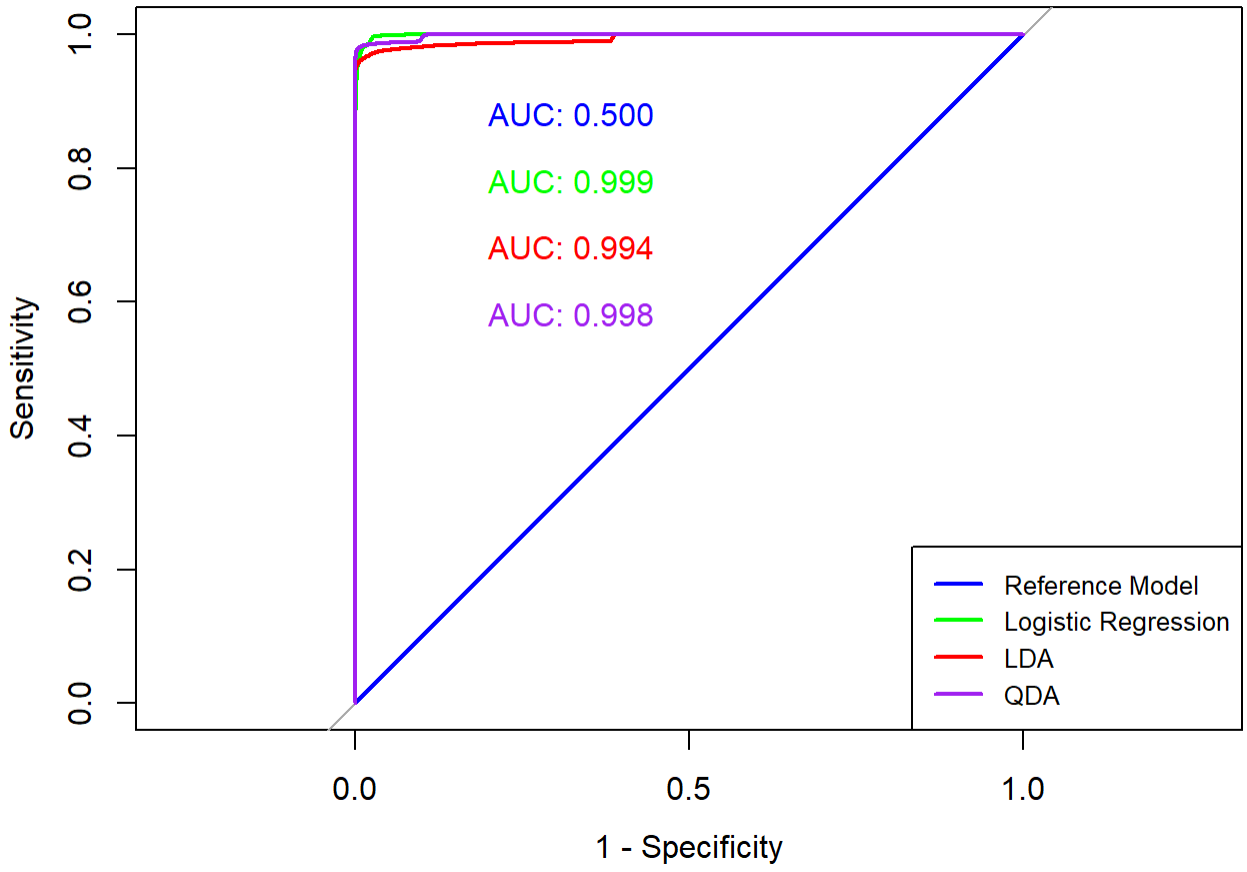


Figure 9. ROC Curve from Cross Validation for Each Classification Model

Figure 10 below overlays the ROC AUC curves for a comparison, for the models fitted on the 10-fold cross validation samples. Similar to the ROC AUC values fitted on the training and test set, the logistic regression model had the highest AUC ROC value.



In order to address the imbalance in our dataset, we performed a threshold selection analysis. Table 3 and Figure 11 below are the results of the analysis. Figure 11 shows a comparison of the ROC AUC curves for each model between the cross validation results with the selected threshold in black and the training data results in red. By optimizing the threshold, our models become more sensitive to the minority class, Blue_Tarps, reducing the false negative rate. We chose thresholds where the F-measure was maximized since the metric takes into account both precision and recall. The F-measure improved for some of the models in Table 3 compared to those in Table 2.

Table 3. Metrics for the Cross Validation Classification Models (With Threshold Selection)

model	threshold	dataset	accuracy	f_meas	sens	spec	kap	roc_auc
Logistic regression	0.25	CV	0.999	0.911	0.912	0.999	0.910	0.999
LDA	0.50	CV	0.985	0.454	0.796	0.986	0.447	0.994
QDA	0.25	CV	0.998	0.893	0.873	0.999	0.893	0.998

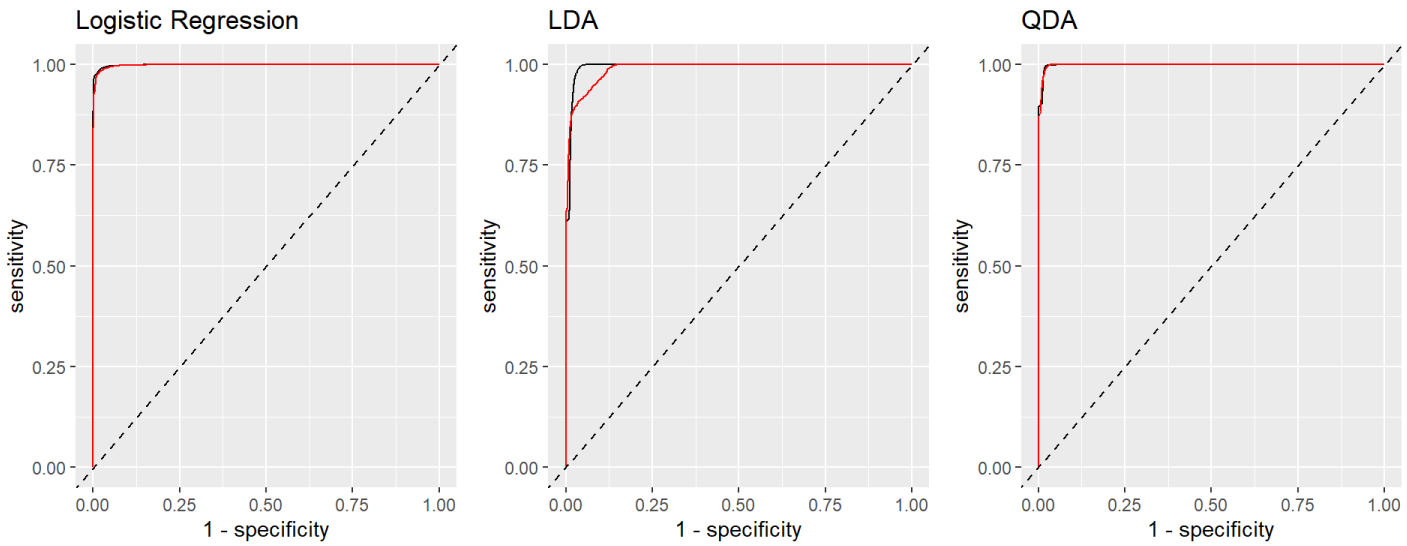


Figure 11. ROC Curve from Cross Validation for Each Classification Model with Threshold Selection

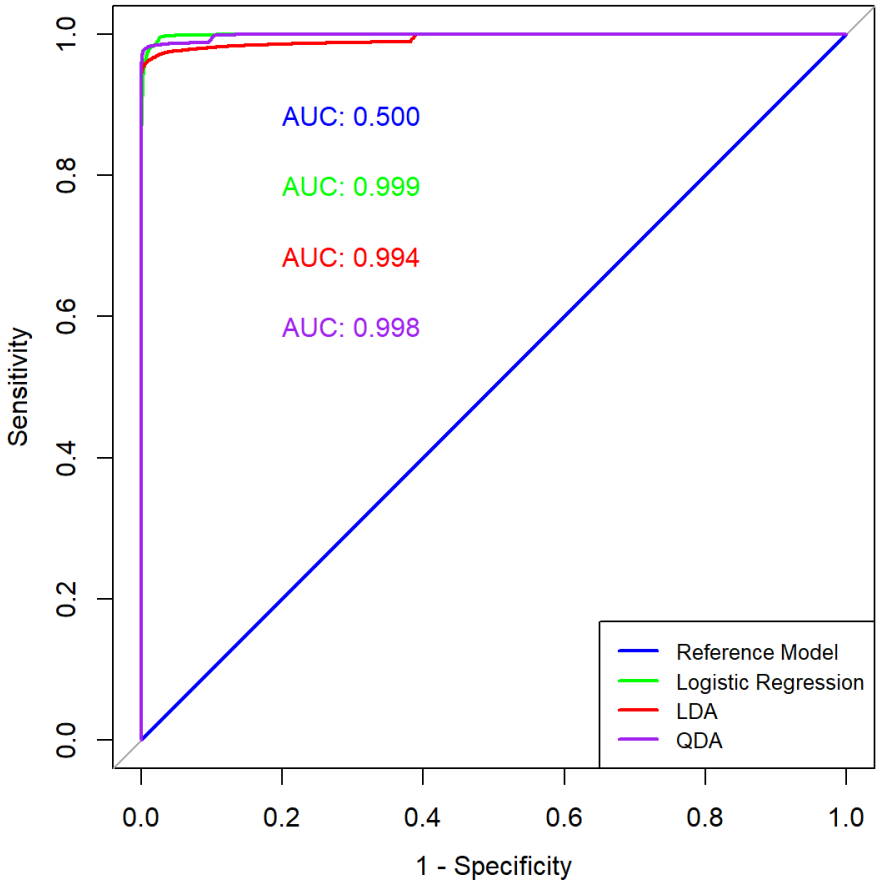


Figure 12. ROC Curve Comparison on Cross Validation at Threshold Selection

In order to address the imbalance in our dataset, we performed a threshold selection analysis. Figure 13 and Table 4 below are the results of the analysis. By adjusting the threshold, our models become more sensitive to the minority class, Blue_Tarps, reducing the false negative rate. We chose thresholds where the F-measure was maximized since the metric takes into account both precision and recall. For the LDA model, Figure 13 shows that there is no significant change in the F-measure value as the threshold value varies. This is likely why the LDA model has a higher threshold value of 0.84, when compared to the logistic regression threshold (0.21) and the QDA threshold value (0.22) that maximizes the F-measure.

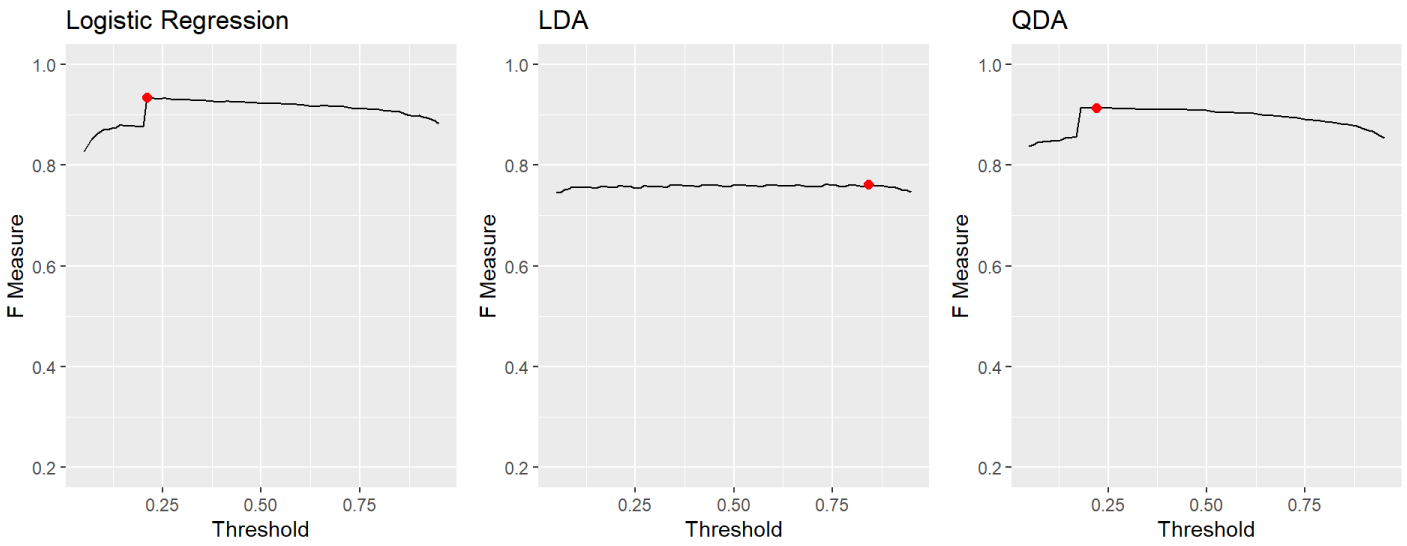


Figure 13. Threshold Scan of Each Classification Model using the F-Measure

Table 4. F-Measure at Threshold

model	threshold	f_meas
Logistic Regression	0.21	0.583
LDA	0.84	0.399
QDA	0.22	0.714

Once the threshold selection analysis was complete, we re-ran each of the classification models using these thresholds. The ROC AUC and accuracy of all three models are high and there is minimal deviation between the metrics for the test and train datasets. Based on these metrics, the logistic regression model performs the best on test data. This is slightly different from the models with no threshold selection where the QDA model outperformed logistic regression based on accuracy.

When evaluating the three models based on the F-measure the QDA model performs the best on the test data. This is consistent with our model evaluation with no threshold selection.

Table 5. Metrics for the Classification Models (With Threshold Selection)

model	dataset	ROC	Threshold	Accuracy	TPR	FPR	Precision	F_meas
Logistic Regression	Test	0.999	0.21	0.965	0.993	0.036	0.168	0.288
LDA	Test	0.992	0.84	0.984	0.755	0.015	0.273	0.401
QDA	Test	0.992	0.22	0.995	0.761	0.003	0.621	0.684
Logistic Regression	Train	0.999	0.21	0.996	0.926	0.002	0.943	0.934
LDA	Train	0.989	0.84	0.985	0.759	0.008	0.766	0.763
QDA	Train	0.998	0.22	0.995	0.867	0.001	0.966	0.914

The threshold scan for maximizing the F-measure lowered the threshold for logistic regression and QDA and raised the threshold for LDA. As seen in the plot of the threshold against F-measure in Figure 13, the F-measure is relatively constant across all thresholds, so the fact that the selected threshold of 0.84 is higher than the other two models does not seem significant. Comparing the confusion matrices for QDA at a threshold of 0.5 and the selected threshold 0.22, demonstrates the lower threshold resulted in a lower false negative rate. The confusion matrix below at a threshold of 0.5 has a false negative rate of 0.305. The sensitivity, or true positive rate, is 0.694.

However, after lowering the threshold to 0.22, we see more observations are predicted to be “Blue_Tarp”, and the false negative rate lowers to 0.239. The sensitivity (TPR) increases to 0.761. Thus, the lower threshold for the QDA model could lead to more displaced people receiving aid, as fewer blue tarps were not classified as not blue tarps.

Due to limited resources, relief workers might also need to consider the effect lowering the threshold has on the false positive rate. The false positive rate, meaning the proportion of non-blue tarp pixels classified as blue tarps, for the 0.5 threshold confusion matrix is 0.0018. The false positive rate increases to 0.0034 when the threshold is lowered to 0.22.

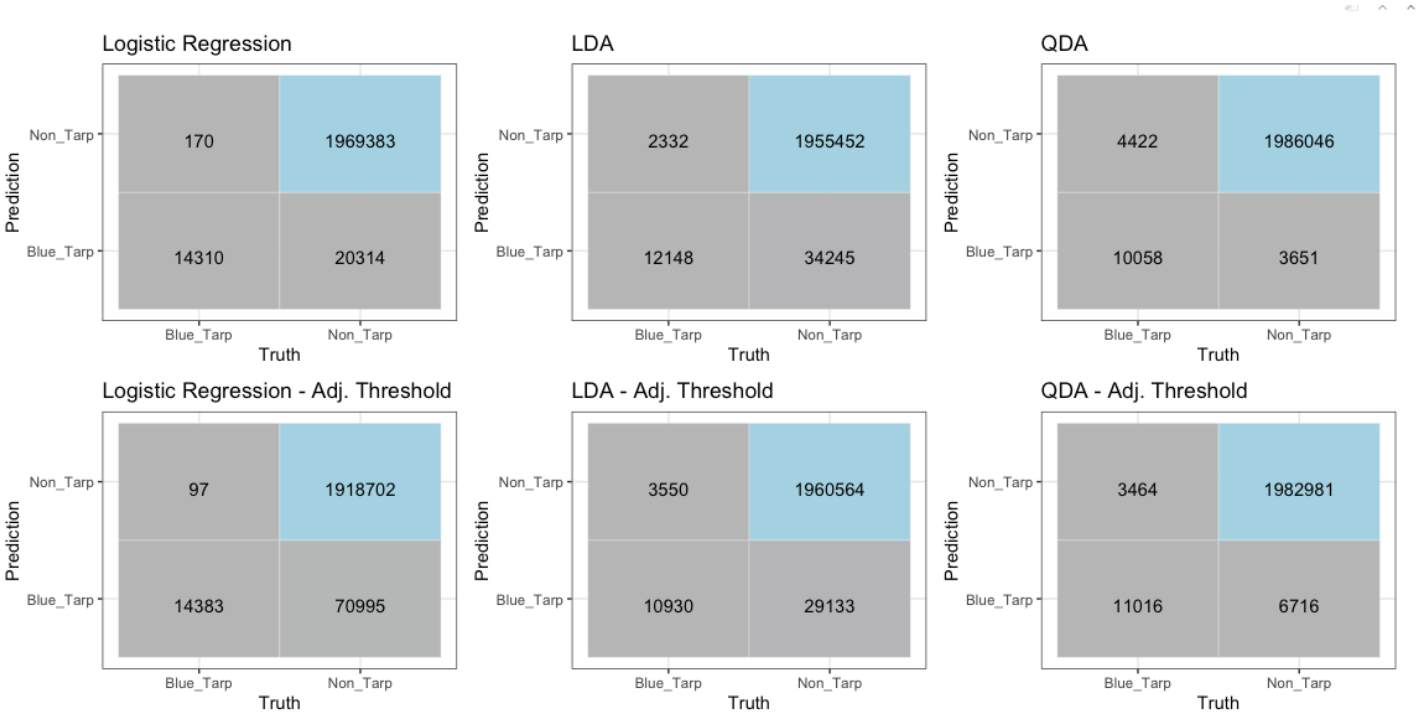


Figure 14. Confusion Matrix for Each Model On Holdout Data

In order to further analyze the precision and sensitivity of the models, we investigated the precision and recall curves. The values in Table 6 for the models on the test data indicate that the logistic model had the highest PR AUC value, followed by the QDA model. Figure 15 shows the comparison between these PR AUC curves, in which the models that are performing better (logistic regression and QDA) have a curve that trends towards the upper right hand corner of the visual.

Table 6. Precision-Recall AUC for the Classification Models on Train and Test Data

model	dataset	PR_AUC
Logistic Regression	Test	0.969
LDA	Test	0.683
QDA	Test	0.762
Logistic Regression	Train	0.975
LDA	Train	0.859
QDA	Train	0.966

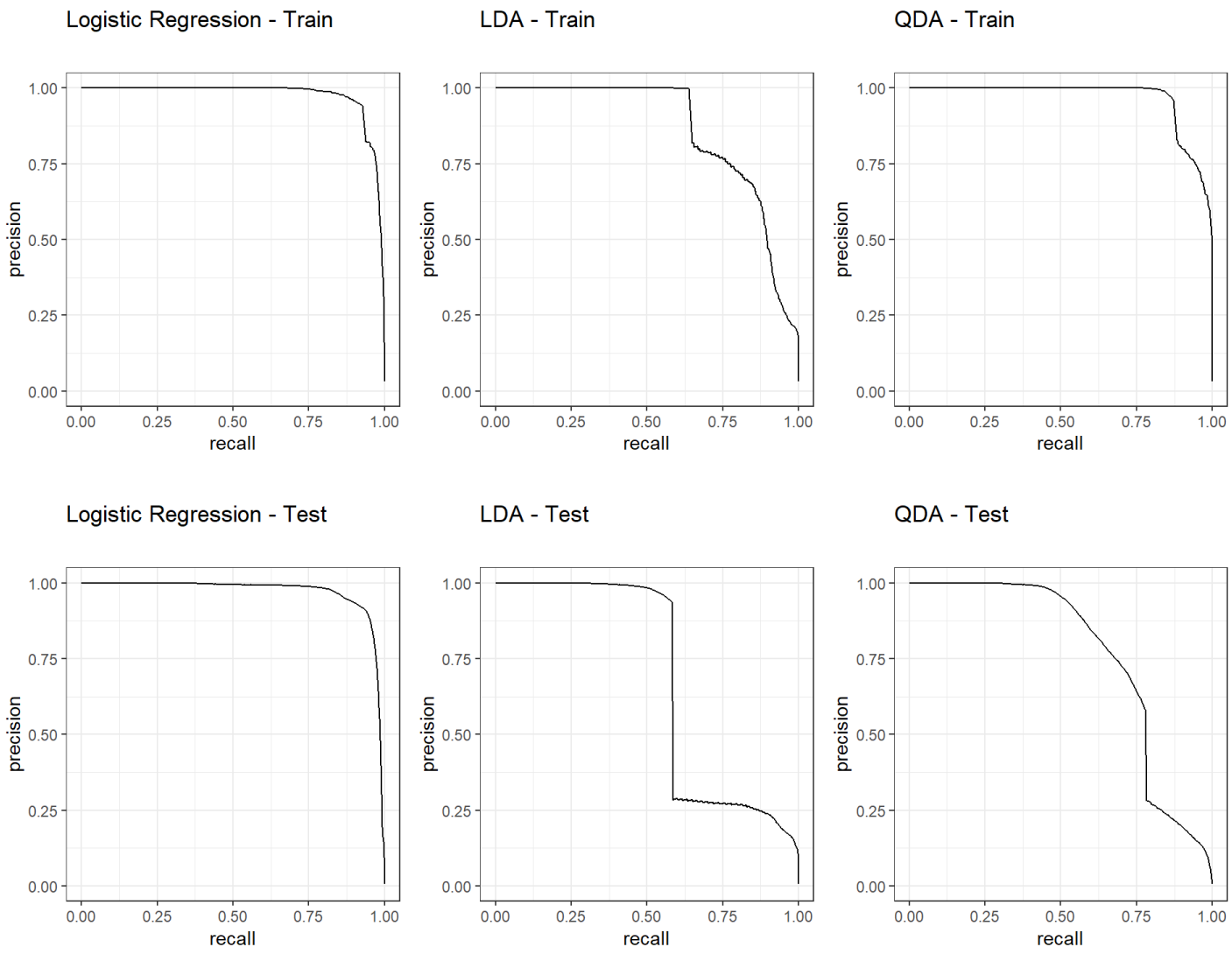


Figure 15. Precision-Recall Curve on Train and Test Data for Each Classification Model

Figures 16 and 17 show a comparison of precision-recall curves on both the train and test data. In both instances, the logistic regression model performed the best for this metric.

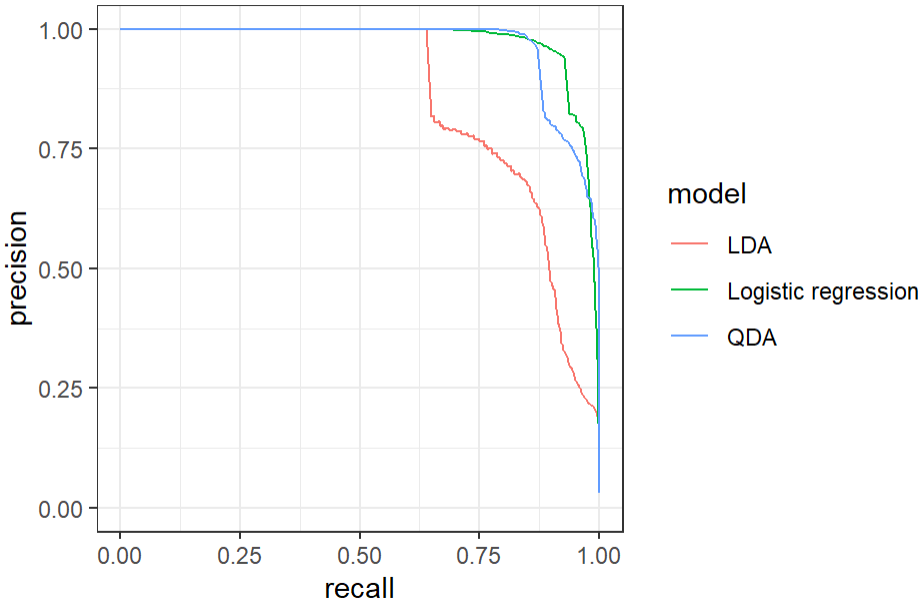


Figure 16. Precision-Recall Curve Comparison on Train Data

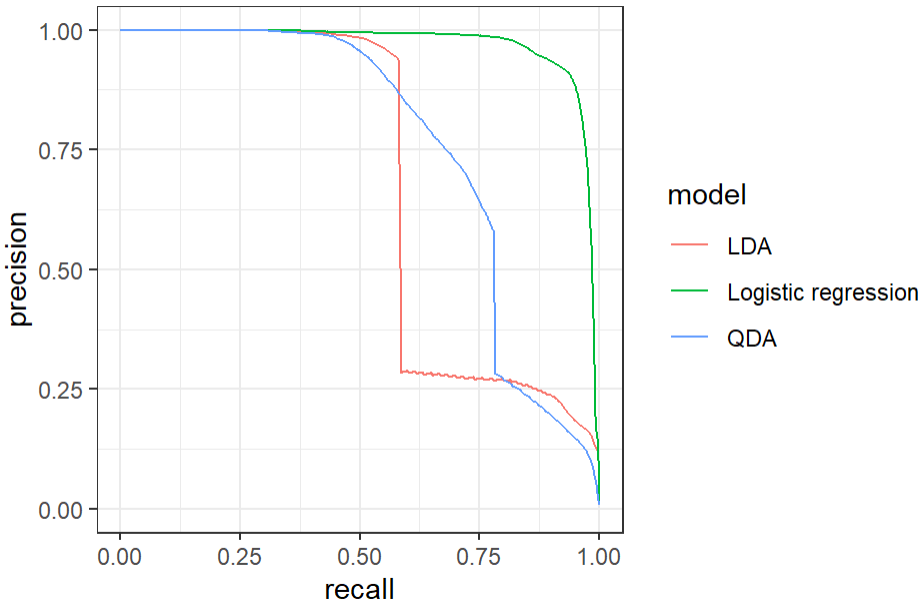


Figure 17. Precision-Recall Curve Comparison on Test Data

We also performed an analysis on the model metrics to see how the metrics from the test datasets performed on each model with and without threshold selection. The LDA and QDA models generally performed better on test data with the selected threshold. The logistic regression model performed worse on test data with the selected threshold.

The logistic regression model with threshold selection performed worse when comparing the accuracy and F-measure. The logistic regression model with selection outperformed the model without selection when considering the false positive rate. The model with threshold selection negligibly outperformed based on the true positive rate and had the same ROC AUC.

The LDA model with threshold selection performed better when comparing the accuracy. The model negligibly outperformed based on the F-measure. The model performed worse based on the true positive rate as well as false positive rate and had the same ROC AUC.

The QDA model with threshold selection performed better when comparing the true positive rate and marginally better when comparing false positive rate. The model negligibly performed worse based on accuracy and F-measure. The model with and without threshold selection had the same ROC AUC.

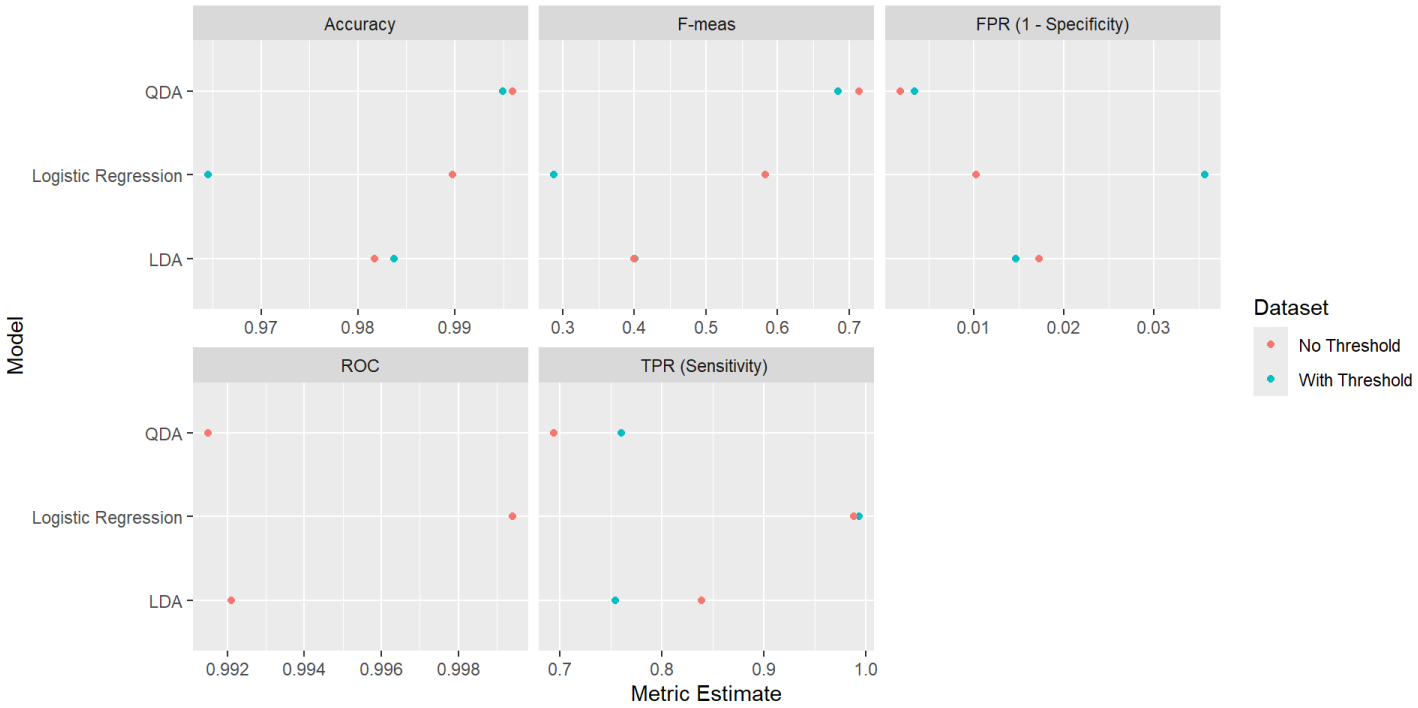


Figure 18. Metrics of the Classification Models With and Without Threshold Selection

V. Conclusion

After fitting several models, predicting probabilities on test data, and analyzing the ROC curves, we identified optimal thresholds that maximized the F-measure, addressing both sensitivity and specificity. This approach ensures that our model's predictions align well with identifying displaced individuals and minimizing the number of blue tarp pixels that were not identified, enhancing the model's utility to identify displaced persons.

The QDA model performed the best with regard to the F-measure, accuracy, true positive rate, and lower false positive rate. The F-measure value for this model was 0.684. This means that the model is effectively minimizing false positives and false negatives, performing well and identifying blue-tarp areas where displaced individuals are located. It also means that fewer resources will be expended to further investigate areas incorrectly classified with a blue tarp by the QDA model. The high TPR and low FPR values instill a fair amount of confidence in the model's classification results. Although the model performed the best, the F-measure value on the test set is only moderately high on a scale from 0 to 1, indicating that the model was only moderately better than guessing. This means that further actions to improve the performance of the model, including the tuning of hyperparameters, would be recommended if there were a limited number of resources available to rescue displaced people.

Compared to the QDA model, the LDA model with threshold selection also performed somewhat well, with a ROC AUC value of 0.992 and a TPR of 0.755 on the test data, which is close to the QDA TPR of 0.761. The F-Measure value of 0.401 was lower than the QDA model, but still indicates more value that the logistic regression F-measure value of 0.288. This means the model could be used as a secondary reference for identifying or predicting where displaced individuals are located.

Overall, the logistic regression model performed worse for all performance metrics on the test data, with lower accuracy, TPR, FPR, precision, and F-measure values, despite a high ROC AUC value of 0.999. This is likely due in part to the fact that the dataset was imbalanced and the ROC AUC metric, which is less sensitive to the performance of the minority class, in this case blue pixels, because it gives equal weight to both positive and negative classes, regardless of imbalanced distributions.

The use of the QDA model to identify displaced people would be somewhat effective to save human life. In this case, a regression model is an efficient and fast option for the extremely large test data set, and the precision value of 0.621 on the test set indicates that the model would accurately identify blue tarp areas, while minimizing the false positives. The QDA model also enables us to account for non-linear boundaries. The approach of this model would also save rescue resources from investigating areas misclassified as a location of a blue tarp that did not actually have any displaced people. However, it is possible that a non-linear classifier, such as K-Nearest Neighbors, could potentially also be effective if the relationship between the red, green, and blue pixels is (1) complex due to the variations in color hue based on time of day, the geographic area, and vegetation and (2) pixels tend to be clustered together.