



Haiti Earthquake Relief Project

GROUP 9

CLAIRE SULLIVAN, SAMANTHA ASEFI,
TERRANCE LUANGRATH, KATIE DUNNING

Background and Purpose

- Destruction caused by size 7 Haiti earthquake has left residents displaced and without resources
- It is known that individuals were taking shelter under blue colored tarps
- Aircraft imagery has been provided by Rochester Institute of Technology
- To supply these individuals with resources and care, need to efficiently supply the location of the blue tarps from images
- Our goal: Create an algorithm that takes RGB values from images to identify blue tarp amongst other terrain.





Haiti Disaster Relief: The Data

Identifying Blue Tarps

Identifying RGB Values

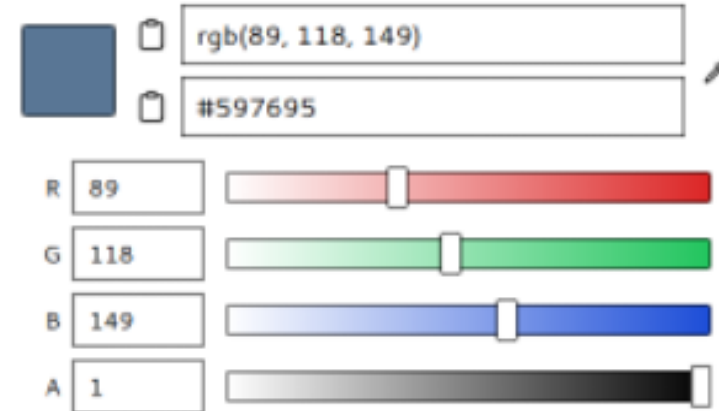
Since we're focusing on blue tarps, determine blue channel.

Select a random blue tarp point from the blue tarp test data: 89 118 146.

Start by assuming RGB are already in the correct places since channel 3 always has highest value in dataset.

Correct color. All channels probably correct, but double check.

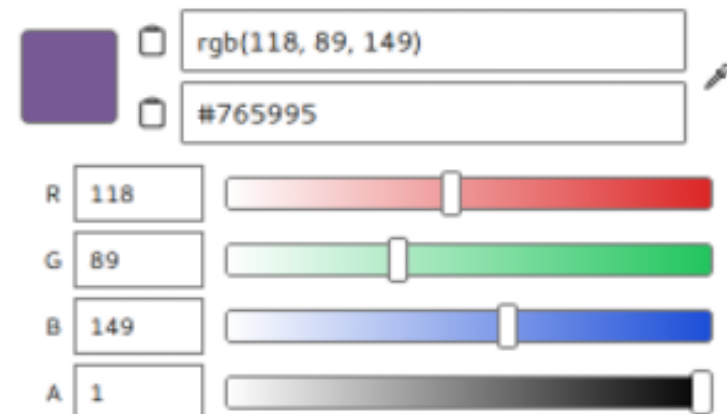
With Correct Color Channels RGB Color Picker



rgb(89, 118, 149)
#597695

R 89
G 118
B 149
A 1

With Green and Red Swapped RGB Color Picker

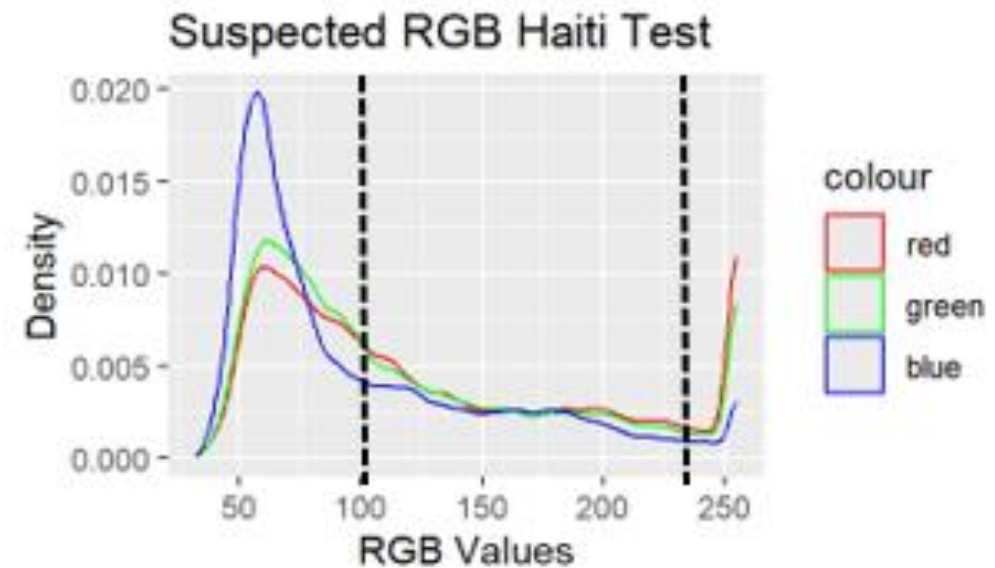
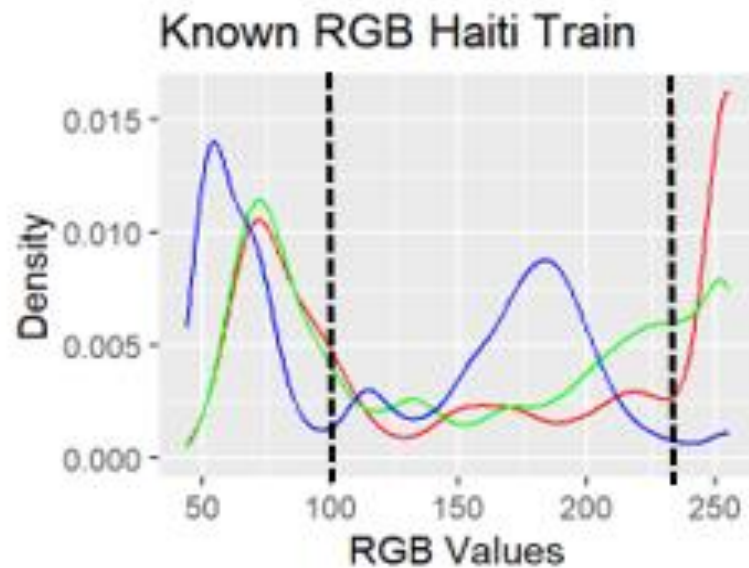


rgb(118, 89, 149)
#765995

R 118
G 89
B 149
A 1

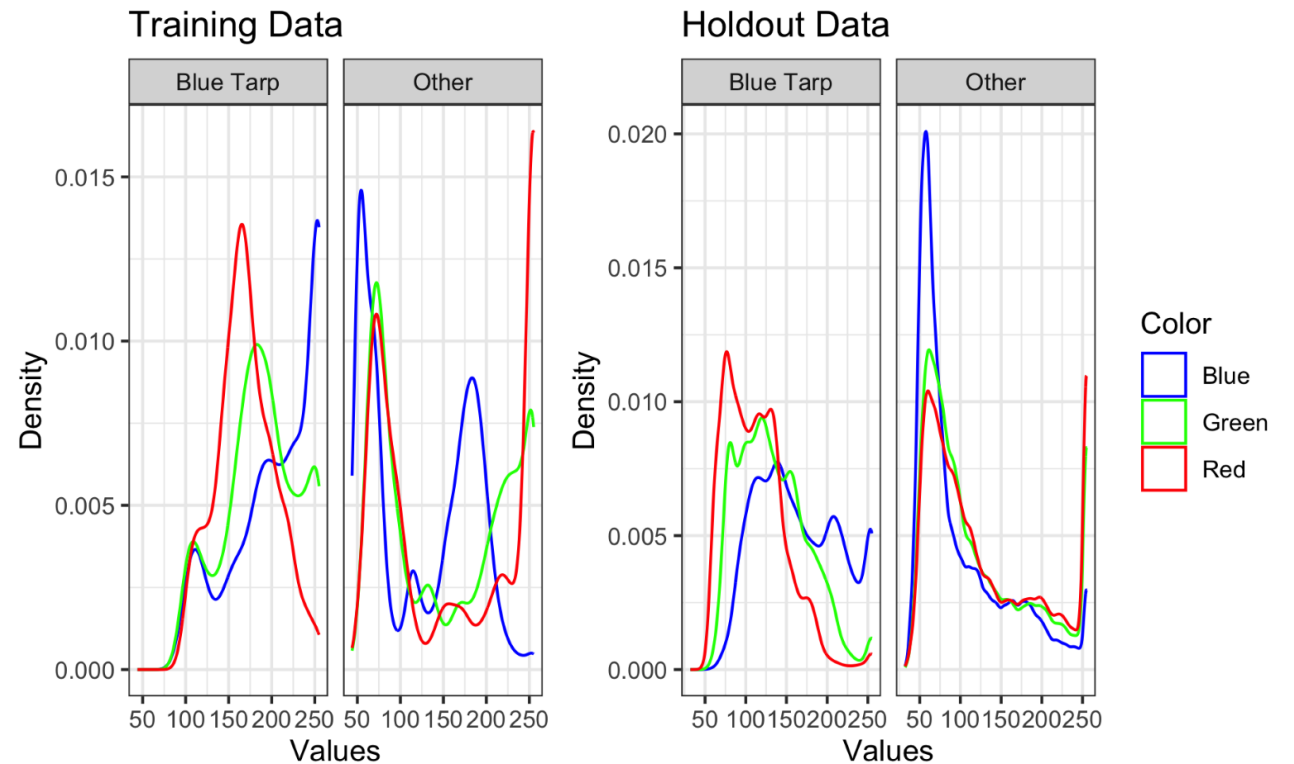
Verifying RGB Channels

- Densities of color should be similar across both datasets due to context.
- RGB given for Haiti training set and blue determined for test set.
- Educated guess on red and green for test set.
- Distinct patterning similarities at the beginning and end of graph confirming selection.



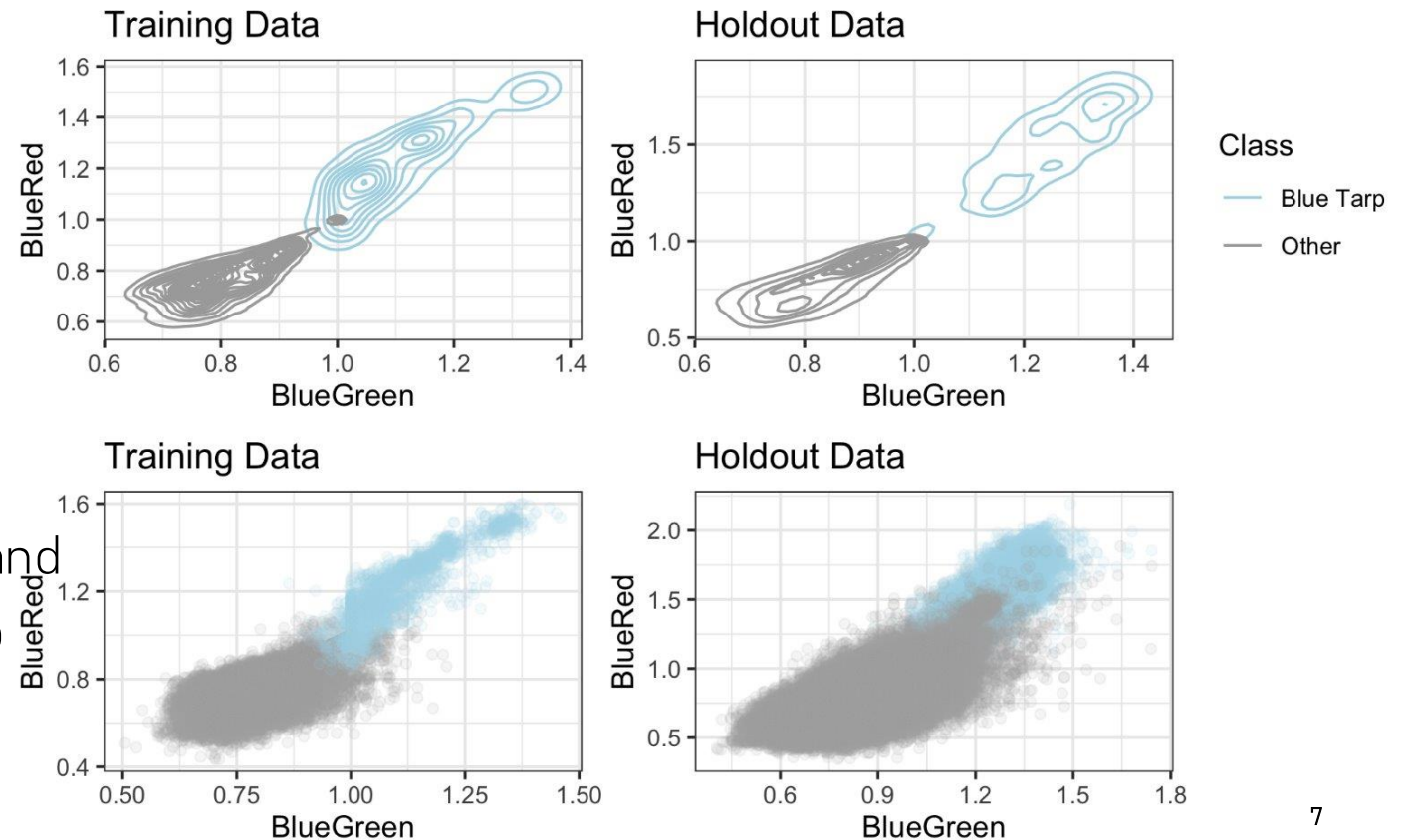
Identifying Blue Tarps using RGB Values

- High frequency of Blue in the 'Blue Tarps' class for training data
- In holdout data, RGB density plots overlap in both 'Blue Tarp' and 'Other'
- RGB values alone are not useful in visualizing blue tarps



Examining RGB Ratios in Blue Tarps

- Created 2 new color indicating variables calculated as:
 - Blue/Red called BlueRed
 - Blue/Green called BlueGreen
- See clearer class separation, with some overlap
- The blue value, relative to the red and green values, indicated a Blue Tarp





Methodology

Modeling Blue Tarp Classification

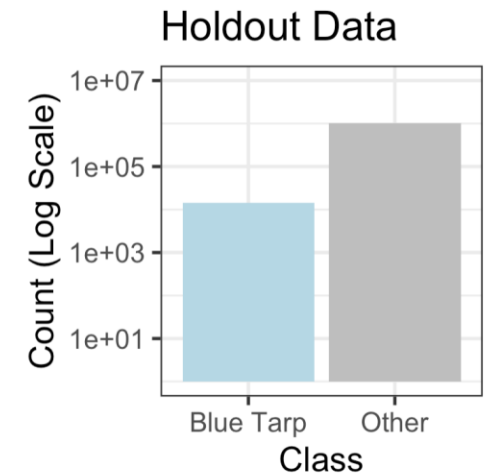
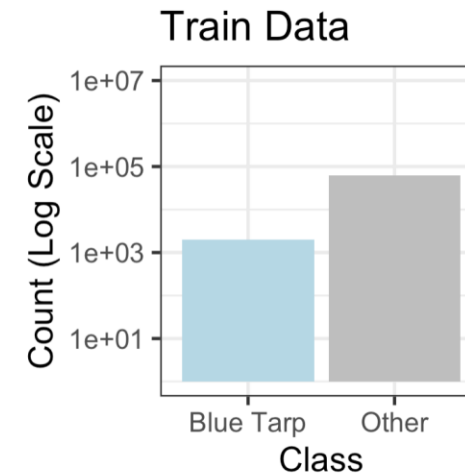
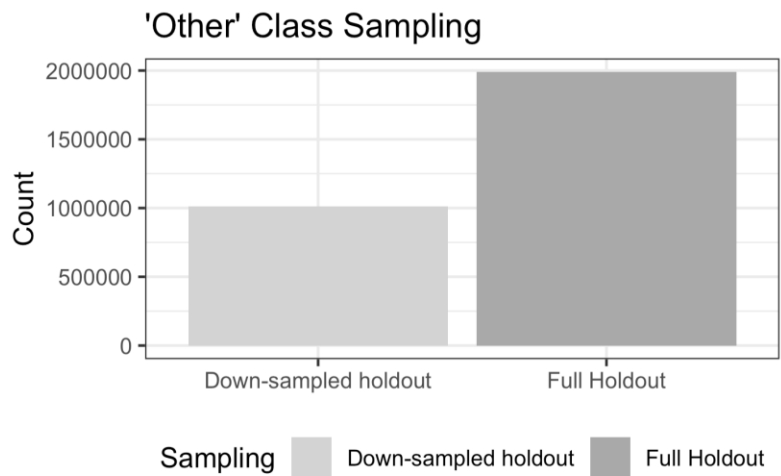


Separating classes: Blue Tarp vs Other

- To simplify the identification of persons in need, the data was reduced to two classes:
 - 'Blue Tarp': indicating displaced people
 - 'Other': indicating all other types of surface structures including vegetation, soil, rooftops, and other non-tarp structures
- Based on this leveling, the classification models seek to predict observations as being either Blue Tarps or other terrain features

Addressing Class Imbalance

- Down-sampled by opting to omit one Non-Blue Tarp data set
 - This dataset numerically was unpaired with a corresponding Blue Tarp dataset
 - Has ~1 million non-blue tarp datapoints.
 - This functions as manual undersampling of the dataset.
- Blue tarps now constitute approximately 1.41% of the test dataset.





Model Fitting

To model the classification of observations as being 'Blue Tarp' or 'Other' based on their RGB channels, the following modeling techniques were chosen:

- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- K-Nearest Neighbors
- Penalized Logistic Regression
- Random forest
- Support Vector Machines (SVM)

Model Tuning

- The following models had hyperparameter tuning based on latin hypercube grids and selected for optimal AUC values:

Model	Tuning Parameter
K-Nearest Neighbors	# Neighbors
Penalized Logistic Regression	Penalty
	Mixture
Random Forest	Mtry
	Minimum N
SVM	Cost
	Scaling Factor
	Degree



Model Thresholding and Metrics Selection

- Thresholds were selected based on maximal f measure (f1 score)
 - Thresholds were bounded below 0.5 to force more liberal models
 - F measure is known for being a reliable statistic for imbalanced data
 - F measure threshold optimization yielded reasonable threshold values (above 0.1)
 - J-index optimization yielded unreasonably low threshold values (below 0.1)
- To assess model performance on predicting the holdout dataset, the following metrics were collected:
 - Accuracy, kappa, f measure, j_index, sensitivity, ROC/AUC
 - To choose a top-performing model, the most weight was put on F measure
- In addition to statistical comparison, logistical conclusions were drawn based on the confusion matrices and classifications of each of the models



Results

Predicting Blue Tarps

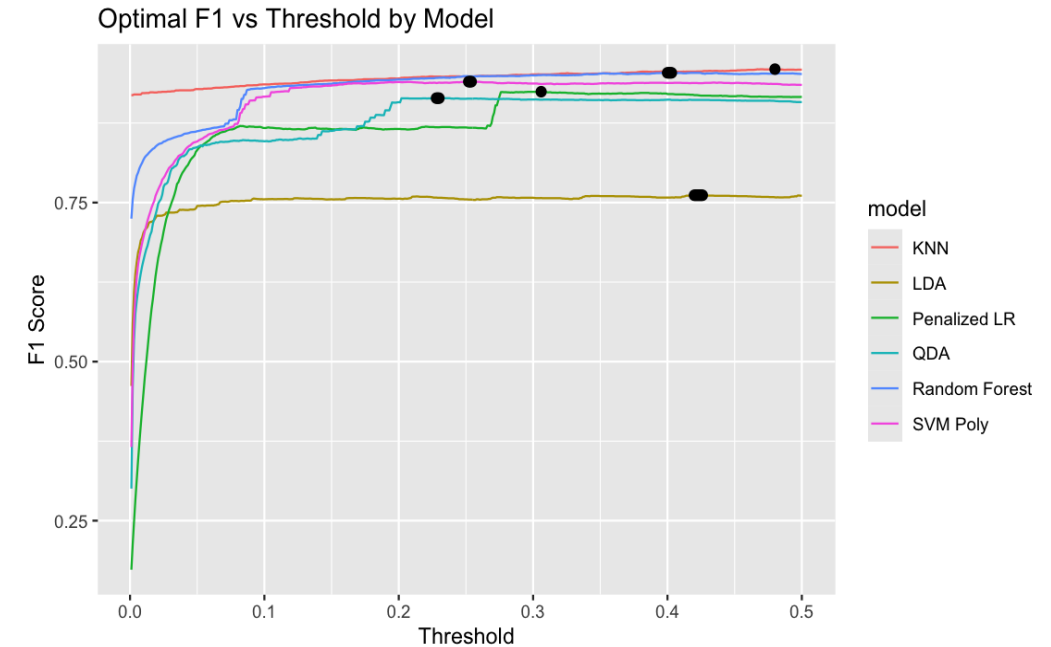
Tuning

- Table with tuning hyperparameters and Model

Model	Tuning Parameter	Value
K-Nearest Neighbors	# Neighbors	17
Penalized Logistic Regression	Penalty	0.0000283
	Mixture	0.643
Random Forest	Mtry	2
	Minimum N	14
SVM	Cost	0.959
	Scaling Factor	0.0746
	Degree	2.45

Threshold Selection

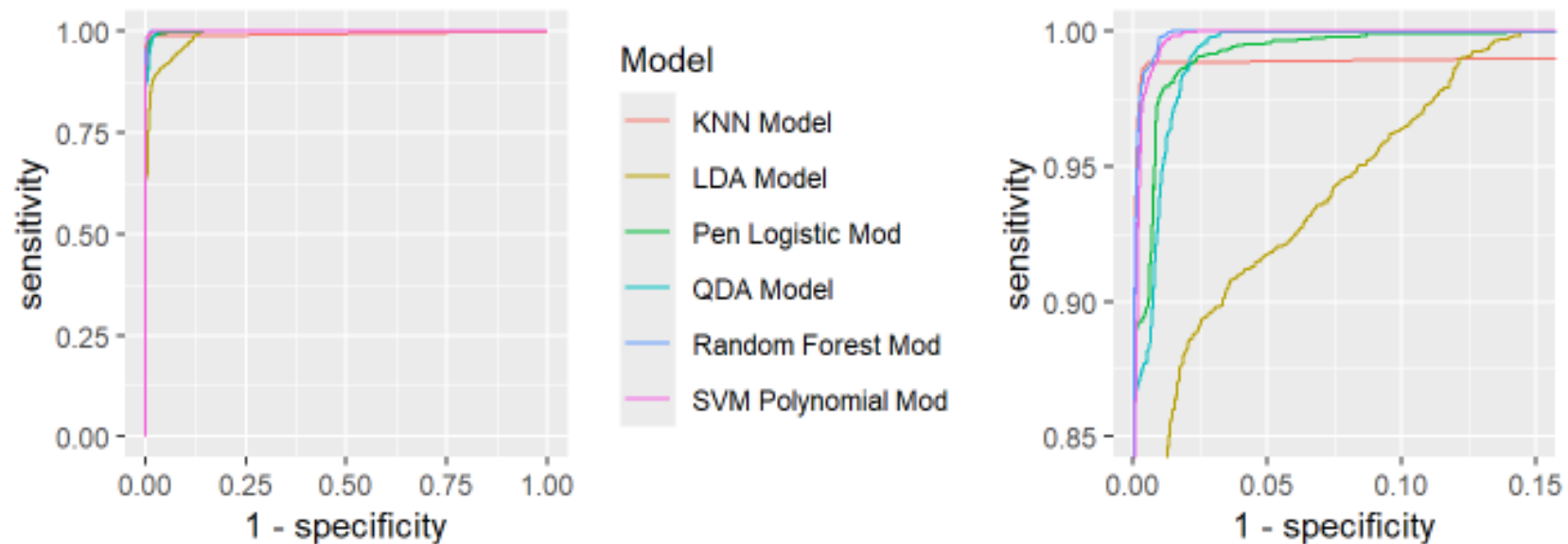
- F-measure thresholds all reasonably valued (none at nearly 0 and none not thresholding at all)
- LDA and KNN trend closer to default threshold, with Random Forest close behind
- Penalized Logistic, SVM polynomial, and QDA more discerning



model	f_meas	Thresholds
LDA	0.761	0.420
QDA	0.908	0.228
KNN	0.959	0.480
Pen Logistic	0.916	0.306
Random Forest	0.954	0.396
SVM Polynomial	0.936	0.251

CV Results

- All models demonstrate high F1-Score and ROC AUC.
- LDA clearly performing worse than other models.
- Best three models by ROC AUC are: Random Forest, SVM Polynomial, and Penalized Logistic Model.
- These are the three models we will focus on most closely.

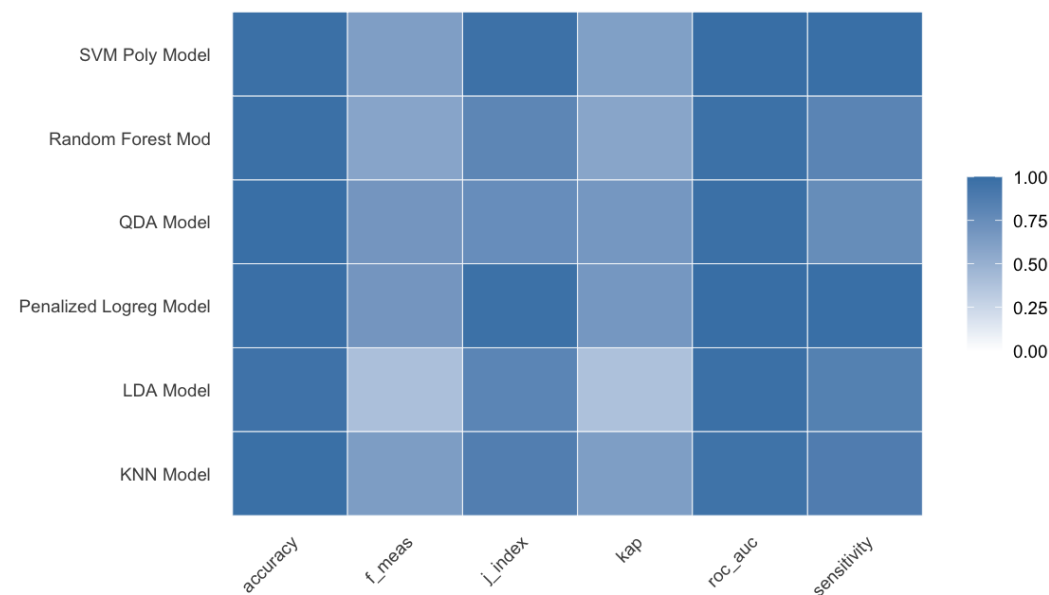


model	f_meas	roc_auc
LDA	0.761	0.989
QDA	0.908	0.998
KNN	0.959	0.994
Pen Logistic	0.916	0.999
Random Forest	0.954	0.9997
SVM Polynomial	0.936	0.9995

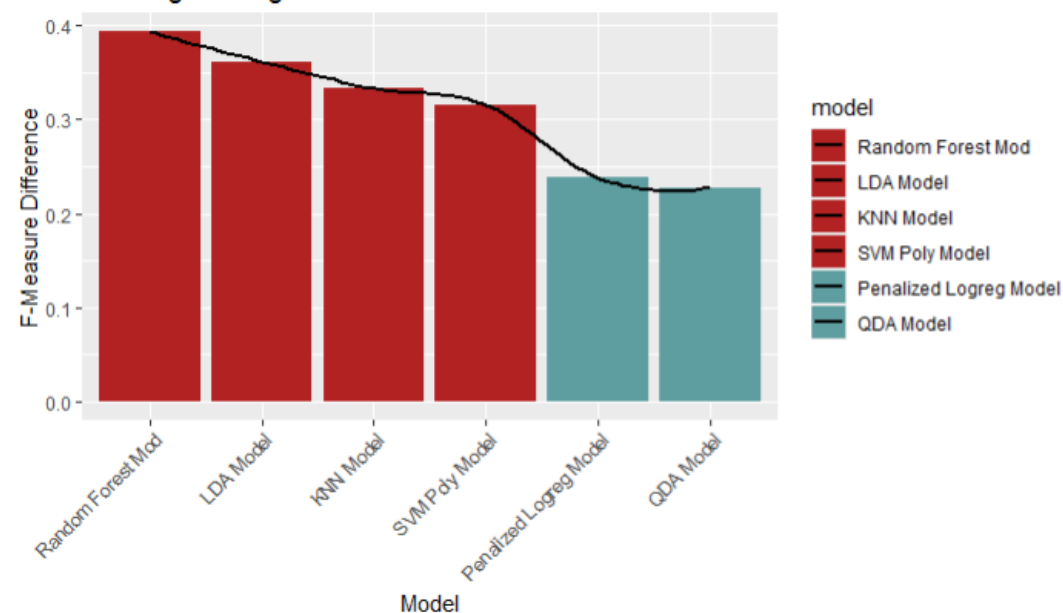
Train & Test Data Results

- Four out of the six models show larger signs of overfitting (and/or changes due to test imbalance) than the other two.
- Random forest, which had the best CV results, has the worst drop in f1-score between train and test performance.
- Based on a combination of CV results and train/test gap evaluation, we would select the **Penalized Logistic Model**.

Test Set Performance by Metric

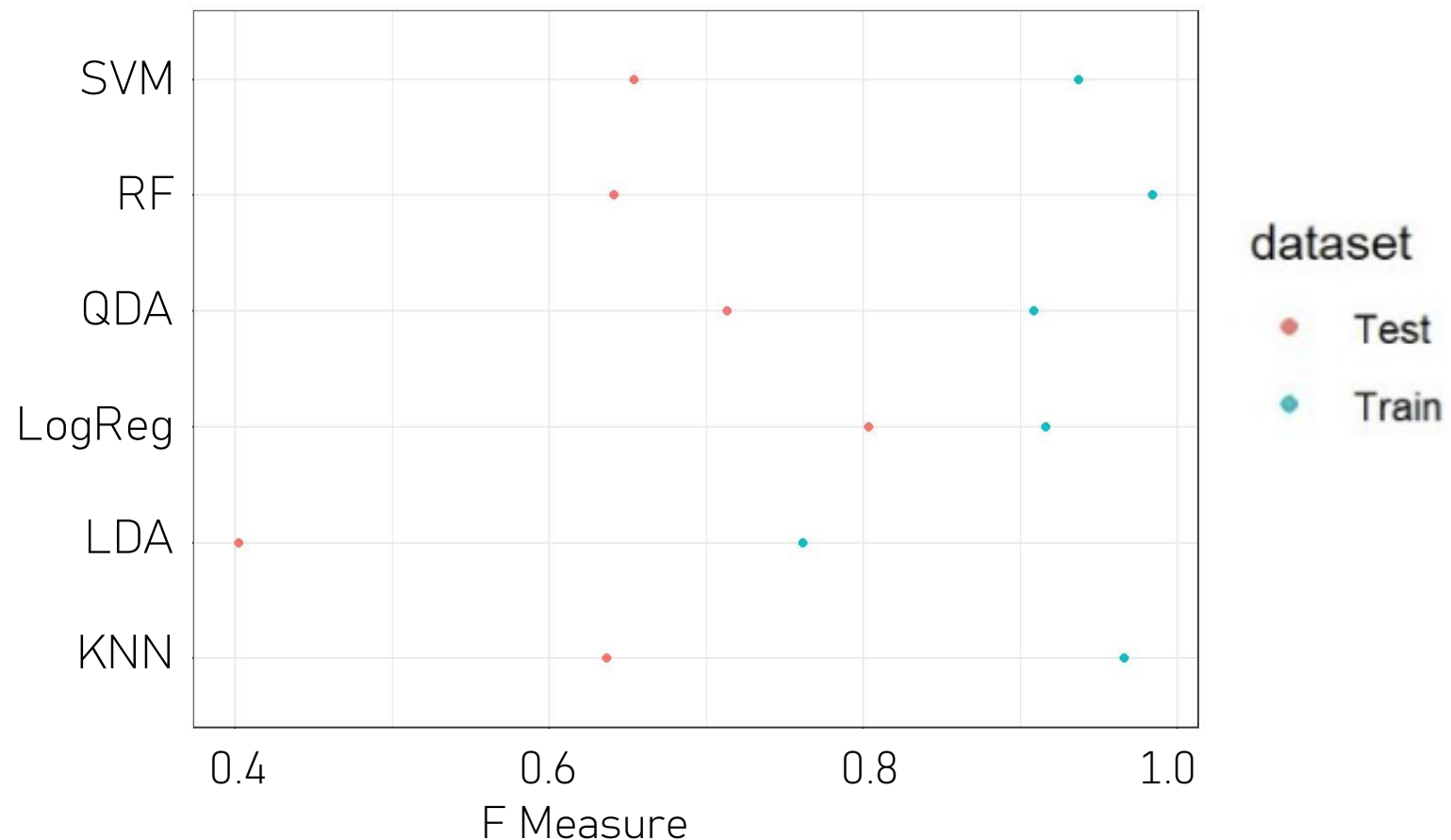


Training/Testing Differences in F-Measure



Test / Train F Measure Performance

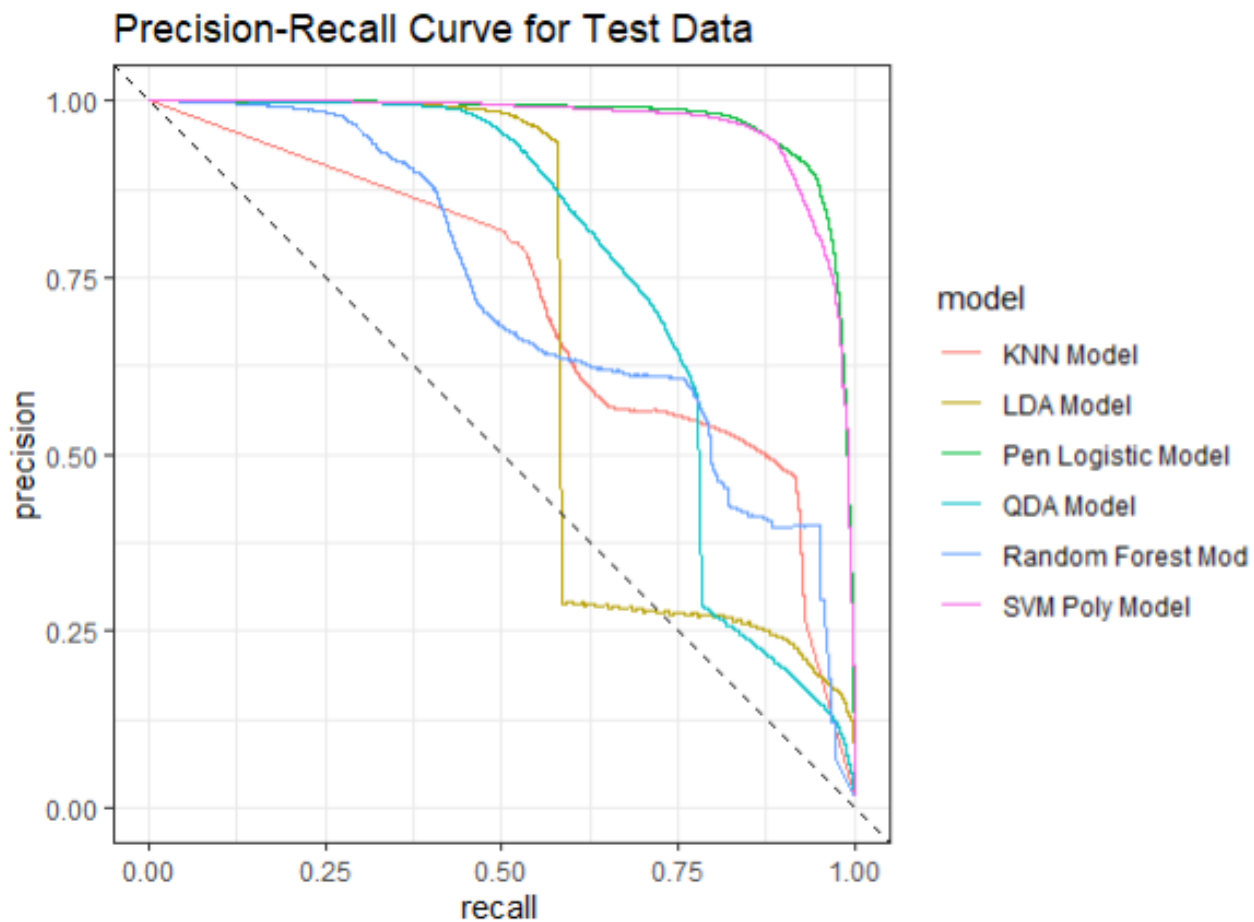
- Another representation of test vs train model performance
- Train performs better for all models
- Log reg most similar metrics
- LDA performs the worst for both test and train
- RF and KNN large difference between test and train
- Log reg best test performance



Precision-Recall Curve

Comparison of All Models

Model	precision	recall	f_meas	roc_auc
LDA	0.265	0.839	0.403	0.985
QDA	0.734	0.695	0.714	0.984
KNN	0.502	0.870	0.637	0.960
Log Reg	0.680	0.982	0.803	0.999
Random Forest	0.539	0.793	0.642	0.978
SVM Poly	0.488	0.993	0.654	0.999





Conclusions

Penalized Logistic Regression Modeling Blue Tarp Classification

- Only incorrectly classifies 149 (~1%) of blue tarps.
- Only incorrectly classifies 12,995 (~1.3%) of other scenery.
- Model is robust and the correct choice for deployment.
- Simpler model than the other best model (Polynomial SVM), with less overfitting

Logreg Matrix

Prediction	Truth	
	Blue_Tarp	Other
Blue_Tarp -	14331	12995
Other -	149	997424



Citations

1. Anthropic. (2025). Claude Sonnet 4 [Large language model]. Used for reduction of run-time and graph compatibility errors.
2. RGB color picker. (n.d.). *RGB color picker*. RGB Color Picker. <https://rgbcolorpicker.com/>