Lab 4: Cloud Detection

Aniket Kesari, UC Berkeley Law 25948127
Mingjia Chen, UC Berkeley Stats 3032130297
Xiaoqi Zhang, UC Berkeley Stats 3032129569

October 31, 2017

1 Introduction

This lab examines the data from "Daytime Arctic Cloud Detection Based on Multi-Angle Satellite Data With Case Studies" by Shi et. al. In this study, the authors are concerned with developing an algorithmic approach for detecting clouds in images taken of the Arctic region. The fundamental problem is that at high latitudes, clouds are hard to distinguish from surrounding snow and ice in satellite imagery because the white surfaces are similarly reflective.

In this lab, we explore the data to evaluate the authors' choice of model features, and develop our own predictive algoirthms. Fundamentally, this is a "classification" problem, which is an example of a "supervised learning" model. This approach differs from the unsupervised clustering through distance similarities because we are interested in using known covariates to predict out-of-sample outcomes, rather than look for for patterns in an underlying data structure.

2 Exploratory Data Analysis

The first step in this sort of statistical analysis is an exploration of the data. In this stage, we are interested in replicating the pre-labeled plots of the three images in the paper, and exploring the relationships between the features and the "angular radiances" recorded by the satellite camera.

2.1 Map the Presence or Absence of Clouds

First, we recreate the plots by using the XY-coordinate pairs provided in the data, and shading each point with its expert label. Curiously, only Images 1 and 2 are reproducible compared to the original paper. "Image 3" does not look similar to the one in the paper.

2.2 Differences Between the Two Classes

Before we go in depth to discuss the relationship between the radiances of angles, first, we explore the relationship between the expert labels against the Angle Radiances with the boxplot Figure 2. It is easy to notice that the "clear" points are mostly concentrated around 200 and 300 radiances, while "cloud" labeled points are more concentrated around 100-250 radiances.

From the below figure we could also see that:

- `clear` labeled points are highly right skewed
- `clear` labeled points generally have higher value than `cloud` labeled points
- For angle DF, the `clear` and `cloud` are very close to each other. This indictes that angle DF might not be an ideal feature to classify the two labels. We will further cover this part in the feature selection session.

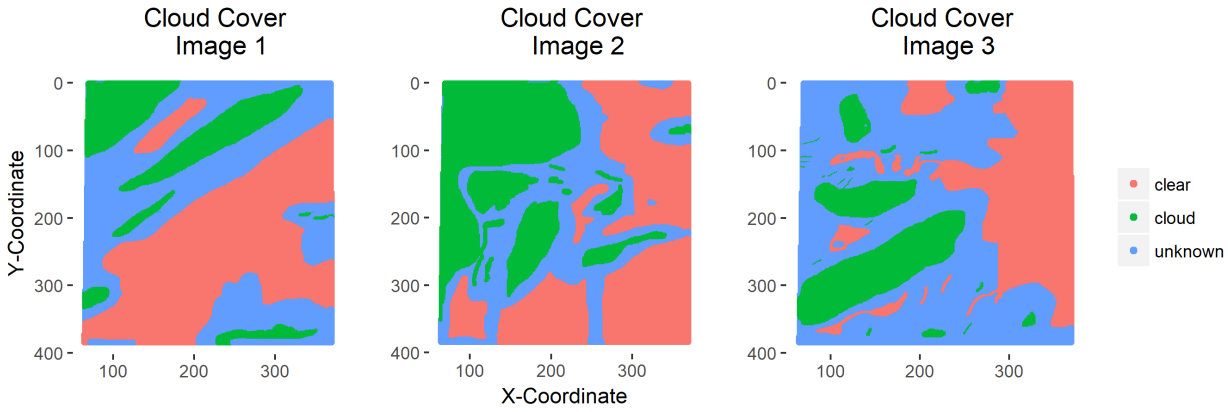


Figure 1: Mapping Clouds with original labels

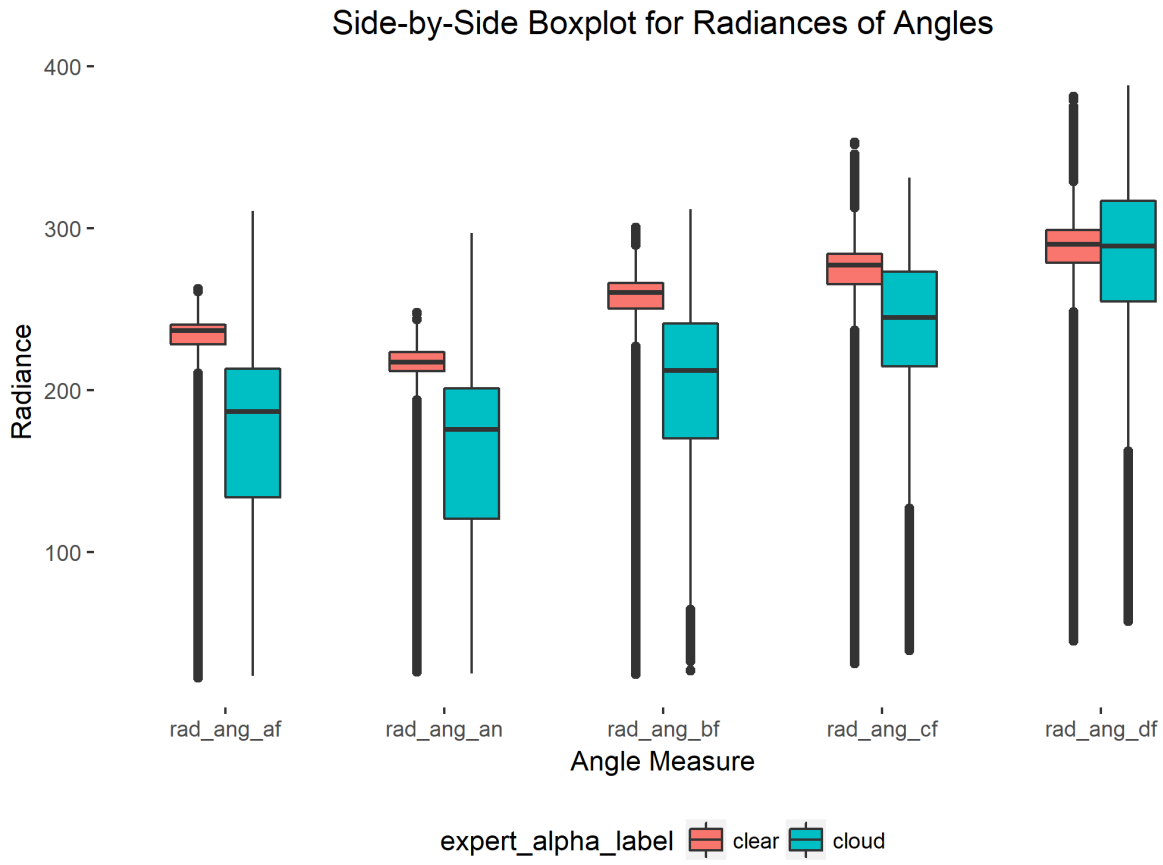


Figure 2: Boxplots for Radiances of Angles

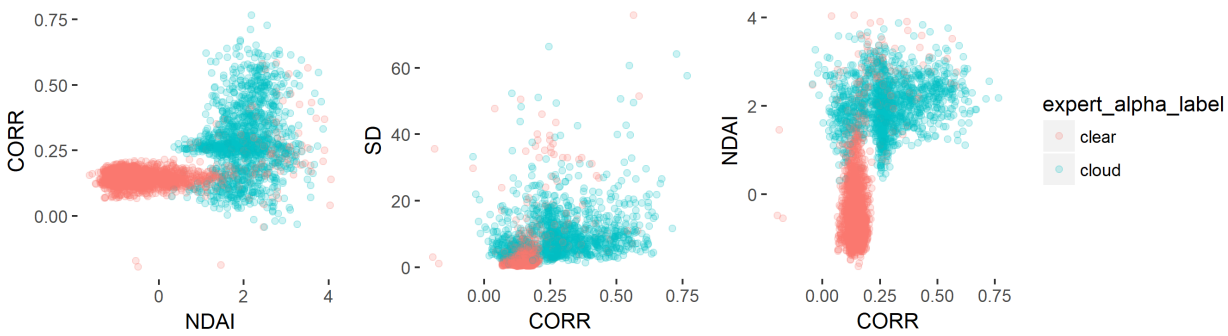
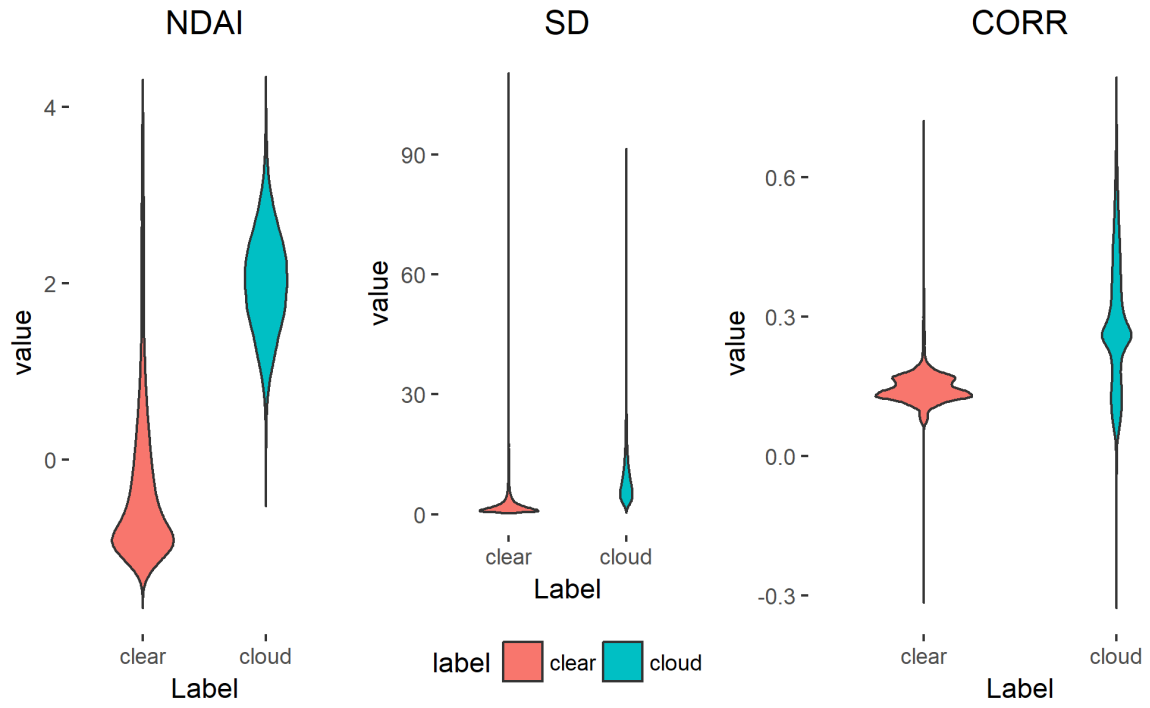


Figure 3: Pairwise scatter plot for features

Next, we will check the label differences based on features. The figure below is a side-by-side violin plot indicating the density curve of each label, distributed against the features. From this figure we see that labels are separated well in NDAI and SD.

- For CORR, although `clear` labels are concentrating well, it is difficult to differentiate the two labels against CORR value.
- For SD, `clear` label is highly skewed and concentrated against the SD feature.
- For NDAI, the two labels concentrated on different values and are easy to separate based on NDAI.



We also create a pairwise scatter plot to check the 2d distributions of labels. All three measures are fairly good at separating out the points from one another. The NDAI measure does the best job of separating the cloudy and clear points from each other, with clusters forming around high NDAI scores with low radiance angles, and around low NDAI scores with high radiance angles.

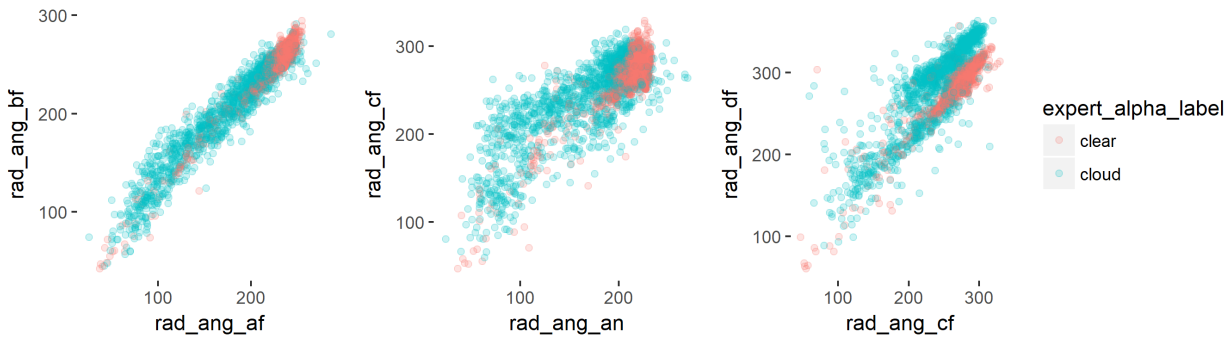


Figure 4: Exploration: scatter plot for radiances of angles

2.3 Explore the Relationship between the Radiances of Different Angles

Moving to exploring the relationship between the radiances of different angles, Figure 2 highlights that both labels are concentrated on different radiance values, with angle DF on the highest radiances and angle AN on the lowest radiances. This aligns with the information provided in the paper, since DF has the smallest angle and AN has the biggest (90 degree).

We plot several pairwise scatterplots on radiances of different angles. We could see that `clear` label also concentrates, and there is some evidence for collinearity.

To gain more insight into collinearity, both visually and quantitatively, we construct a correlation matrix, and create a pairwise scatter plot.. The correlation matrix would help us identify highly correlated columns in our data and thus help in feature selection. Figure 5 is a both quantified and visualized correlation matrix. Using a cutoff correlation of .75, we find that the "af," "an," "bf," and "cf" measures are highly redundant with one another, while DF has quite high correlation with CF and BF but less correlation with AF and AN.

3 Modeling

3.1 Best Predictors

To select the best features, we first create a "correlation matrix" that reports which features are highly correlated, and therefore probably redundant. This correlation matrix is also visualized in Figure 5. As mentinoned in the previous section, we used a cutoff correlation of .75, and found that the "af," "an," "bf," and "cf" measures are highly redundant with one another. So, we will only keep one feature among these four features for consideration before proceeding. Among these four, angle AN has the lowest correlation with the other features which we would like to keep, so we would keep angle AN, together with NDAI, Standard Deviation (SD), Correlation (CORR), and the Radiance Angle DF (rad_ang_df) measure for model building.

Next, we used caret implementations to determine what the most important features are for prediction. For feature selection, we employ the "Cross-Validation Learning Vector Quantization" method. This method runs a repeated cross validation, and returns scores for each feature that range between 0 (not important) and 1 (important).

After running this analysis, we find that of the four features we identified, "Radiance Angle DF" was the least important, with "Radiance Angle AN" being the fourth most important feature. Unsurprisingly, NDAI was the most important feature. This plot suggests that the three best features are "NDAI," "SD," and "CORR". Figure 4 also lends support to the idea that cloudy and clear points can be split quite well when mapped onto NDAI, CORR, and SD measures.

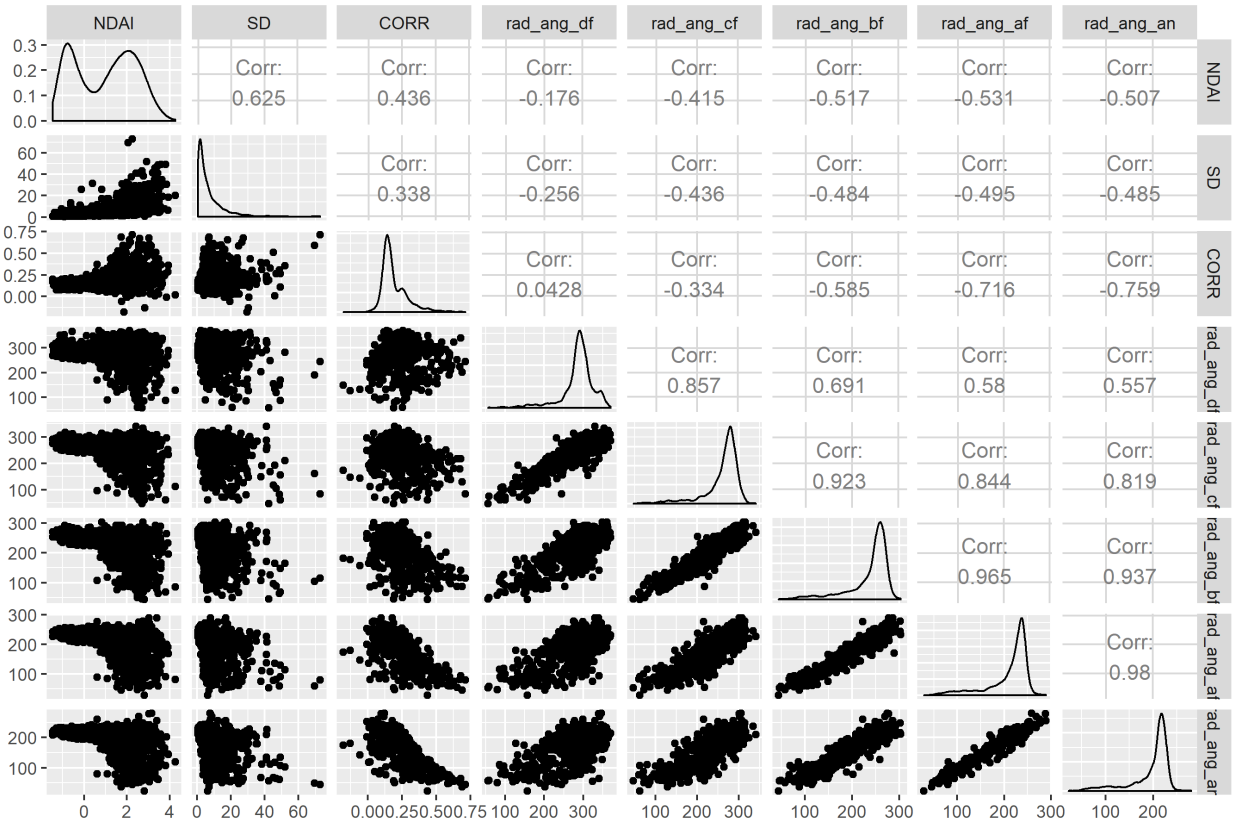


Figure 5: Correlation matrix

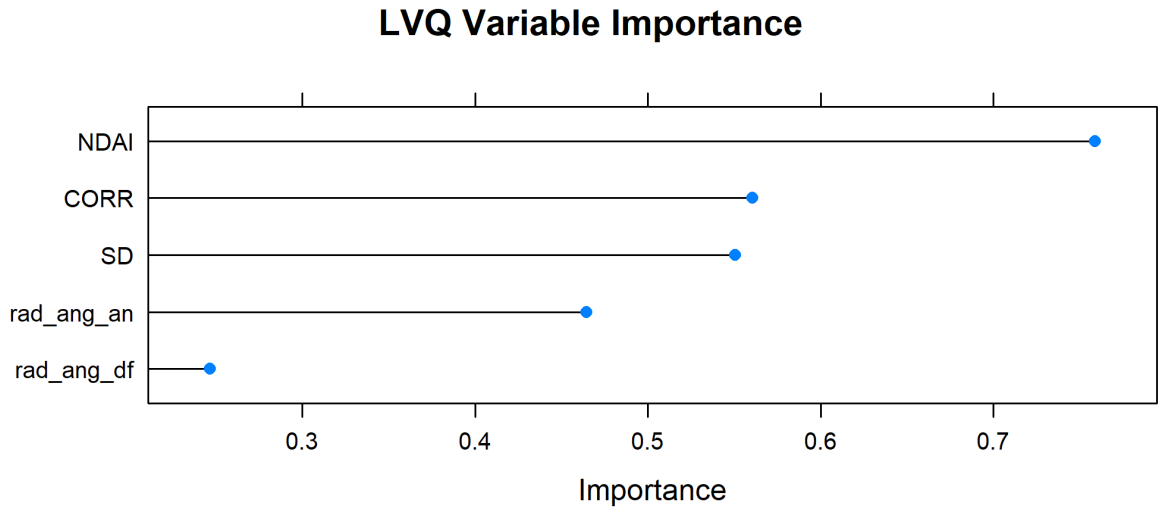


Figure 6: Feature selection

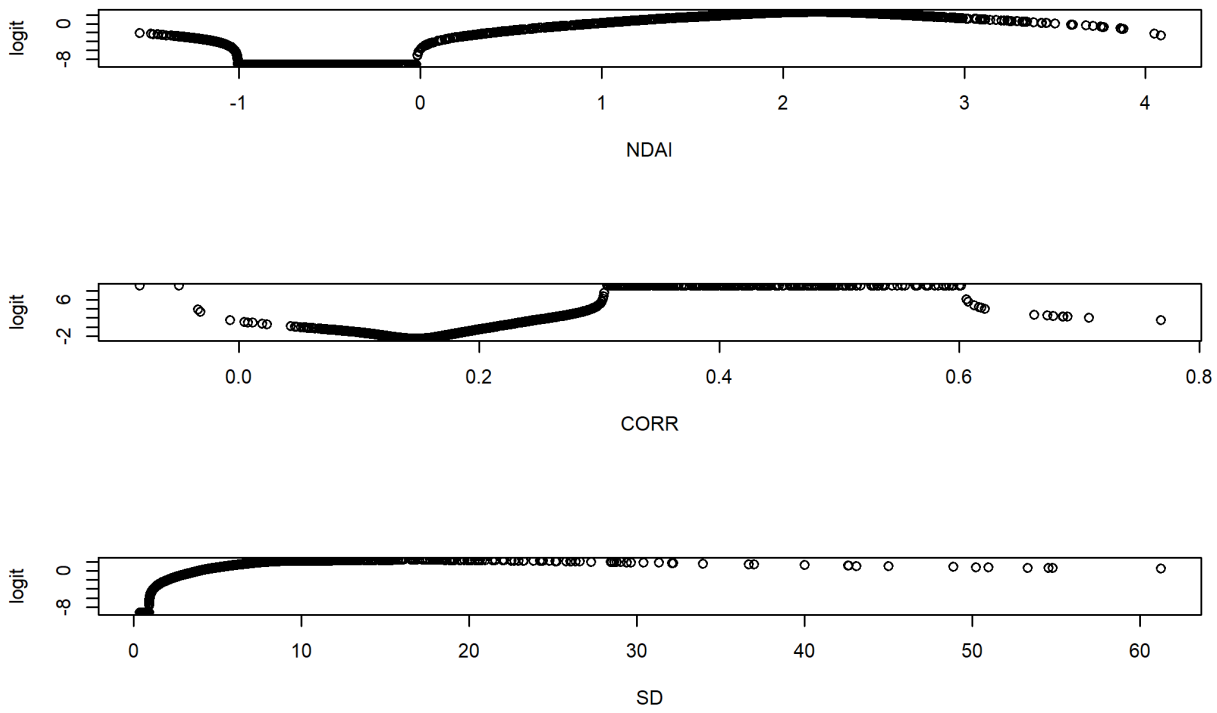


Figure 7: Logit of output against features

3.2 Classifiers

In this section, we develop several 0/1 classifiers to accurately classify expert labels. First, we remove unlabeled points in order to develop a classification rule that can distinguish between binarized cloudy/clear data points. We then try a series of models, each using different assumptions. The three classifiers we plan to pick are Logistic Regression, Support Vector Machine, and Quadratic Discrimnant Analysis (QDA). Here are are going to inspect the assumptions for each of them.

3.2.1 Assumptions for Logistic Regression

Logistic regression has always been a typical 0-1 classifier. Assumptions of logistic regression are:

- Error terms to be independent: we created covariance matrix to check for collinearity. Among the three features we picked - SD, CORR and NDAI, all the pairwise correlations didn't exceed 0.75, and we are confident to say this assumption is fulfilled.
- The observations are linearly related to the log odds. To check this assumption, we could make a plot for the logit against each feature, which is illustrated in Figure 7. In Figure 7, we could see that it is not very much linear related.
- Sample size cannot be too small, usually should be bigger than 30 cases. We already have 230,339 observations, which we believe should be sufficient to fulfill this assumption.

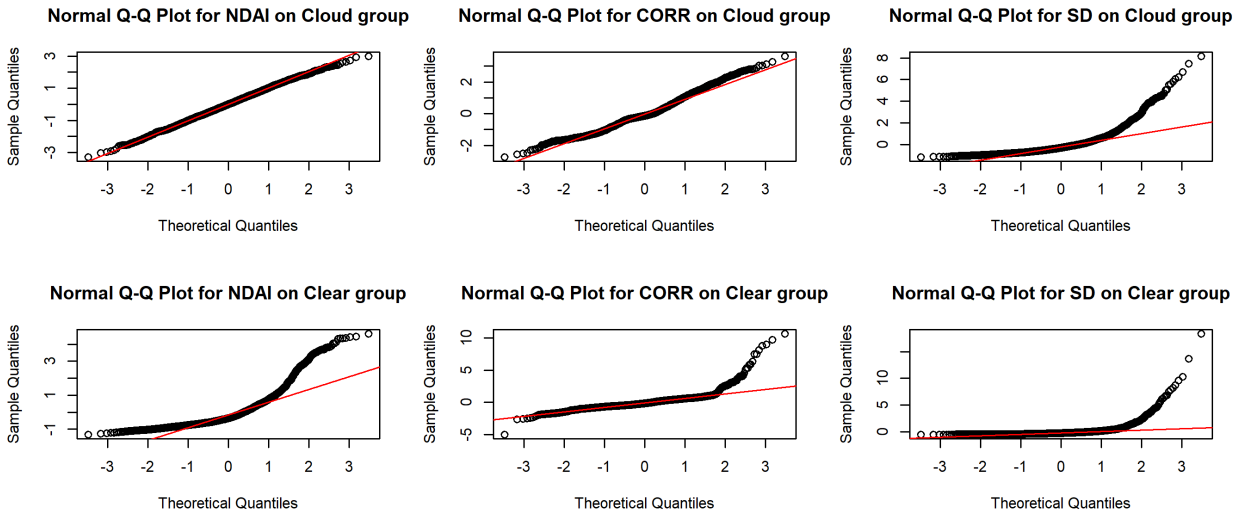


Figure 8: qq plot for features

3.2.2 Assumptions for SVM

While choosing SVM, we assume the two labeled groups can be linearly separated in an n dimensional space. In Figure 4 we illlustrated that NDAI, SD and CORR are actually good features on which the labeled groups can be separated. We also ran SVM using Gaussian kernel to raise the classifier into higher dimension spaces, which could separate these points better.

3.2.3 Assumptions for QDA

In QDA, we assume that our predictors are drawn from multivariate Gaussian distribution. We have several ways to illustrate this. First, from Figure 3 in EDA part, we already created a violin plot to check density curve of each labeled group among these three features. The clear group does not have a normal distribution - it is skewed and sometimes has multiple peaks - but the cloud group looks better. To further check whether the distributions are Gaussian, we created QQ plots for each labeled group on each feature. In Figure 8 we see that the cloud group is following the normal line pretty well, while clear group is still skewed and not very normal. QDA also requires number of predictors n to be smaller than the number of observations, which in our case, we have only three predictors and more than 230k observations, so this requirement is fulfilled.

3.3 Model fit and cross validation

To save time for knitting this report, we trained our modeled in a separate R script file `ml_models.R` which could be found in the `R` folder. The results of each trained model are saved into `.rds` files, and we would load into this report if needed. Before proceeding, we expect that the assumptions for SVM are the most realistic, and expect to get the best results from it.

3.3.1 Cross Validation

After understanding the nature of our data, which is generated from image, we noticed that it would be best to group our observations based on their x and y locations. In this way, we make sure points that are close to

each other stays in the same group, and we believe this grouping cross validation method would benefit our hyperparameter tuning process and proevent overfitting.

Generally, we divde image into 20 by 20 small pieces, and when we tuned hyperparameter for SVM, this grouped cross validation method was used, In below, we created a table to show the result svm hyperparameter tuning:

Table 1: Hyperparamter Tuning

sigma	C	ROC	Sens	Spec	ROCSD	SensSD	SpecSD
3.188308	0.25	0.9755456	0.9327805	0.9692404	0.0082312	0.0177027	0.0105594
3.188308	0.50	0.9756682	0.9341102	0.9687655	0.0083707	0.0183125	0.0079473
3.188308	1.00	0.9754397	0.9350942	0.9670862	0.0085532	0.0176365	0.0078837
3.188308	2.00	0.9744945	0.9356569	0.9655218	0.0084216	0.0169887	0.0086019
3.188308	4.00	0.9735965	0.9366158	0.9639290	0.0076180	0.0178832	0.0099365
3.188308	8.00	0.9726680	0.9370793	0.9640548	0.0075010	0.0180720	0.0104337
3.188308	16.00	0.9725309	0.9368305	0.9638068	0.0067155	0.0179708	0.0102447
3.188308	32.00	0.8771172	0.9333203	0.9673776	0.2156806	0.0194068	0.0077202
3.188308	64.00	0.9735740	0.9357482	0.9651670	0.0056767	0.0186278	0.0096716

3.3.2 Model fit

To compare the model to one another, we chose to run Receiver Operating Characteristic (ROC) options in the training model. This is advantageous because it allows us to compare the ratio of accurate predictions to innaccurate ones, which is the central concern of this paper. Below we summarize each model along with useful measures

Table 2: Model Selection Table

	ROC	Sens	Spec	ROCSD	SensSD	SpecSD
Logistic Regression	0.9591091	0.9242101	0.9240978	0.0134908	0.0127743	0.0094025
QDA	0.9650968	0.9287806	0.9439658	0.0131754	0.0148934	0.0053968
SVM	0.9756682	0.9341102	0.9687655	0.0083707	0.0183125	0.0079473

Next we plot a confusion matrix and the ROC under the curve for the SVM predicted classes. High values in the cells corresponding to "True Positive" (top left) and "True Negative" (bottom right) are desirable in this case as they indicate overall model accuracy through frequency counts. Then we plot the ROC curve for the SVM model, and it has an area under the curve of about .85 (with close to 1 being descirable).

Table 3: Confusion Matrix

	clear	cloud
clear	29255	6295
cloud	4497	14949

3.4 Model Diagnostics

Below we conduct some basic model diagnostics on our SVM model (Figure 10). The density of the residuals plot (bottom right) suggests a bimodal distribution of the residuals. This indicates non-normality in the

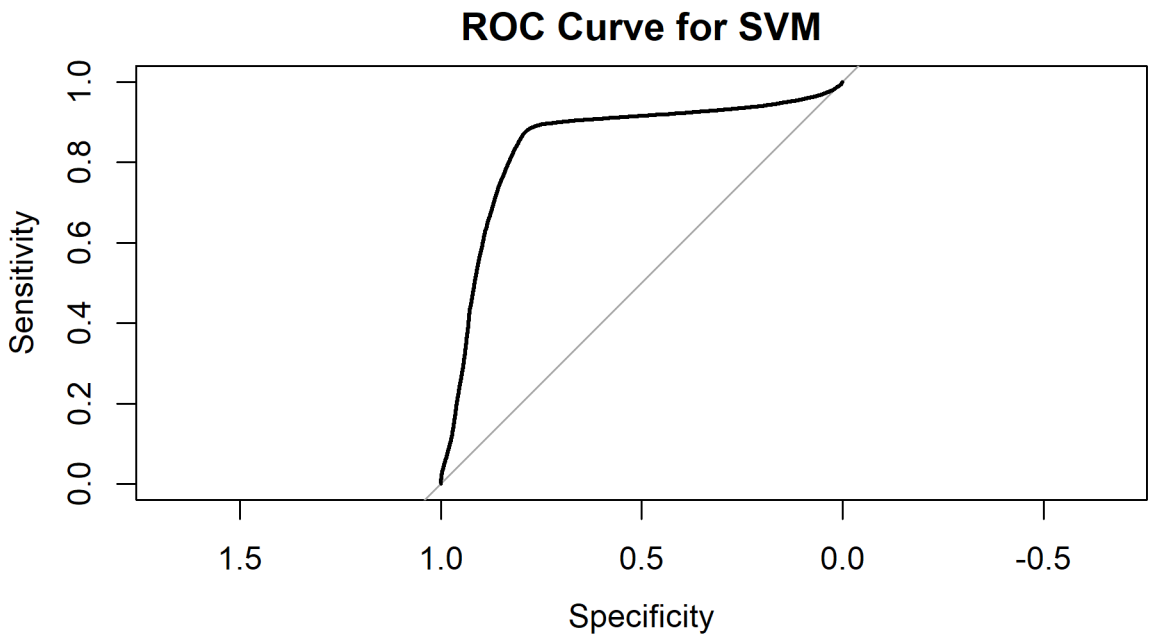


Figure 9: ROC Curve for SVM

data, which could be problematic for out-of-sample inference.

3.5 Misclassification Errors

Overall, the SVM model performed the best in classifying points properly. The one exception comes in Image 1 (Figure 11) in the oval-shaped region that lies between 100 and 200 on the x-coordinate axis and about 0 and 100 on the y-coordinate axis, where it classifies most "clear" points as cloudy. Subsetting to the misclassified data and plotting density plots of the distributions of the NDAI scores, shaded by whether it was truly cloudy or clear or predicted to be so, reveals a significant disparity.

Most notably, the points misclassified by SVM have a unimodal shape centered around 1 (Figure 12). However, while looking at all data points, there was no peak, and in that area, cloudy and clear points shared similar proportions. We believe that the points around 1 were misclassified because neither the cloud nor clear classes dominated the population around that range, so the classifier is prone to make mistakes there.

3.6 Performance on data without expert labels

To test performance on an out-of-sample dataset, we purposely withheld image 3 from the training data so that we could test the algorithms on it. The SVM model had an accuracy rate about .8 in this new dataset (visualized in Figure 13). While this is a fairly encouraging rate, it may not be adequate for real-world use. The most likely culprit for the mistakes was that the algorithm predicted a fairly uniform distribution of NDAI scores (Figure 14), whereas in the actual dataset, it was unimodal around a score of 1.5.

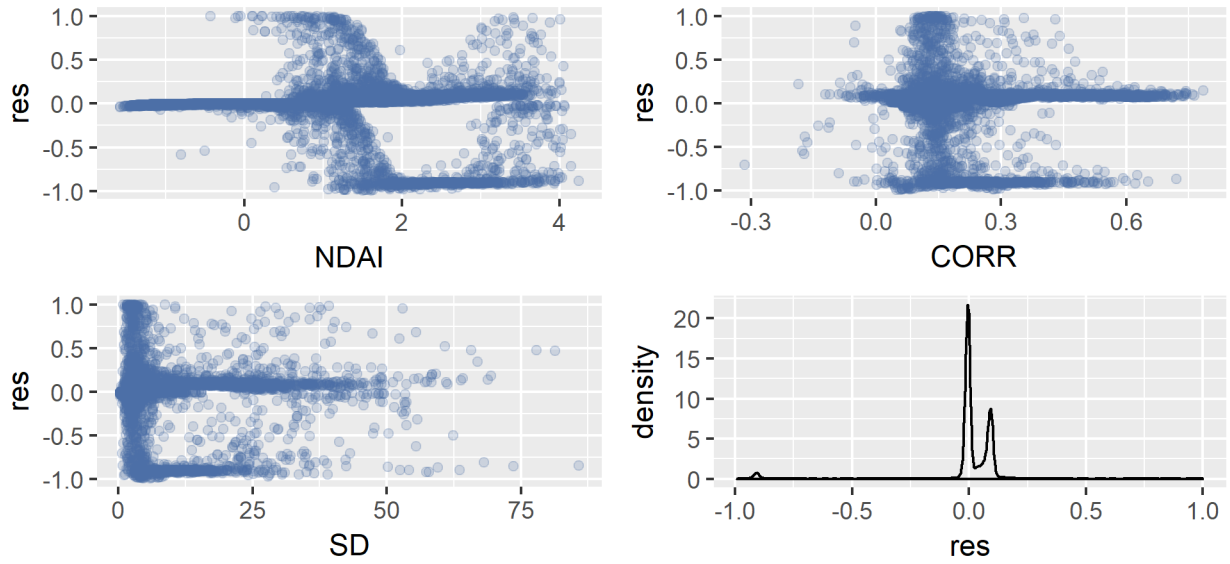


Figure 10: SVM Diagnostics Plot

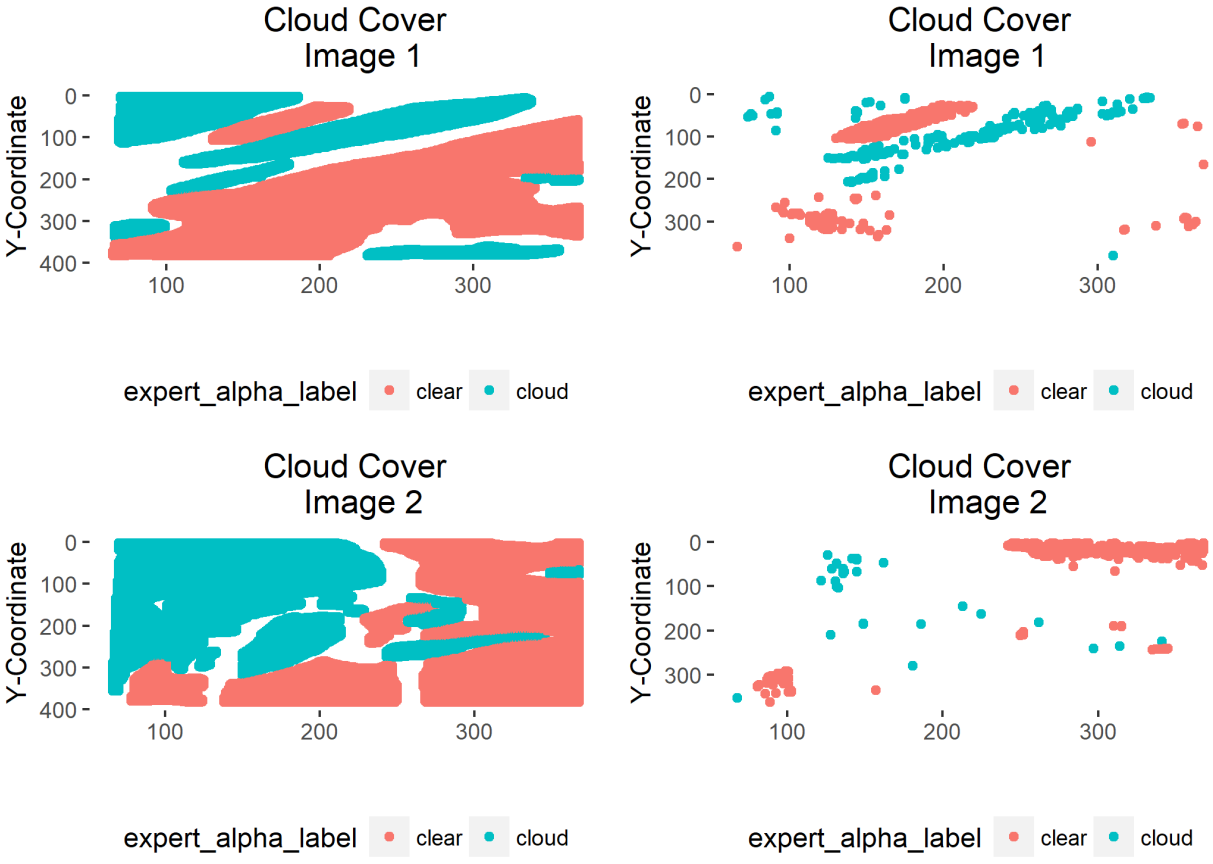


Figure 11: Image 1 True v. Predicted Cloud Cover

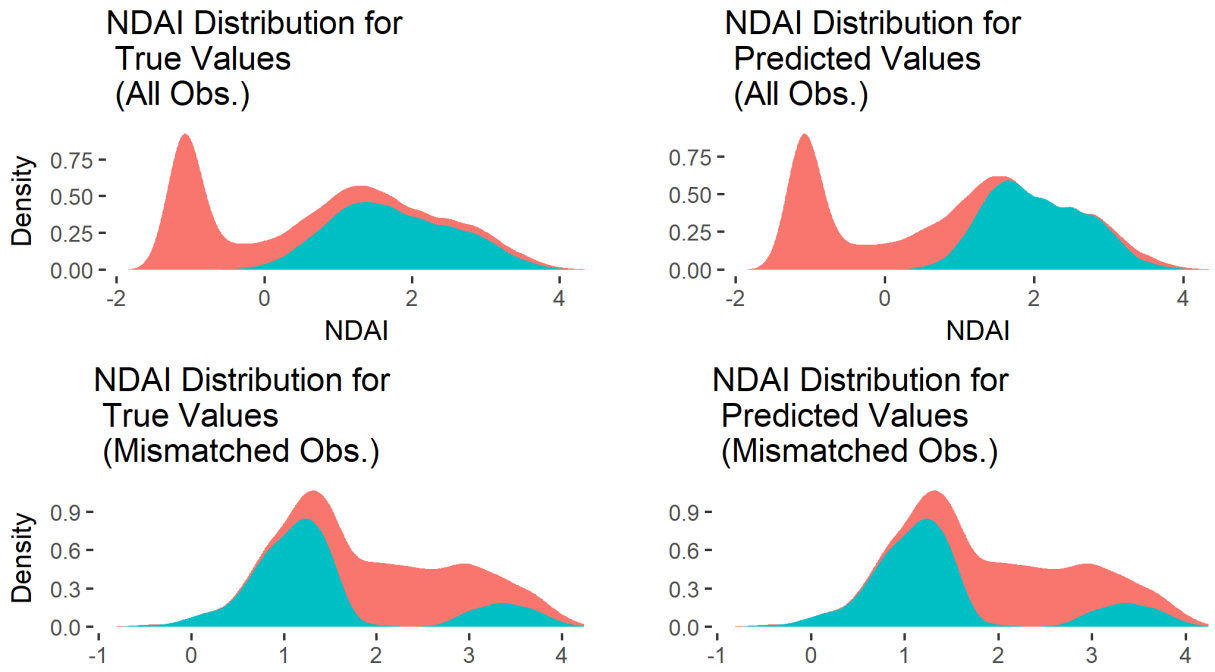


Figure 12: NDAI Distributions for True vs. Predicted

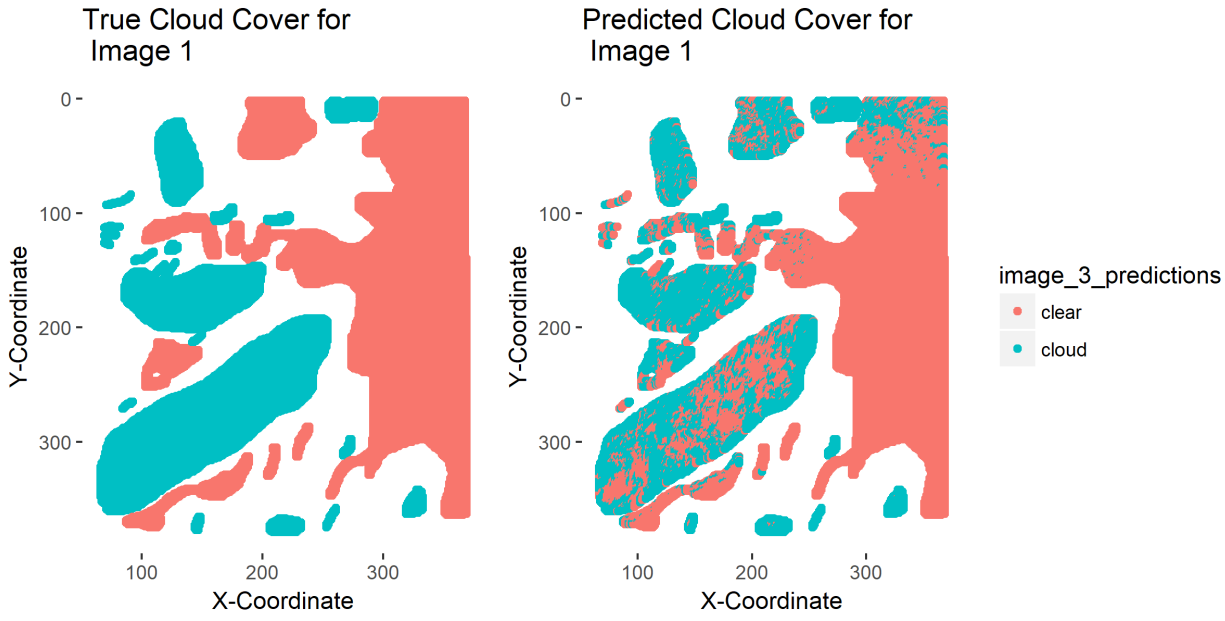


Figure 13: Image 3 True v. Predicted

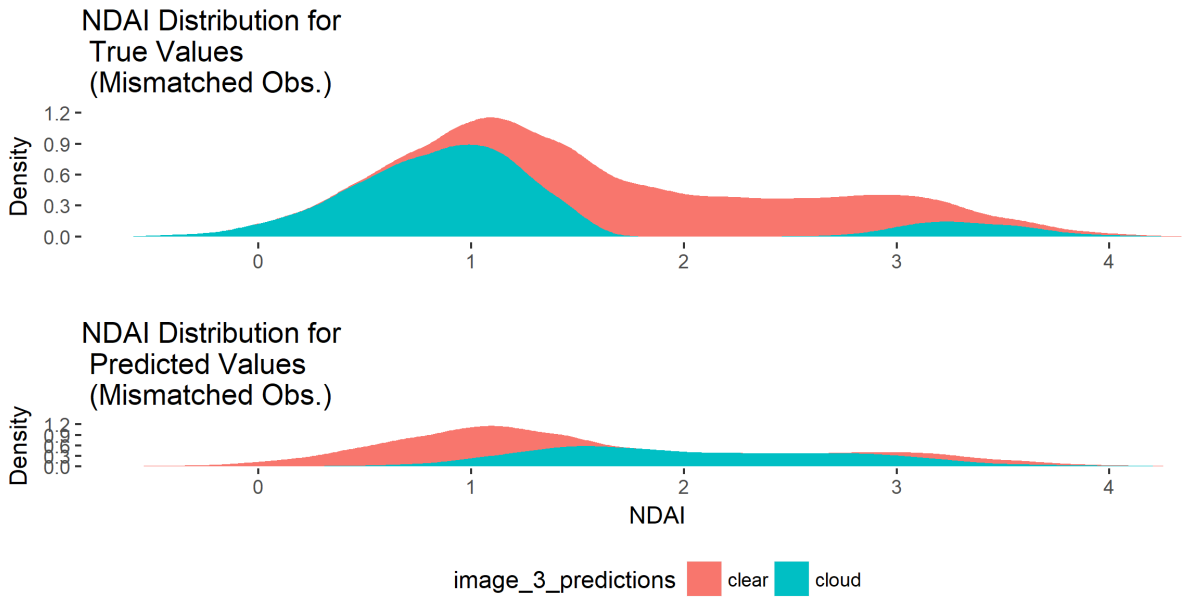


Figure 14: Image 3 NDAI dist.

4 Conclusion

In this lab, we developed predictive models for identifying cloudy regions in satellite imagery. To do so, we first explored the relationship between camera angles and radiation, selected covariates, and built several candidate models. Overall, non-parametric methods performed quite well within sample (both in the training and test sets), and was able to predict points in a new dataset with about 80% accuracy. We checked the SVM model against other candidate models, and our initial hunch that it would have the best performance was confirmed.