The Clouds and the Trees: Using Random Forest to Predict Cloud Coverage

Christine Giang (3032409225) & Jessica Yu (26027217)

Contributions: Team collaborated on implementing the two methods of splitting the data, ROC analysis, and producing the written report and README file. Christine produced pairwise relationships, plots for feature selection, cross validation/test accuracy tables, diagnostic and misclassification plots, and model stability analyses. Jessica examined relationship between expert labels and features, model fit and improvement, and the impacts of PCA transformed LDA, QDA, logistic regression, and random forest.

GITHUB LINK: https://github.com/christinegiang/stat154

SECTION 1

part a)

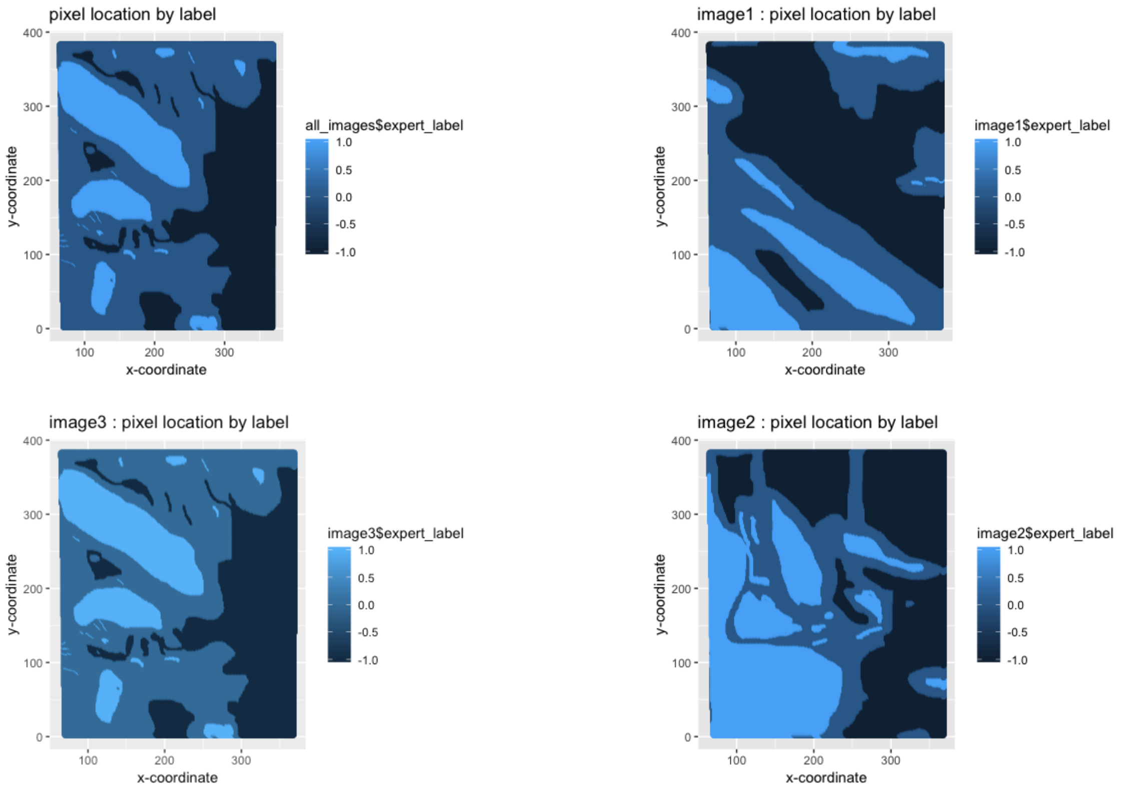
The paper, *Daytime Arctic Cloud Detection*, explores a study of the sensitivity of Earth's climate to increasing amounts of atmospheric carbon dioxide, and focuses on the impacts of cloud coverage in the Arctic. In 1999, NASA launched the Multiangle Imaging SpectroRadiometer (MISR), capturing electromagnetic radiation measurements. However, the amount of visible and infrared electromagnetic radiation emanating from clouds compared to snow- and ice-covered surfaces resemble each other, leading to issues with cloud detection in the Arctic. With the goal of efficiently combining classification and clustering methods, statisticians collaborated with scientists at the Jet Propulsion Laboratory and used an approach focused on searching for cloud-free instead of cloudy ice/snow-covered surface image pixels. The algorithms used in the efforts were based on three physical features. CORR measures the correlation of MISR images of the same scene from different viewing directions and indicates the scattering properties of ice/snow-covered surfaces. SD measures the standard deviation of MISR nadir camera pixel values across a scene. NDAI measures the normalized difference angular index that characterizes the changes in a scene produced by changes in the MISR view direction. An enhanced linear correlation matching (ELCM) algorithm was based off these features. Results were used to train QDA to provide probability labels for partly cloudy scenes. Overall, this research improved scientists' understanding of the flow of visible and infrared radiation through the atmosphere, so that they may study the response of clouds to changes in the arctic climate and their feedback on it. Ultimately, these studies will enable the scientific community to see how changing cloud properties may enhance any initial changes in the arctic brought about by increasing concentrations of atmospheric carbon dioxide.

part b)

There are three images, with 345,556 pixels in total. Experts labeled 23.43% as clouds, 36.78% as non-clouds, with the remaining 39.79% left unlabeled. A highly interesting pattern is that the unlabeled pixels seem to completely wrap around all the parts that are definitively labeled as clouds. This verifies that the cloud labels are correct because the edges of the clouds fade out completely. This is a fact that can be useful in the analysis of our classifiers.

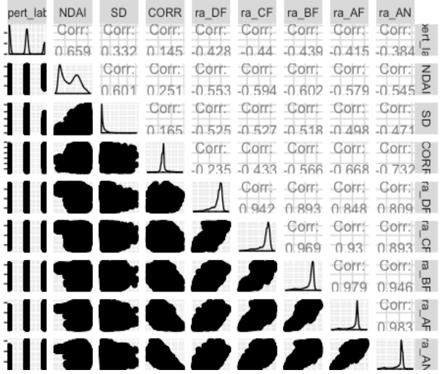
Another trend is that pixels assigned to a certain expert-label tend to be clustered with other pixels with that same expert-label. The expert labels do not appear in random, isolated locations. We observe this by visually examining the three images, where we see that pixels assigned a given expert label are grouped together as indicated by color. For example, pixels classified as -1 tend to be clustered with other pixels classified as -1 in a dark blue mass on the image.

Given these observations, an i.i.d. assumption would not be appropriate here. In addition, from a scientific and intuitive perspective, cloud particles are not isolated points in space but instead cluster to form a whole cloud altogether that spreads across a certain amount of space.



part c)

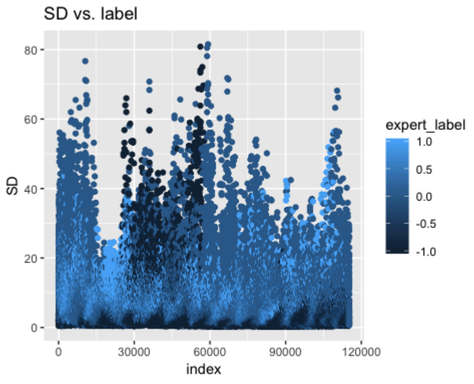
(i) Pairwise relationships: We see from the pairs plot that SD is highly correlated with NDAI (0.601), but not as highly as the correlations between the five radiance angles, which are ~0.85 up to almost 1. There is high collinearity between the radiance

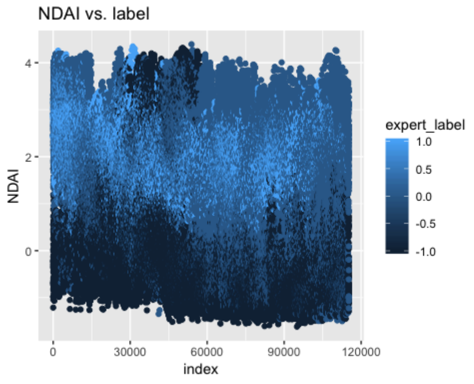


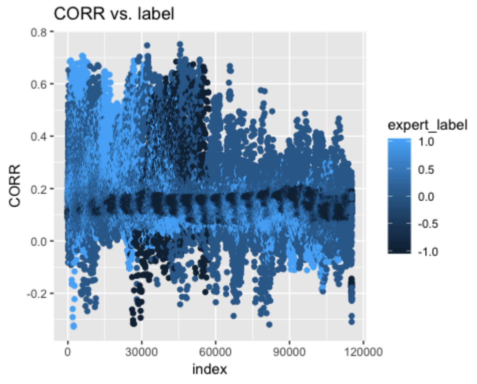
angles. CORR is weakly correlated with SD and NDAI, at 0.16 and 0.251 respectively.

	image1	image2	image3
expert_label	1.0000000	1.0000000	1.0000000
NDAI	0.6591295	0.6825384	0.49887924
SD	0.3324615	0.3509872	0.23597170
CORR	0.1448060	0.6922682	0.34274487
ra_DF	-0.4277770	0.2608798	0.14241384
ra_CF	-0.4399456	-0.2174091	0.02168176
ra_BF	-0.4387482	-0.4594755	-0.05722629
ra_AF	-0.4153346	-0.5258646	-0.12841936
ra_AN	-0.3838326	-0.5167218	-0.17290429

(ii) Relationship between expert labels and individual features: the correlation matrix above generates correlation values between the expert labels and NDAI, SD, CORR, and the radiance angles. We see that NDAI is highly correlated with the expert labels across all three images, particularly in Image 1 and Image 2. It appears that the usefulness of the five radiance angles varies across the three images. The radiance angles appear to be weakly correlated with the expert labels in Image 3, and may not be very useful in classification. In Image 2, it appears that ra_BF, ra_AF, and ra_AN have moderate correlation with the expert labels. However, for Image 1, it appears that a different set of radiance angles (ra_DF, ra_CF, and ra_BF) have the highest correlations with the expert labels out of the five radiance angles. The strength and consistency of the correlations of the radiance angles with the expert labels is not very promising.


SD vs. label


NDAI vs. label


CORR vs. label

The expert labels have been plotted against all of the variables. These three features (SD, CORR and NDAI) have been displayed because the distribution of the three classes look as if have some patterns and seem to be separable

SECTION 2

part a)

SPLIT METHOD 1

Our first method of splitting the data fairly considering that the observations are not i.i.d, is to use each of the three images for a set. For example, the first can be used as the training set, the second for validation, and the third for testing. This does not produce an ideal ratio in which the training set would comprise the majority of the total observations. Nonetheless, this is a relatively unbiased way of splitting the data.

SPLIT METHOD 2

Our second method of splitting the data involved separating the range of the points into grids and producing a coordinate map of grids. By randomly sampling from these grids, we assumed that the general location of a pixel on the coordinate map does not matter, but relative location to nearby pixels is important (non-i.i.d. assumption). We divided each image into 100 grids of 40 by 40 pixels. Then, the grids were randomly split into three groups: 80% for training, and 10% each for validation and test sets. A shortcoming of this method is that not all of the grids contain an equal amount of points. However, we believe that this method for splitting is still valid and useful because the points in a grid are mostly dependent on each other and less on points outside of the grid, which is more important than the slight unevenness in the quantity of points in each grid. Overall, the abundance of points allowed for training, validation, and test set proportions that came out to be very close to the intended 80/10/10% split, at 79.49%, 10.87%, and 9.64% respectively.

part b)

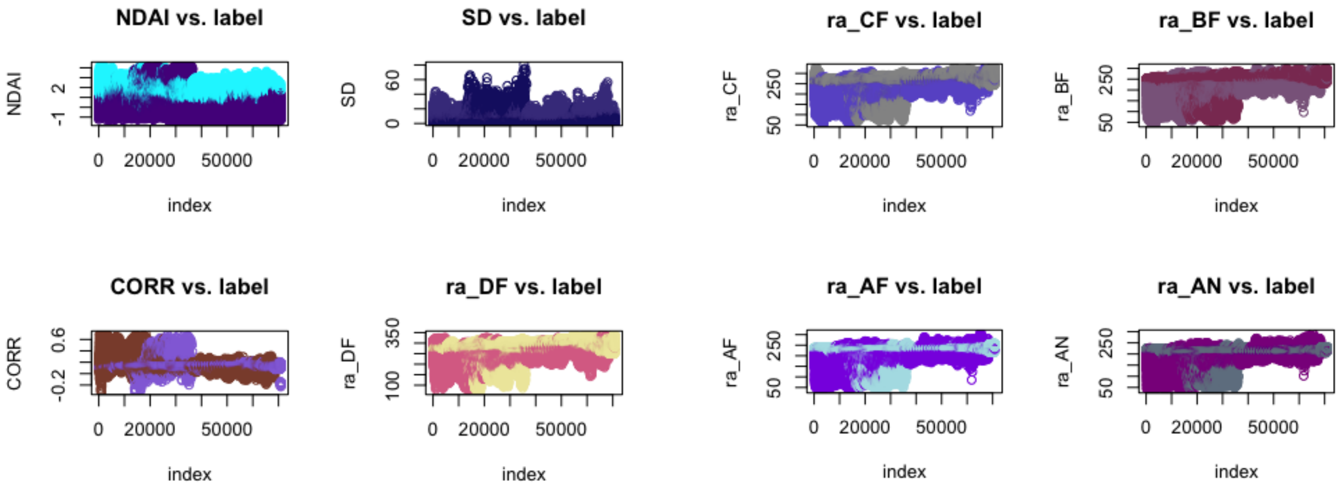
The trivial classifier that predicted all labels as -1 performed as expected. The prediction accuracies were equal to the proportion of labels in the validation and test sets that were actually equal to -1, so 41.88% and 25.02% respectively. This classifier would have a high average accuracy when the validation and test sets happen to contain more cloud-free data points. When using the grids method of splitting the data, we would obtain a higher test accuracy when the grids randomly selected for the validation and test sets happen to have lower cloud coverage. When we split the data by assigning one of the images as the test set, we would obtain a higher test accuracy when the selected image has fewer clouds compared to the other two images.

part c)

To select features for classification, we produced a correlation matrix in which the rows show the relationship of each variable to the expert labels and the columns represent each of the three images, the last one being all three combined as one large encompassing dataset. The zeroes have been removed before analysis because only the variables important for classification will be relevant and the classifiers we will implement are all binary, as predicting "unlabeled" is not very useful.

	image1	image2	image3	all_images
expert_label	1.0000000	1.0000000	1.0000000	1.0000000
NDAI	0.7697836	0.8193795	0.64446287	0.7584104
SD	0.4423335	0.4736005	0.36854162	0.4360264
CORR	0.1965444	0.7500088	0.52604971	0.5510043
ra_DF	-0.5845621	0.3090899	0.21192096	0.0107873
ra_CF	-0.5856738	-0.2534827	0.01070498	-0.2827573
ra_BF	-0.5761878	-0.5154060	-0.12935262	-0.4476648
ra_AF	-0.5485383	-0.5842979	-0.24533528	-0.5073210
ra_AN	-0.5089311	-0.5726166	-0.31502792	-0.5045998

The ggpairs plot from Section 1 shows that all the radiance angles are highly correlated, which means we shouldn't use more than one of them due to collinearity. Also the marginal plots below of each variable against the labels show that all angles seem to have the same shape. Since NDAI takes the average of the radiation measurements as stated in the paper, we felt it was best to use NDAI to be representative of all the radiance angles. Additionally, the correlation matrix shows that the correlation between NDAI is the highest across all three images individually and also when combined. The correlation of the CORR variable to the expert label is also relatively high across the images, as is SD. Therefore, our exploratory data analysis led us to select the same three variables as in the research paper.



part d)

Our **CVgeneric** function takes in a classifier, training features, training labels, a test set, number of folds K and a loss function as arguments. The function is compatible with the four classification methods we used in our research: logistic regression, LDA, QDA, and random forest. Additionally, there is a custom loss function written to report model accuracy under the name ***accuracy_loss_fcn***. **CVgeneric** returns two vectors: one of cross-validation accuracies and one of test accuracies across *k* folds.

SECTION 3

part a)

Before making any models, we removed all observations that were unlabeled. Then, all non-cloud values were changed from -1 to zero, making our classification a binary problem. We implemented four models: logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and random forest. The CV and test accuracies stayed very consistent across each fold, regardless of the model used. Across the models, CV and test accuracy scores were very similar. Random forest had the highest CV and test accuracies. Using CVgeneric to compare each of our models, we selected random forest as our best performing model.

TUNING PARAMETERS:

The default parameters for random forest were too slow to run, so we decided to tune them slightly. The nodesize defines the smallest size of a node in which the trees will stop splitting. Since the default size is 1, random forest would split into *N* nodes, meaning there could be as many nodes as there are observations. This node size splits too far,wouldn't necessarily provide useful information, and would likely lead to overfitting. The nodesize should be small enough so that the model can still pick out relevant small groups. The size we decided on was 19, which is 0.01% of the training size.

SPLIT METHOD 1

Cross Validation Accuracies

k	LOGISTIC REGRESSION	LINEAR DISCRIMINANT ANALYSIS	QUADRATIC DISCRIMINANT ANALYSIS	RANDOM FOREST
1	0.8366755	0.8702696	0.8610571	0.9027916
2	0.8401763	0.8678526	0.8589581	0.9040661
3	0.8417123	0.8706219	0.8600588	0.9074736
4	0.8394027	0.8679560	0.8608078	0.9037767
5	0.8369932	0.8703093	0.8583171	0.9039034

Test Accuracies

k	LOGISTIC REGRESSION	LINEAR DISCRIMINANT ANALYSIS	QUADRATIC DISCRIMINANT ANALYSIS	RANDOM FOREST
1	0.8987316	0.9361762	0.9500170	0.9366874
2	0.8983907	0.9358353	0.9496640	0.9374057
3	0.8976847	0.9355797	0.9499562	0.9362614
4	0.8996445	0.9362005	0.9504431	0.9347154
5	0.8971856	0.9356040	0.9496397	0.9373326

SPLIT METHOD 2 (GRIDS)

Cross Validation Accuracies

k	LOGISTIC REGRESSION	LINEAR DISCRIMINANT ANALYSIS	QUADRATIC DISCRIMINANT ANALYSIS	RANDOM FOREST
1	0.8741831	0.8970160	0.8958087	0.9173031
2	0.8737599	0.8968033	0.8952811	0.9203979
3	0.8702989	0.8929218	0.8907698	0.9168045
4	0.8730775	0.8964096	0.8952548	0.9174322
5	0.8726543	0.8954358	0.8934674	0.9177974

Test Accuracies

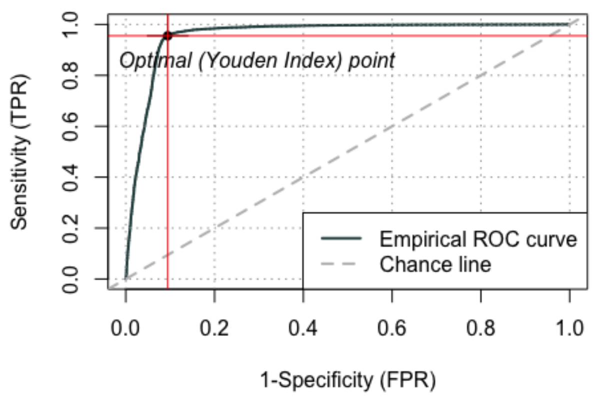
k	LOGISTIC REGRESSION	LINEAR DISCRIMINANT ANALYSIS	QUADRATIC DISCRIMINANT ANALYSIS	RANDOM FOREST
1	0.8510541	0.9049573	0.9022792	0.9463248
2	0.8514530	0.9050712	0.9027350	0.9475783
3	0.8515100	0.9052991	0.9029630	0.9467806
4	0.8512251	0.9046154	0.9026211	0.9487179
5	0.8513390	0.9051852	0.9025071	0.9463248

Random forest performed better on the first split method with a mean CV accuracy of 90.4%, although the test accuracy is not the highest of the four methods. However, the test accuracies are pretty close, with the average of 93.65% being almost identical to LDA (93.59%) and performing just slightly worse than QDA (94.99%). Additionally, our team believes that split method two is better because the training size for the first method is only one-third of the data, which we do not believe is large enough for training. We are confident that the grid split method is a better way of randomizing the data while also accounting for relational dependence. The second split method yielded the best results associated with random forest for cross validation and test accuracies, Random forest achieved test accuracy of 94.71%, yielding a roughly 5% improvement over LDA & QDA test accuracies and nearly 10% over logistic regression test accuracy.

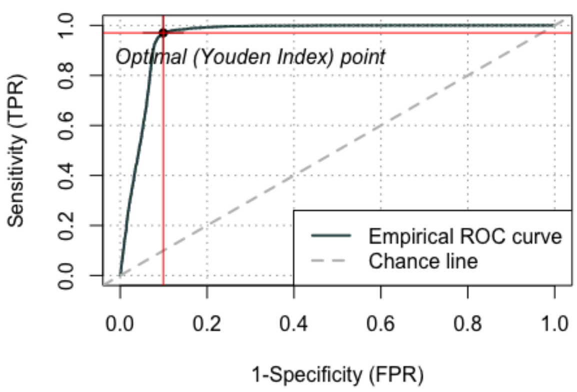
part b)

We used Youden's J statistic to determine the ROC cutoff level. The statistic measures the difference between the true positive (TP) and false positive (FP) rates. We picked the point that maximizes this point, because this value indicates the best trade-off between TP and FP rates. The cutoff values were accessed through the ROCit object produced for each model, and the optimal value was found by using the index of Youden's J statistic.

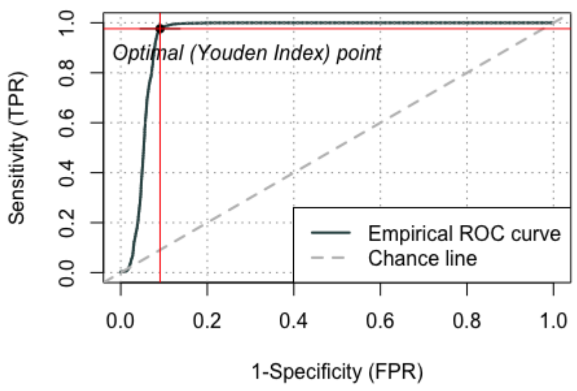
Logistic Regression



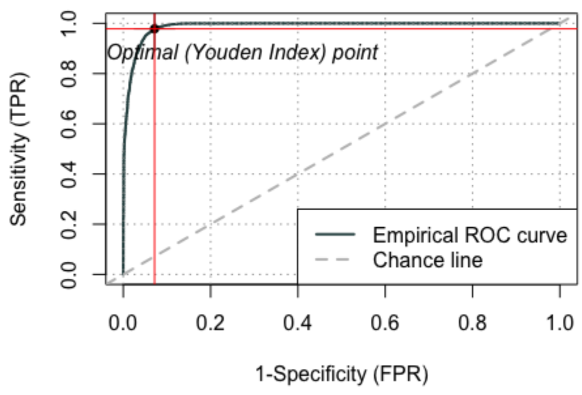
LDA



QDA



Random Forest



ROC CUTOFF VALUES

Logistic Regression	0.2381142
Linear Discriminant Analysis	0.2007537
Quadratic Discriminant Analysis	0.1081531
Random Forest	0.2833333

part c)

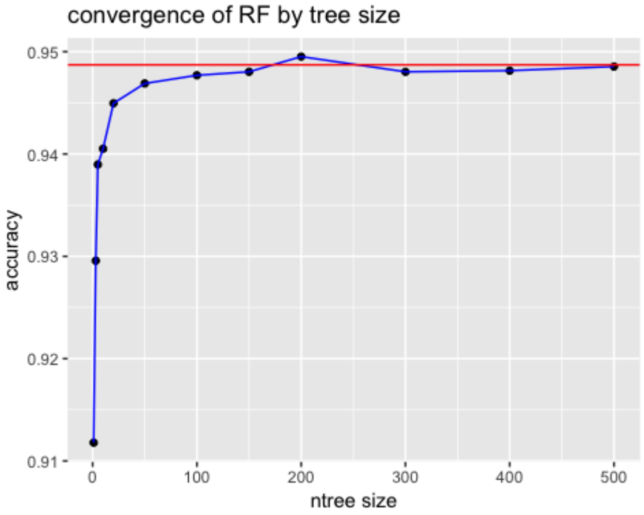
We evaluated the AUC scores, which indicate the area under the ROC curve. The AUC score is a single measure that helps measure model performance, with a higher AUC score indicating better fit. As seen below, each of the models has a high AUC score around 0.95, with random forest producing the highest AUC score at .9884.

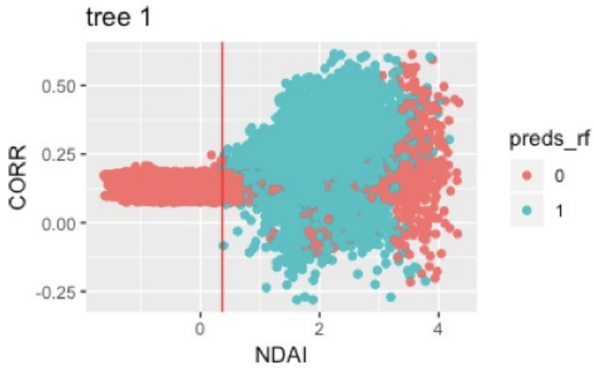
Next, we evaluated the confusion matrices for the number of true positives, true negatives, false negatives and false positives. Random forest produced the highest accuracy at 0.9722, and logistic regression performed the worst at 0.8876.

SECTION 4

part a)

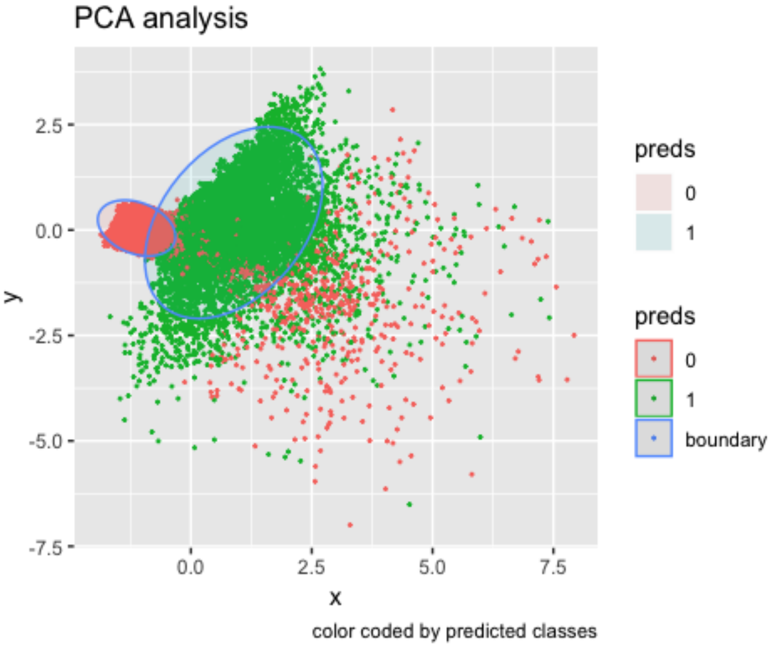
To the right is a plot of the accuracies changing as the ntree size (number of trees to grow) increases from 1 up until size 500. The plot shows that the accuracy converges pretty quickly and reaches our original test accuracy of 94.87% at around ntree size of 100.

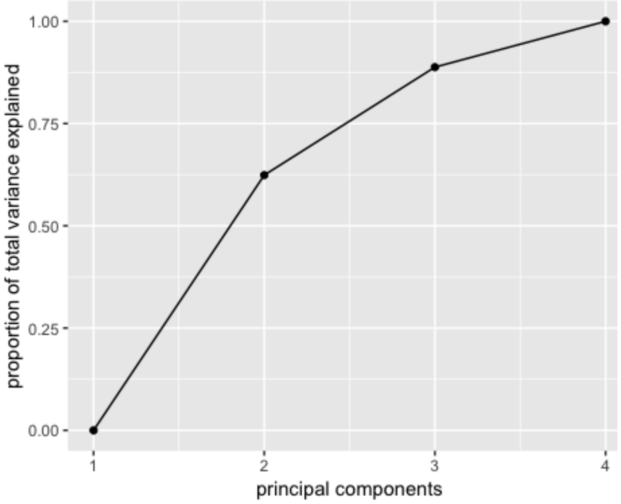




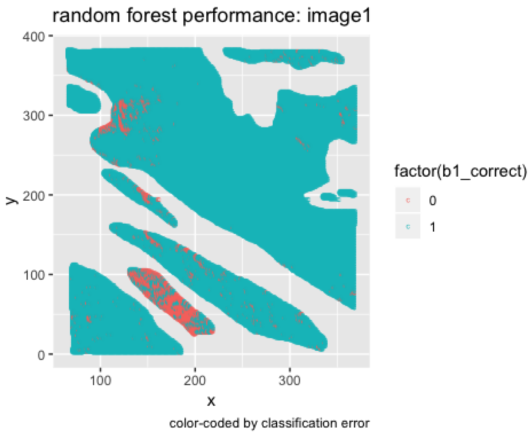
This graph displays the first tree in the random forest model, and shows that the first main split separates the two classes well upon visual inspection when we plot the two variables that the tree split on.

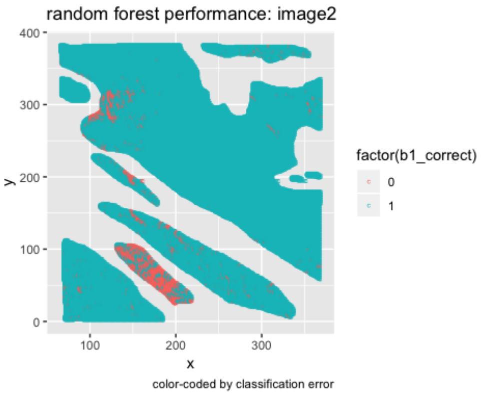
We also used principal component analysis (PCA) to look at a lower dimensional representation of the data in which we could compare our results. The first two principal components cover 88% of the variation, and is thus sufficient for representation. at the 85% threshold. They have been plotted against each other, overlaid with the predicted classes from out random forest model. The bulk of both of the classes are pretty well separated which indicates that our model did a good job distinguishing between the clouds and non-clouds.

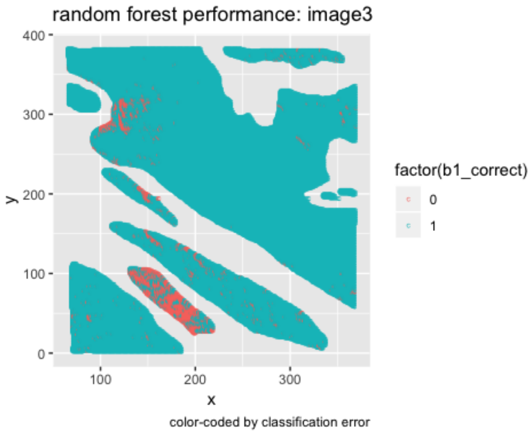




part b)







The bulk of the misclassified pixels seem to be in the same two areas in all three images: towards the upper left corner and in the bottom mid-left. Our assumption is that there is probably a lot of snow/ice cover in these locations, making it more challenging to distinguish clouds. We pulled out the misclassified observations to see their distribution within the expert labels and it seems that a much larger proportion of non-clouds (75.23%) are misclassified as clouds. This is further evidence that these areas might just look very similar to clouds.

part c)

Running LDA, QDA, logistic regression, and random forest on the three PCA-transformed features did not improve prediction accuracy very much when we included the first two PCs in our models. In fact, prediction accuracy slightly decreased. When we expanded our feature selection to include the five radiance angles in addition to NDAI, SD, and CORR for PCA transformation, we obtained similar results as when we included the first three PCs in our model. This indicates that the radiance angles are not very significant in classification, most likely because the information in the radiance angles are essentially already captured in NDAI. Without expert labels, we expect the random forest model to continue achieving high accuracy rates as it already consistently achieves high accuracy rates across all three images. It would be helpful to train the model across more than three images.

part d)

The training data was bootstrap sampled with replacement 5 times, split into two smaller sets, one of which was bootstrapped and the other permuted randomly to check the performance of the random forest model. The accuracy remained consistent which shows that the model was robust through all of these perturbations of the data, and ranged only between 94.28% and 94.91%.

Another way we modified our data split was by permuting the order in which we assigned images to the training/test/validation sets in our first splitting method. There are five additional ways to order them, aside from the original assignment. All these permutations also performed very well with accuracies ranging only between 95.17 and 95.21%.

part e)

CONCLUSION

The random forest model performs well on the selected original features as evidenced by its high classification accuracy and the stability of its performance even when the data is split differently or perturbed. We found that we can train a useful model using only the three features NDAI, SD, and CORR. Adding the radiance angles does not provide very much additional usefulness since NDAI already effectively summarizes the information provided by the five radiance angles. PCA transforming the raw features did not provide higher accuracies. Thus, using random forest on three original features can help produce a classification model that scientists may utilize in studying cloud coverage in the Arctic and its effects upon the concentration of atmospheric carbon dioxide.

Resources

- https://towardsdatascience.com/linear-discriminant-analysis-lda-101-using-r-6a97217a55a6
- https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5
- https://cran.r-project.org/web/packages/ROCit/ROCit.pdf