Lab 4 - Cloud Data, Stat 215A, Fall 2022

Sizhu Lu, Kaitlin Smith, Ghafar Yerima

November 13, 2022

1 Introduction

Global warming is characterized by the rapid increase in Earth's surface temperature over the past century. It is due primarily to human activities including fossil fuels burning, which increases greenhouse gases (Rexford 2020). The Earth's surface temperature has increased by nearly 1 degree Celsius between 1905 and 2005, and is predicted to increase between 2-6 degree Celsius by the end of the 21st century (Rexford 2020). Models have predicted that the strongest dependencies between greenhouse gases level and earth's surface temperature will occur at the Arctic (Shi et al. 2008). As the Arctic warms faster than the rest of the globe, alterations in the properties and distribution of ice-and snow-covered surfaces, atmospheric vapor, and clouds can lead to further warming, and more susceptibility to greenhouse gases levels (Shi et al. 2008). To understand these intricate relationships, a study of the cloud coverage at the Arctic is necessary as clouds play a crucial role in tuning the sensitivity of the Arctic to the rise of surface air temperatures (Shi et al. 2008).

However, studying Arctic clouds is challenging because liquid and ice water cloud particles have similar physical properties to ice and snow surfaces particles. Thus, distinguishing Arctic clouds from water or ice using their electromagnetic radiation is challenging. Thanks to a National Aeronautics, and Space Administration (NASA) effort, a Multiangle Imaging SpectroRadiometer (MISR) was launched on the Terra satellite in 1999. This radiometer produces novel radiation measurements from nine view angles. The nine view zenith angles of the cameras are 70.5" (Df), 60.0" (Cf), 45.6" (Bf), and 26.1" (Af) in the forward direction; 0.0" (An) in the nadir direction and 26.1" (Aa), 45.6" (Ba), 60.0" (Ca), and 70.5" (Da) in the aft direction. Thus, the MISR produces a large amount of data thanks to its 360 km-wide swath coverage on the Earth's surface that extends from the Arctic down to Antarctica in approximately 45 minutes.

Although the MISR produces a high resolution data, its operational cloud detection system was designed before its launch and proved to not be effective for detecting clouds over bright surfaces in polar regions. Solving the polar cloud detection problem requires efficient statistical models, capable of handling the massive data set while incorporating scientific and operational considerations.

In this paper, we explored a MISR data of polar regions and implemented various polar cloud detection algorithms based on expert labels. The algorithms implemented include logistic regression (with and without regularization), KNN, RF, QDA, and SVM. Ultimately, we aimed to design algorithms that will perform well on data sets lacking any expert labels.

2 Data

Our data comprises 3 images which represent pictures taken from the MISR. Each image text file comprises 11 columns. The first two columns represent the y and x coordinates of the pixel, and the third column represent the expert label (+1= cloud, -1= not cloud, 0= unlabeled). The fourth, fifth, and sixth columns were features engineered by Shi et. al. They are: the Normalized Difference Angular Index (NDAI) which characterizes the variations in a scene with the variations in the MISR view direction, the standard deviation

(SD) of MISR nadir camera pixel values across a scene, and the correlation (CORR) of MISR images of the same scene from different MISR viewing directions (Shi et al. 2008). Columns 7 to 11 represent the radiance angles DF, CF, BF, AF, and AN from the MISR.

2.1 Data Collection

The data used in our study is a subset of the MISR data set from Shi et. al. It was collected from 10 MISR orbits of path 26 over the Arctic, Northern Greenland, and Baffin Bay (Shi et al. 2008). As the MISR collects data from 233 geographically distinct but overlapping paths on a repeat cycle over 16 days, the 10 orbits spanned approximately 144 days from April 28 to September 19, 2002. Path 26 was chosen for the richness of its surface features and their relative stability over the study period. From the 6 data units per orbits included in their study, 3 were excluded as they were open water after the ice melted over the summer.

2.2 Data Exploration

2.2.1 1. Expert labels

First, we'll look at the the expert labels for the presece of clouds:

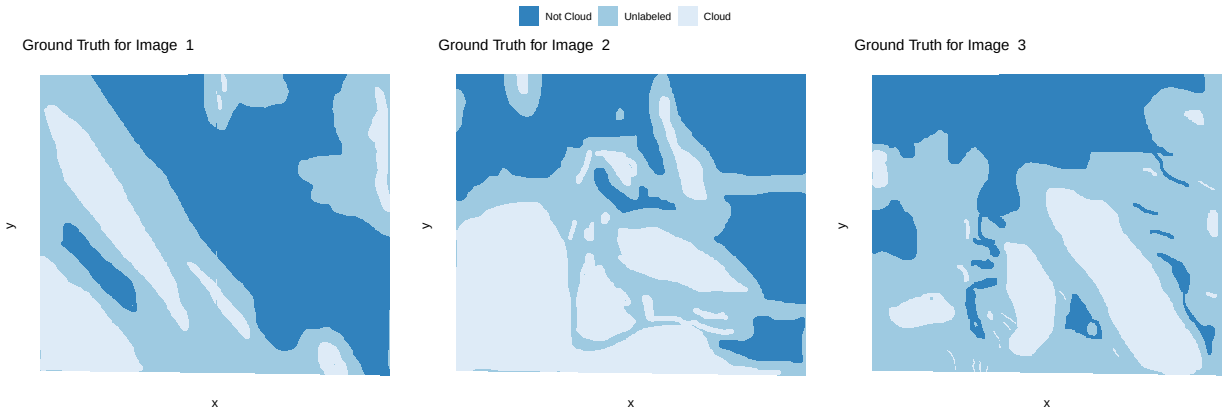

Figure 1: Expert Cloud Labels

Looking at the initial expert labels for image 1 and 3, we notice that the clouds occupy roughly the third of the map. However, the proportion of labeled clouds on image 2 is substantially larger. The different radiance angles seem to be correlated for the most part. They show approximately the same distribution and intensity of pixels.

2.2.2 Rational for Converting Unlabeled pixels to "Cloud" Labels

Next, we explored the distribution of the cloud labels. It appeared that the cloud labels is contained in the labels that were unlabeled. For this reason, we decided to consider the unlabeled pixels as cloudy pixels. Following this conversion, we notice that our images now contain more cloudy pixels than non-cloudy ones.

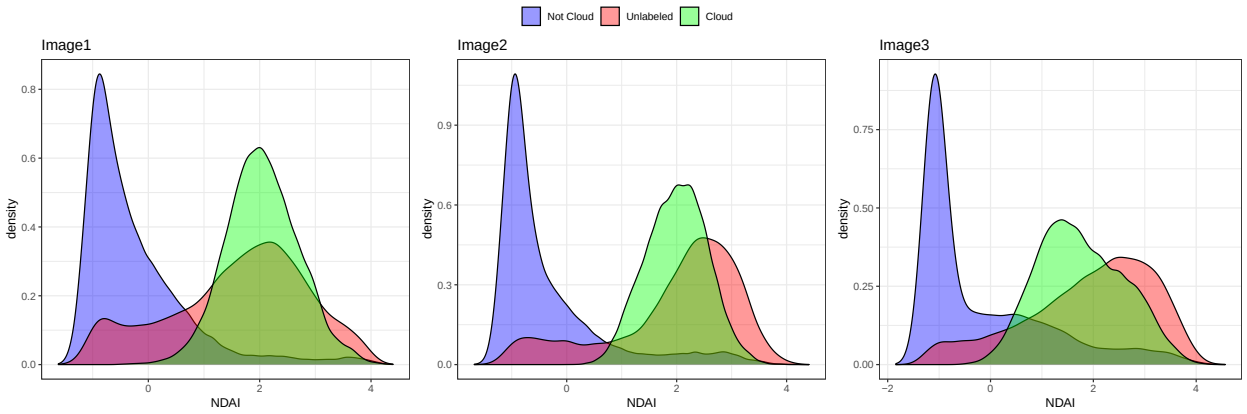

Figure 2: Plotting distributions of cloud expert labels vs NDAI

2.2.3 2.

First, we'll look at relationships between the radiances of different angles visually.

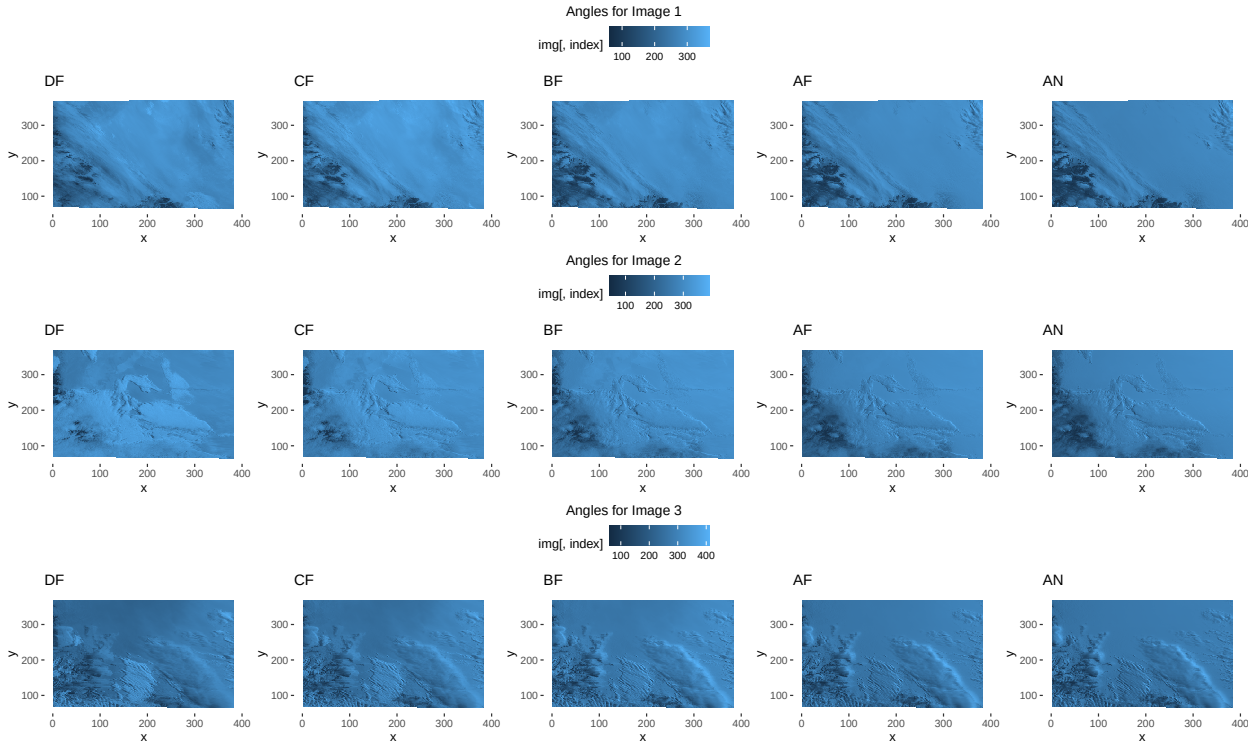

Figure 3: Plotting angles on maps

Visually, it appears like the DF and AN bands are inverses of one another. Particularly for Image 2, When the DF band registers higher values, the AN band registers lower quantities. Comparing these plots to the expert labels, it appears like clouds correspond to high values of DF and CF, which is particularly apparent in image 2.

Now, looking at the CORR, NDAI, and SD features, we see stronger relationships between the presence of clouds and these features:

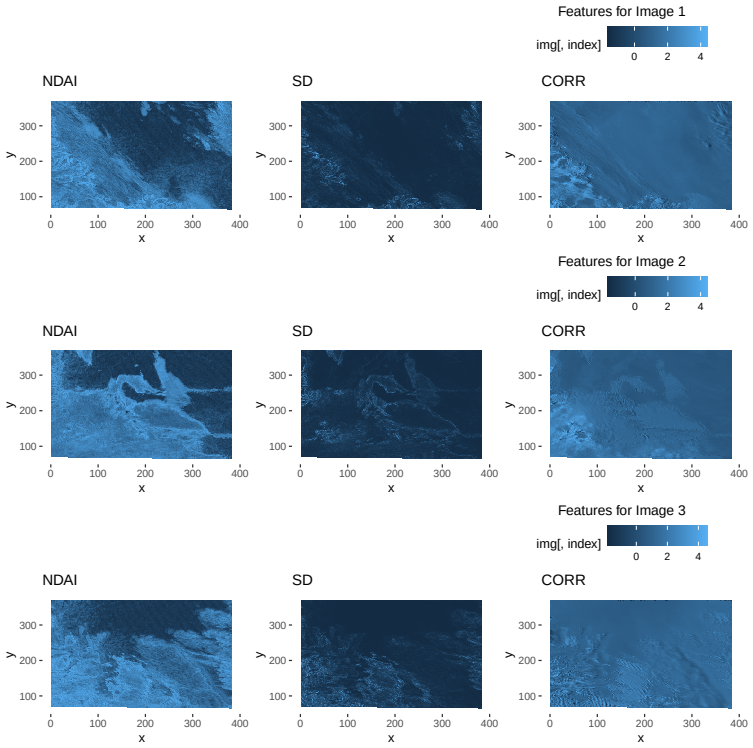

Figure 4: Plotting features on maps

It appears as though high values of NDAI indicate the presence of clouds. However, the highest values of NDAI correspond to the regions that the expert failed to label. More subtlety, you can see that higher values of CORR also indicate the presence of clouds.

Our next step was exploring the features quantitatively. Thus, we investigated the correlation of our features to the labels for each image. As shown below, it appears that NDAI has the strongest correlation with the labels. For image1 and image2, it appears that the angles of radiance also have a negative correlation. However, image3 does not have a meaningful negative correlation between the angles of radiance and the labels. However, the engineered features, NDAI, SD, and CORR, have a positive correlation with the labels between all of the images.

Figure 5: Plotting two-way correlation matrices of the features for images 1,2 and 3.

3 Feature Engineering

Although many features were engineered by combining and transforming the above features in various combinations, we'll only present the feature that outperformed CORR, NDAI, and SD.

3.1 Average NDAI

From the correlation matrix, it appeared that NDAI was the feature that was most correlated with the labels. Intuitively, the spacial aspect of this data is important. We expect "cloud" pixels and "not cloud" pixels to be clustered together. Or, if a label is "cloud", we would expect the other pixels around it to also be labeled as "cloud". Thus, we constructed a feature whose values for a given pixel is the average of the NDAI values for the 16 pixels around the pixel. If a pixel is on the edge of the image, the minimum value of $(x, mod(x+4, max(xval)))$ is taken. Where x is the index for a given pixel and $xval$ is the number of pixels in that row or column. Since the data is not square, $xval$ changes from row to row and column to column.

4 Modeling

4.1 Selecting The Three Best Features

After engineering the specified feature, we now re-examine the relationship between the expert labels and these new features. To reduce runtime and file size, 5000 points from the entire dataset were sampled.

Figure 6: Pair Plot to Select Best Features

After comparing these results with the results presented above, we conclude that Average NDAI is the feature that is most correlated with the expert labels. While both NDAI and Average NDAI separate the pixels into cloud and not cloud distributions, the Average NDAI feature further provides a greater degree of separation. The CORR feature also separates the 2 classes into 2 distributions, but to a lesser degree than Average NDAI and NDAI. Hence, we conclude that Average NDAI, NDAI, and CORR are the three best features.

4.1.1 2. Classifiers

Logistic regression model: Logistic regression is a standard and canonical choice of classification model when the outcome is binary. In the logistic model, we assume the data points are independently and identically distributed from a model where the probability of outcome being 1 is equal to a non-linear (sigmoid) transform of some linear combinations of the covariates. Both the generalized linearity and the IID assumptions are very strong. Specifically, the independence assumption is likely to be violated in our dataset due to spatial correlation. We need to be careful when interpreting the regression coefficients. However, in this lab, our main goal is the prediction problem instead of estimating the coefficients in the specified logistic model, we focus less on the model assumption verification and evaluate the models mainly from the perspective of prediction performance. We also fit logistic regression with l1 regularization penalty to avoid over-fitting.

KNN: KNN is a classification model where we predict the binary outcome of a single point by referring to the values of its k nearest neighbors, where k is a hyper-parameter we need to tune. This non-parametric model does not make assumptions about independence between data points or the specific probability distributions of the data. Therefore, we are able to use it even though we observe high spatial correlations in our cloud data.

Random forest: Random forest is an ensemble machine learning model for classification and prediction. We construct various decision trees from the bootstrapped subsample of training data and summarize the results to a predicted value of the outcome from the random forest. For each decision tree, we can train it on only a subset of all covariates to determine splits that best separate the two classes. Random forest is also a non-parametric model that makes minimal assumptions on the distribution of the data points.

Quadratic Discriminant Analysis (QDA): QDA is a version of Bayesian classification model which finds the linear separation in the second-order polynomial basis expansion of the covariates. It captures the lower-order linear interactions between covariates that separate the positive vs negative outcomes. The main model assumption QDA makes is that the covariates from the two classes have normal probability distributions with (possibly) different covariance matrices. As shown in our ggpair plot, not all covariates are obviously normally distributed thus it is likely the assumption is not perfectly satisfied in our data.

Support Vector Machines (SVM): SVM efficiently performs a non-linear classification by implicitly mapping the inputs into high-dimensional covariate spaces and then constructing a hyperplane separating the two classes of data points. It also makes the independence assumption between observations, which is likely to be violated in our dataset. Moreover, it assumes that in the true underlying data generating process, the two classes are linearly separable, which is another strong assumption in our dataset.

4.1.2 3. Fitting Models

Before the models were fit to the data, we separated each image into a grid of 4 sub regions. We then randomly sampled 4 of these sub regions to use as a test test. The remaining 8 regions were then randomly split into training and validation sets with 4 sub regions each.

After experimenting with the inputs to the logistic regression model, we found that computing the logistic regression with all of the features and angles yielded the best results. The results from this model are presented here.

Due to computability constraints, we ran SVM on a random subsample of 50,000 datapoints from the training data, but the model was evalutated on the full test set. We used the default hyperparameters for SVM as we did not have the computational power to utilize cross validation to fine-tune the parameters.

We fine-tuned the hyperparameters for the KNN and Random Forest models using cross validation. For each proposed value of the hyperparameter, the given model was fit to the training set before the model was evaluated on the validation set. The hyperparameters that resulted in the model with the highest accuracy were chosen.

In the plot below, you can see that accuracy for KNN was maximized at $k = 50$ nearest neighbors.

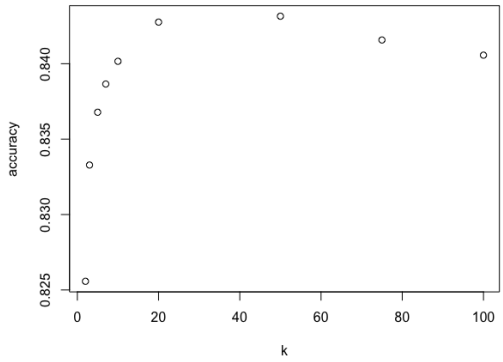

Figure 7: Hyperparameter tuning for KNN

For the random forest classifier, we chose to fine tune the hyperparameter that controls the number of variables that are randomly sampled as candidates in each split. From the plot below, we can see that a value of 2 maximized the accuracy of the random forest classifier.

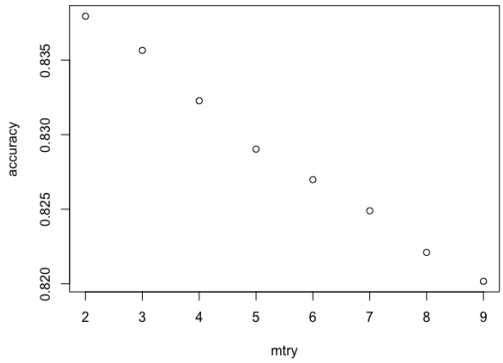

Figure 8: Hyperparameter tuning for Random Forest

Finally, we used cross-fit to tune the hyper parameter, λ of the regularized logistic regression.

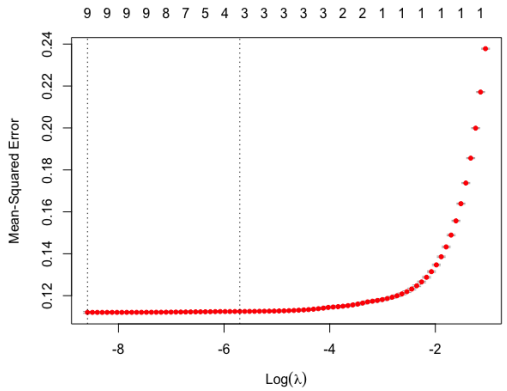

Figure 9: Hyperparameter tuning for Logistic Regression

As you can see, the mean squared error is minimized when $\lambda = 0.00019$. Thus, we conclude that regularization does not benefit this model.

The results for the models on the test set are presented below.

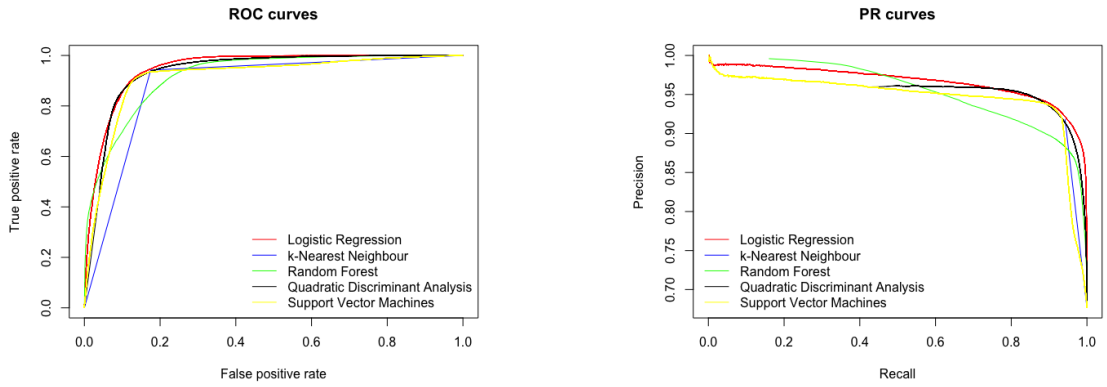

Figure 10: ROC and PR Curves for Models

The plots of the ROC and PR curves demonstrate that Logistic Regression performs the best of all of the classifiers implemented. However, as you can see in the table below, the results of the models are all comparable, with only minor differences in accuracy between them.

method	accuracy	precision	recall	fpr	fnr	f1_score
logit	0.9087902	0.9666281	0.08139827	0.08139827	0.09497597	0.1501524
knn	0.9040900	0.9418500	0.12848209	0.12848209	0.08153408	0.2261183
rf	0.8781607	0.9644667	0.09628726	0.09628726	0.13034198	0.1750941
qda	0.8996945	0.9328316	0.14471824	0.14471824	0.07997513	0.2505643
svm	0.8949508	0.9365625	0.14107919	0.14107919	0.08929532	0.2452197

Looking at these metrics, we can see that the models have high accuracy with balanced proportions of false positives and false negatives. Logistic regression maintains the highest accuracy and precision, but the KNN

model has similar accuracy and higher recall. Since in this application we are more concerned with false positives, or ground being labeled as clouds, logistic regression is the preferred model.

4.1.3 4. Random Forest Convergence and Feature Importance

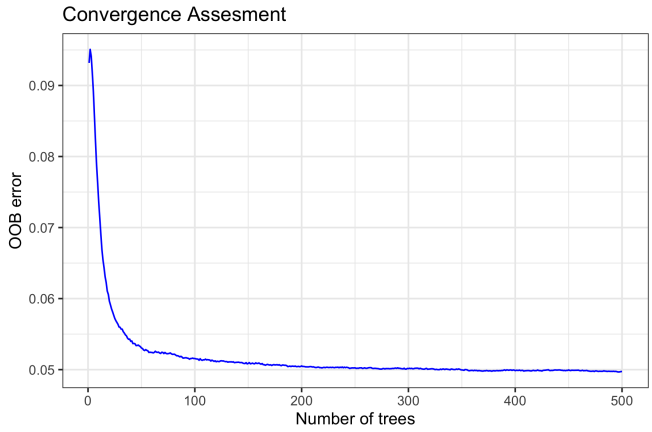

Figure 11: Convergence of the Random Forest model

Here, we are assessing the convergence of our Random Forest model using the Out of Bag (OOB) error. As shown on the figure, the OOB error decreases as more trees are added to the model. Convergence seems to be achieved after adding approximately 250 trees.

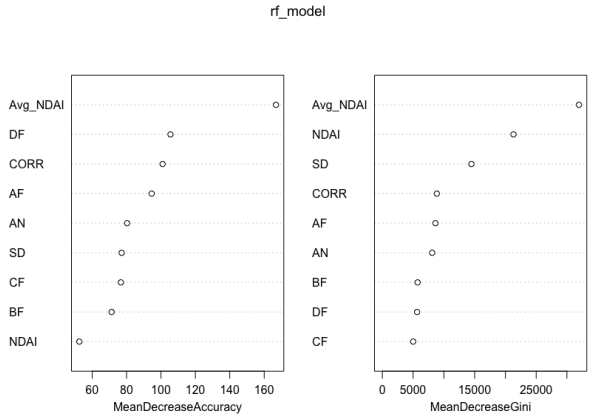

Figure 12: Ploting the Feature Importance of the Random Forest model

Next, we show the feature importance of the Random Forest Model through the Mean Decrease Accuracy and Mean Decrease Gini. The Mean Decrease Accuracy shows how much accuracy the model loses by removing a feature and the Mean Decrease Gini expresses the average gain of homogeneity by splits of a given variable. The higher these metrics for each feature, the more important the feature is. As shown on the figure, the most important feature seems to be the Average NDAI.

4.1.4 5. Post-Hoc EDA

Here, we present visual and quantitative analysis of the misclassified data for the logistic regression model.

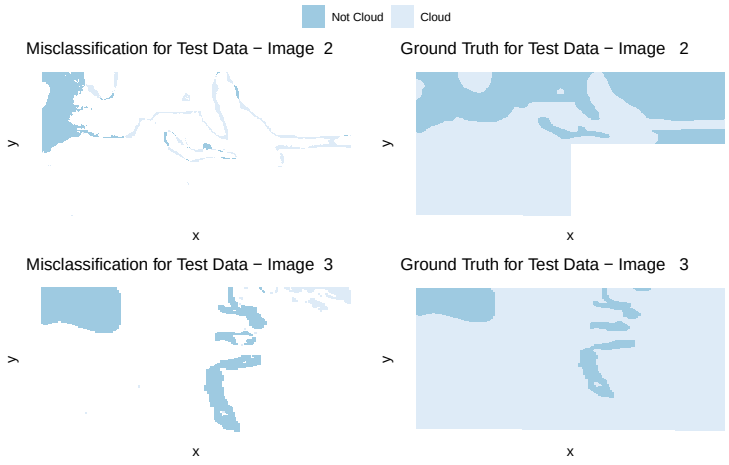

Figure 13: Comparing missclassified test data to ground truth

From the plot, and especially for image 2, we notice that the model has trouble classifying pixels that are on the borders between "cloud" and "not cloud" regions. Most of the misclassified pixels seem to be "not cloud" regions. When comparing these plots to the expert labels, we notice that the misclassified labels for "not cloud" regions are next to unlabeled regions.

We also investigated the distribution of each feature for misclassified and correctly classified data.

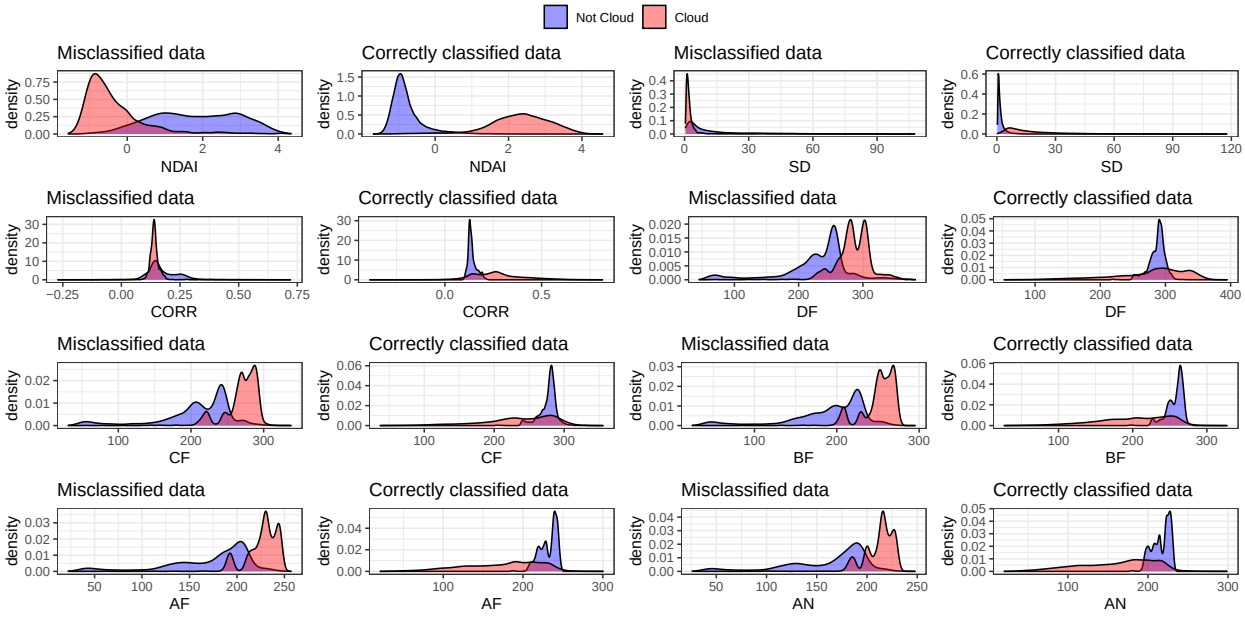

Figure 14: Plotting distributions of test data for all features

By looking at the plot, we notice that the feature range and distribution for the misclassified data seem to be unusual; They seem to be opposite to the general trend of the correctly classified data.

Next, we investigate the predictions quantitatively.

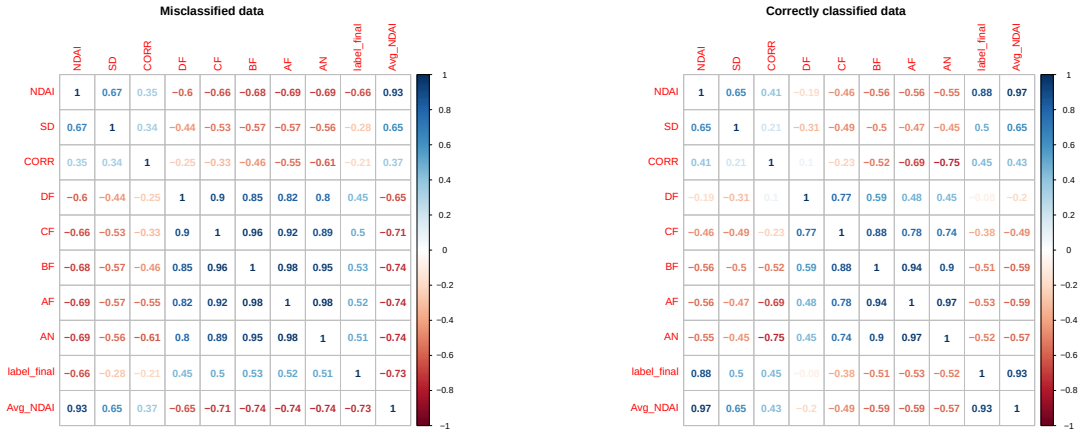

Figure 15: Plotting two-way correlation matrices of the features of predictions.

The correlation metrics show that the final label for the misclassiied data are negatively correlated with NDAI, SD, CORR, Avg_NDAI, while they are positively correlated to the same features for the correctly classified data. Moreover, the final label is positively correlated to the radiance angles, while they are negatively correlated for the correctly classified data.

4.1.5 6. Implication of Results in an Unsupervised Setting

The efficacy of these models on future data depends on the similarity of the future data to the images that these models were trained on. Ultimately, this comes down to the comparability principle. If supposed future images are comparable to this dataset, then we believe that the model would work well. However, if MISR data is taken on a different instrument, or even taken with the the same satellite that has been calibrated differently, these models would not provide adequate results. In addition, it is important to note that these images are from one path of the 233 geographically distinct MISR swaths. Although we could hypothesize that the data from path 26 would be comparable to the 232 other paths, more analysis would be needed to support this claim.

5 Academic honesty statement

We certify that the work above is our own, and collaborators and references have been cited as appropriate.

6 Collaborators

We'd like to thank our GSI, Theo Saarinen, for his support and direction on feature engineering, specifically for mentioning to consider the spatial aspect of the NDAI feature.

7 Bibliography

Ahima S. Rexford (2020) Global warming threatens human thermoregulation and survival, J Clin Invest, 130(2):559-561, DOI: 10.1172/JCI135006.

Tao Shi, Bin Yu, Eugene E Clothiaux & Amy J Braverman (2008) Daytime Arctic Cloud Detection Based on Multi-Angle Satellite Data With Case Studies, Journal of the American Statistical Association, 103:482, 584-593, DOI: 10.1198/016214507000001283.