HW154_Project2

Huy Le, RJ Lee
(3032370043,3034269840)

1. Data Collection and Exploration

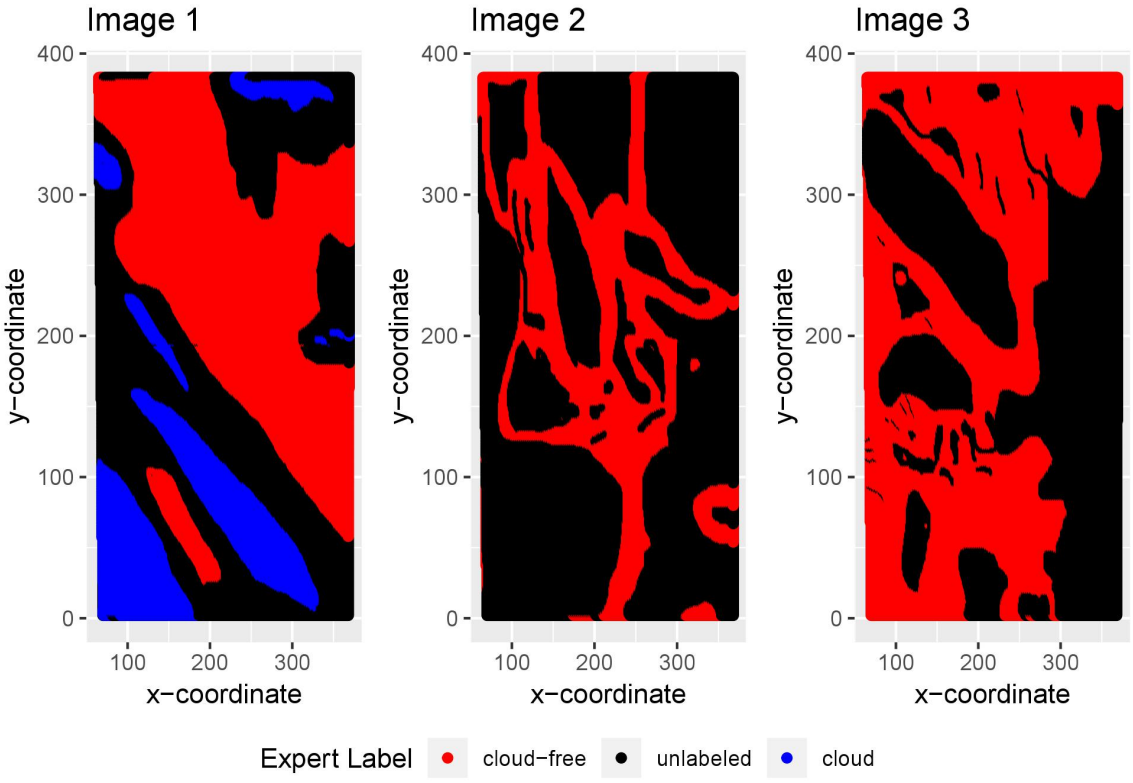
Question 1 - A

The purpose of this paper is to do cloud detection in the Arctic because of the increasing atmospheric carbon dioxide levels occur in this area, which relates directly to global warming. However, due to the similar remote sensing characteristics of clouds, ice- and snow-covered surfaces, it is hard to apply the MISR operational algorithms. The data from this study was obtained from 10 MISR orbits of path 26 over the Arctic, Northern Greenland, and the Baffin Bay, which is approximately 144 days from April 28 through September 19, 2002. There are six data units, which are MISR blocks 11 – 13, 14 – 16, 17 – 19, 20 – 22, 23 – 25, and 26 – 28 from each orbit are included in this study. However, three of the 60 data units were excluded from this study because the surfaces were open water after the sea ice melted in the summer, and the MISR operational algorithm detects clouds over water well. The research demonstrated with three features, linear correlation of radiation measurements from different MISR view directions (CORR), the standard deviation of MISR nadir red radiation measurements within a small region (SD_{An}), and a normalized difference angular index (NDAI), are enough to sufficiently identify clouds, ice- and snow-covered surfaces from each other. This identification can be obtained by using ELCM algorithm that combines classification and clustering frameworks. The method is as follows: construct features based on EDA, use ELCM to produce the first cloud detection product, and then predict the probability of cloudiness by training QDA. After getting the results, they can be used to train QDA to provide probability labels for partly cloudy scenes. The work in this study is useful in determining cloud coverage, which can contribute to the accuracy of current global climate models. Additionally, this study is also significant for statistics which demonstrates the power of statistical thinking and contributions as well as the role of statistics in solving current science questions.

Text

Question 1 - B

As you can see from the three horizontally aligned maps below, the three maps look quite different. These satellite map visualizations are highly dependent on the time and day each satellite image was taken. For example, if two satellite images were taken soon after the other, then the two satellite images are much more likely to resemble each other compared to another image taken few days earlier at the same location. Since these satellite images are highly dependent on spatial and time components, Independent & Identically Distribution assumption is not justified for this dataset.



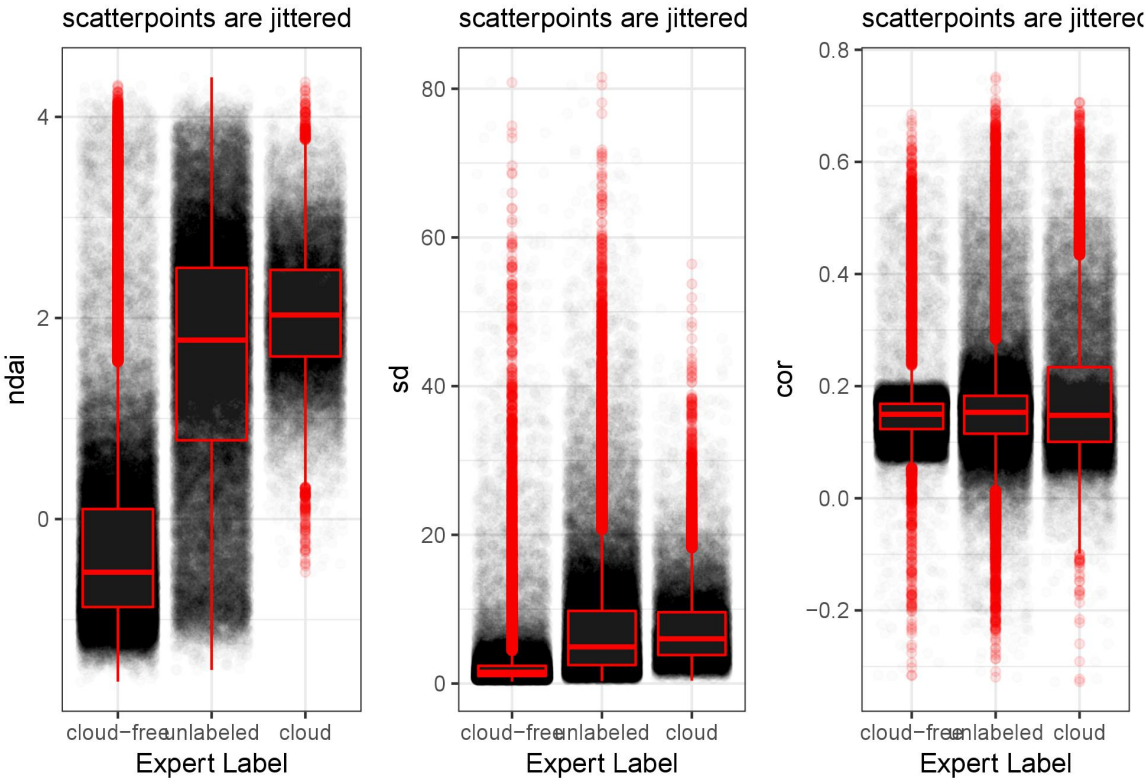
Question 1 - C

For this particular section, image1.txt was chosen for the analyses. The reason for picking image1 is that from Part 1-B's EDA map visualization, we see that image1 is the only map data that contains non-neglibiable observations that correspond to all three unique Expert Labels. The map visualizations correponding to image2 and image3 does not seem to have any "cloud" label with bare eye.

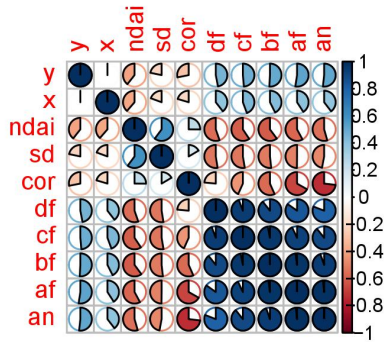
NDAI against Expert Labels clearly shows a clear relationship. When Expert Label is "cloud free", the large proportion of NDAI's distribution is concentrated on the negative real value, in fact seventy-fifth percentile is close to 0. However, the NDAI values corresponding to when Expert Label is "cloud" are mostly positive number.

SD against Expert Label displays interesting pattern as well. As you can see from the overlying boxplots, when the Expert Label "cloud", SD is much more spread out than when the Expert Label is "cloud-free". The median is smaller when Expert Label is "cloud_free". In both Expert Label cases, there are outliers whose values are much greater than the seventy-fifth percentile.

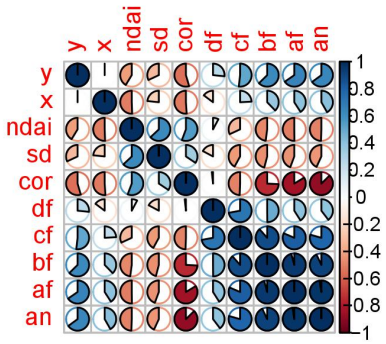
Finally, CORR against Expert Label shows the following pattern: the distribution of COR is more spread out when Expert Label is "cloud" than when Expert Label is "cloud free". In both labels, there are outliers much smaller than twenty-fifth percentile value and outliers much greater than seventy-fifth percentile.



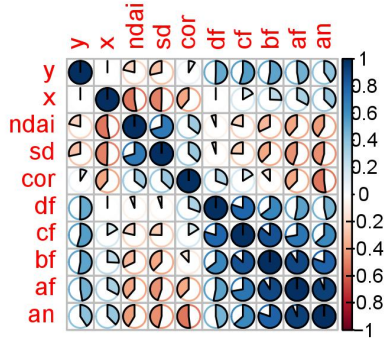
Corr matrix of Image 1 data



Corr matrix of Image 2 data



Corr matrix of Image 3 data



2. Preparation

Question 2 - A: (Data Split)

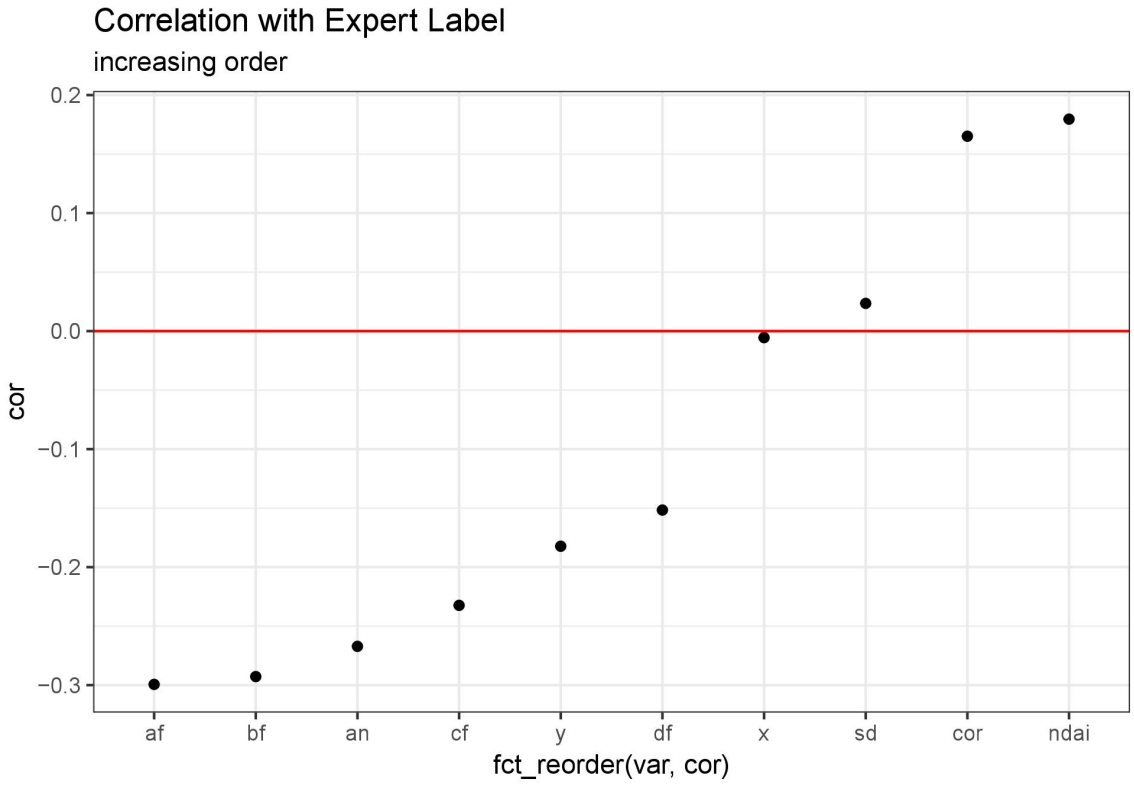
As mentioned above, Independent & Identically Distribution assumption is not justified for this dataset. Therefore, data splitting was done separately for each of the three datasets before they were merged. To be more specific, image1 data was split into training, validation and test data in 70%, 20%, and 10% ratio. The same splitting was done for image2 and image3. Then, the training data sets of image1, image2, and image3 were rowbinded to create one (big) training dataset caled `image_train`. Same method was used to create `image_val` and `image_test`.

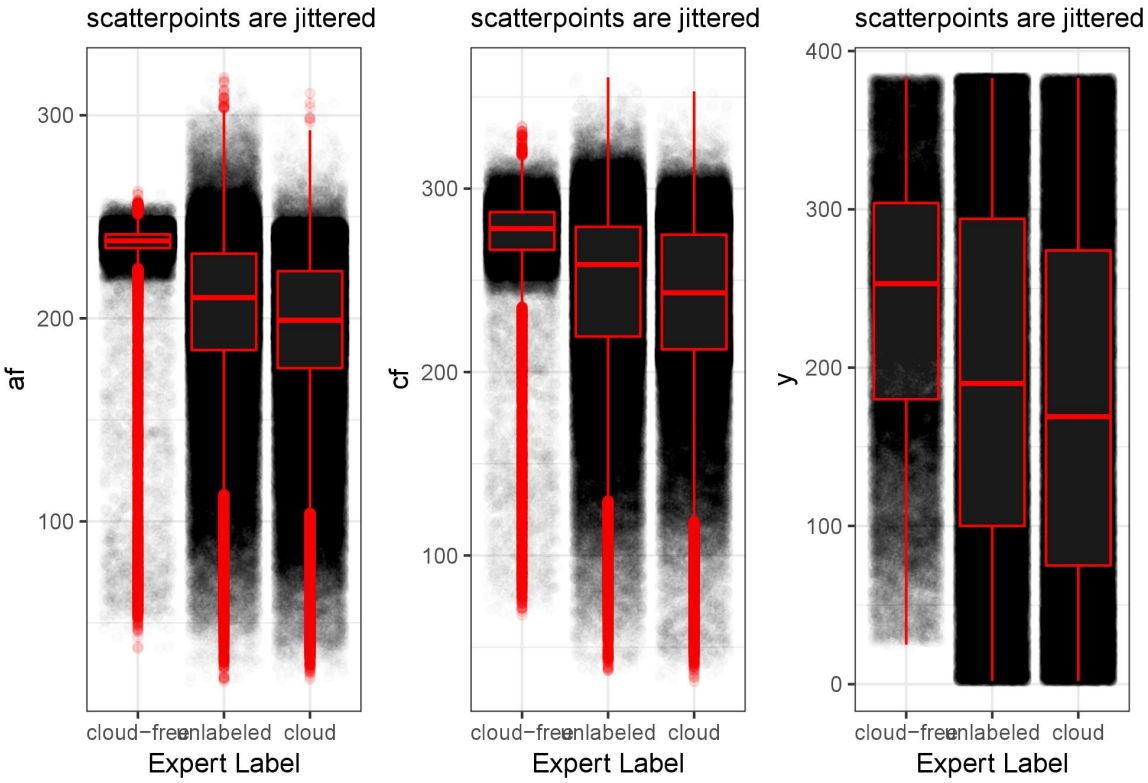
Another suggested method is the following: the ratio we are using here is 70 (training), 20 (validation), 10 (testing). In order to account that the data is not i.i.d. we can first concatenate all three image data sets, then we will split base on the expert labels i.e. -1, 0, 1. Specifically, we split observations with expert label 1 using the ratio 70-20-10 for the train-valid-test set, and similarly for 0 and -1. After that, we can concatenate them by set.

Question 2 - B: (Baseline)

The accuracy of the trivial classifier which sets all labels to -1 achieves accuracies of 12.2% and 3.7% on validation set and test set, respectively. Such trivial classifier will achieve high accuracy when the proportion of -1 is high in the Expert Label field in the validation and test sets.

Question 2 - C: (First order importance)





Since no classification is done here, the entire dataset - created by merging image1, image2 and image3 - was used to suggest three of the "best" features.

The suggested three best features for correctly classifying Expert Labels are 'AF', 'CF', and 'Y'. Firstly, the correlations between Expert Labels and all the other features were obtained. The four features with the largest aboslute value correlation values are the four radiance angles: AF, BF, AN, and CF. However, multicollinearity needs to be considered. The correlation between AF and AN is 0.97, which means both of the mcontain essentially the same information. Hence if AF is used as one feature to predict Expert Labels, AN will not supply any additional new information to better predict Expert Labels. The same multicollinearity problem occurs with AF and BF. Therefore, both AN and BF features will not be included. Thus, the third suggested feature is 'Y', which had displayed fifth highest correlation value with Expert Label.

Question 2 - D: (K-fold CV)

Uploaded on GitHub, "CVgeneric".

3. Modeling

Question 3 - A: several classification methods

5 fold cross-validation was used to evaluate different models, such as Logistic regression, Probit regression, Linear regression, LDA, and QDA.

For Logistic and Probit regression, the response variable Expert Label had to be modified so that the response variable fall in between 0 and 1. The researchers in this study are most interested in finding cloud-free regions which are originally labeled as -1. All the cloud-free regions' labels were **re-labeled** to 1 and the other two labels (not cloud -1 & unlabeled 0) were all **re-labeled** to 0. Note however, after this modification, roughly

84.2% of Expert Label is 0, so we want to make sure that the classification can beat this relatively high threshold.

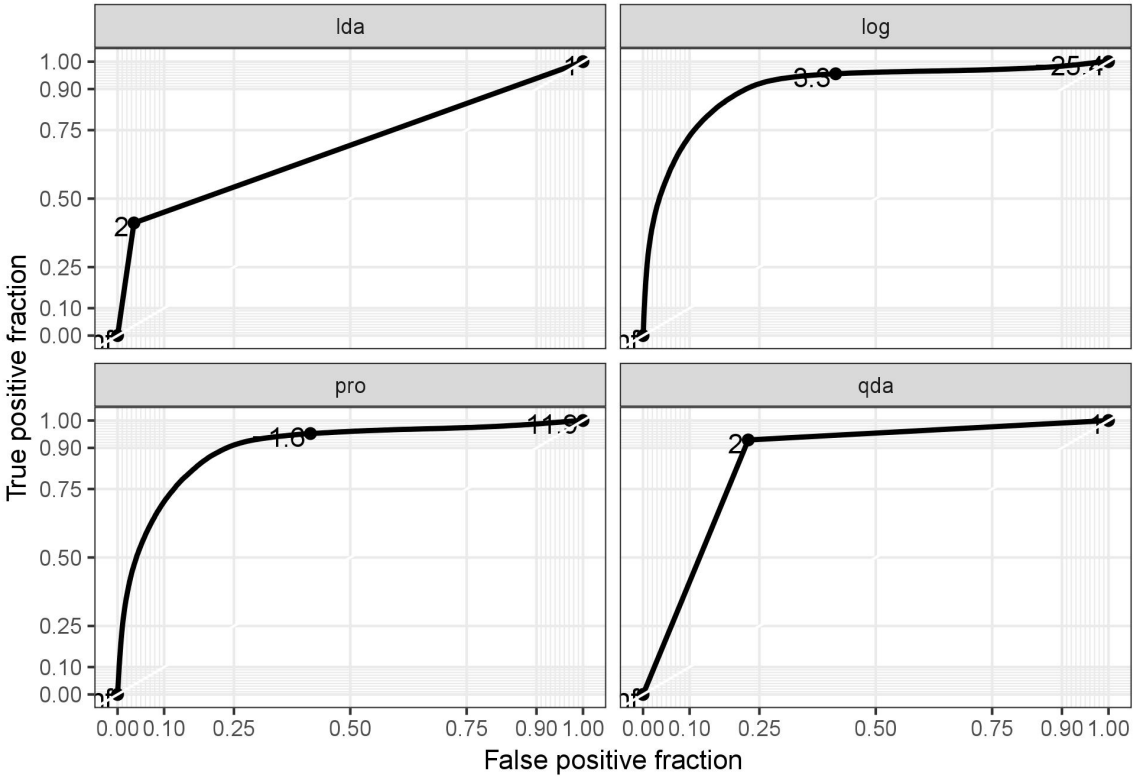
5-fold cross validation was conducted for Logistic and Probit regressions first. The training accuracy for both regressions was higher than 87% threshold mentioned above: Logistic regression's training classification had an average accuracy of 89.6%, while Probit regression's training classification had an average accuracy of 88.2%. The corresponding test classification accuracies for Logistic and Probit regression were 96.29% and 96.34%, respectively. In other words, they performed equally well on unseen test datasets. However, these high proportion is **misleading**.

Next, simple linear regression classification was ran, which resulted in average training accuracy of 52.3% and test classification accuracy of 52.2%.

The last two classification methods modeled were LDA and QDA, both of which performed better than the simple linear regression classification. The average training accuracy for LDA was higher than that of QDA; LDA had an average training accuracy of 95.6%, while QDA had an average training accuracy of 87.9%. The test accuracy for LDA and QDA were 70.7% and 64.2%, repectively. So in both training and test datasets, LDA performed better.

Question 3 - B: ROC curves

The strategy the researchers are using in this paper is to detect cloud-free region, indicated by "-1" in the dataset. Therefore, as was mentioned before, it is not unreasonable to do **one vs all** where we group cloudy and unlabeled responses into one group (re-labeled as 0) and cloud-free as the other group (re-labeled as 1).



As you can see from the plots above, LDA and QDA show similar ROC curves, and Logistic and Probit regression show similar ROC curves. In fact, the ROC curves for Logistic and Probit regressions are almost on top of each other. Their True positive rate increases very fast until it reaches 0.9 and then False positive cases start to follow up. The ROC curves for LDA and QDA looks like a piecewise linear graphs where true

positive rate very rapidly increases up to a certain point, after which the false positive rate increases very rapidly.

In order to find the cutoff point, we will usually choose the point on the ROC curve that is furthest than the 45 degrees (diagonal) line. However, in this situation, between classifying cloud and cloud-free, we are willing to make an error to misclassify cloud to cloud-free. The cutoff points in LDA and QDA are based on the first criteria.

Question 3 - C: Bonus

Cohen's Kappa statistic: Kappa statistic can adjust accuracy by taking into account of the fact that correct predictions can be made from chance. For example, in the above case, I have modified expert labeling column so that 1 is cloud-free and 0 is otherwise. So there are 270602 cases where expert label is equal to 0, which is roughly 87% of the case. Hence, just guessing 0 will give an accuracy of 87%. Kappa's value, which ranges from 0 to 1, will increase when the classifier performs better than this simple way of picking the most frequent class.

The formula for Kappa statistic is as follows:

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)},$$

where $P(a)$ refers to the proportion of actual agreement from classification, and $P(e)$ denotes expected agreement under random chance only. I am going to use a package called Visualizing Categorical Data `vcd` which contains a command that can calculate Kappa statistic for us.

Model	Kappa
Logistic	0.4426628
Probit	0.2473138
Linear	0.0750131
LDA	0.4656745
QDA	0.3804665

Cohen's Kappa statistics on test set were obtained above using the model that was trained on the merged dataset that concatenated both the training set and the validation set. We see that the Linear regression have the smallest Kappa statistics; its Kappa Statistic value is less than 0.1, which means it is only slightly better than the baseline classifier.

However, we observe that LDA and QDA can correctly classify better than the linear regression based on Kappa statistic. When LDA, QDA, and linear regressions were evaluated using cross validation in section 3-A, LDA performed the best, then QDA and lastly Linear regression. We see the same performance order through Kappa statistic as well: the Kappa statistic for LDA is the largest, then followed by QDA and then lastly Linear.

Both Logistic and Probit had very similar ROC curves, but Logistic regression had better cross validation accuracy. Fortunately, the Kappa statistic analysis also provides the same result: the Kappa statistic for Logistic regression is greater than that of Probit. However, we should wary of giving too much confidence in these two models. When the two models were trained, Expert Labels had to be modified so that there were only two levels, 0 and 1. As mentioned above when such modification is implemented, the majority of the labels end up becoming 0 and thus a baseline classifier that just labels 0 for all observations will still give high accuracy. Therefore, the Kappa statistics for Logistic regression and Probit regression is **not** comparable to the Kappa statistics of the other classification models.

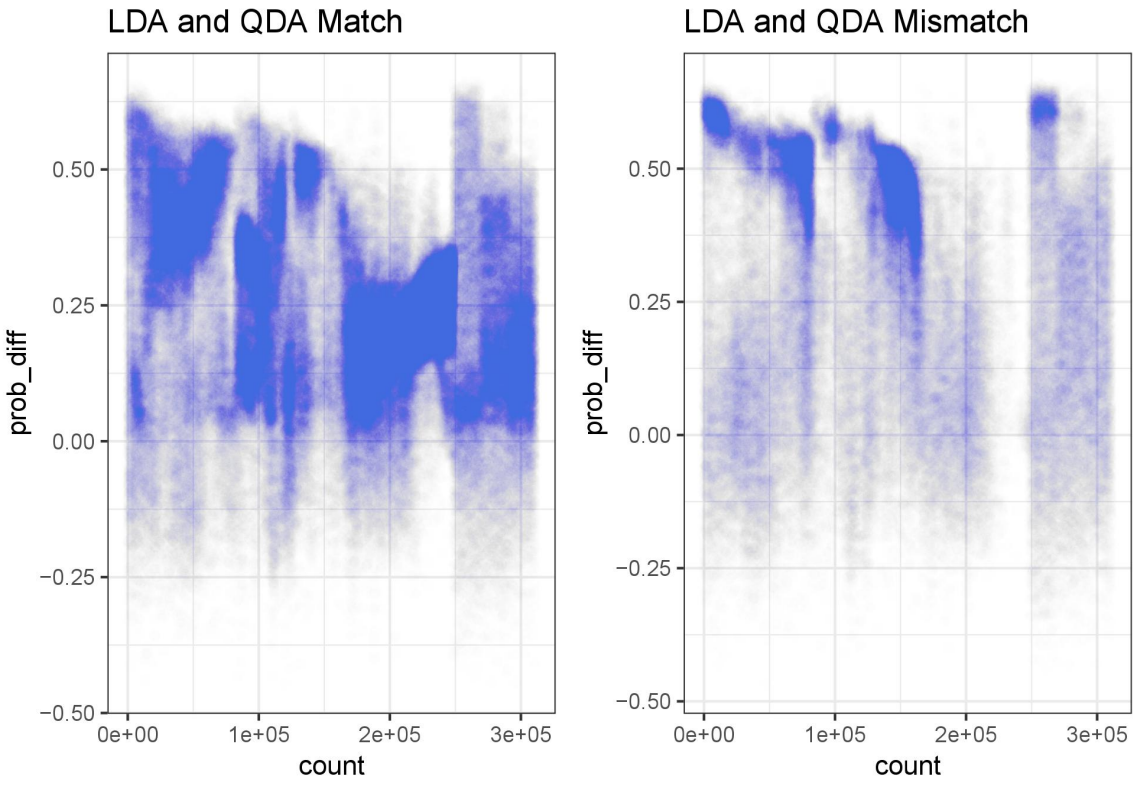
4. Diagnostics

Question 4 - A

Compare Contrast of Discriminant Analysis: LDA vs QDA

```
## # A tibble: 2 x 11
##    `1F`  `2F`  `3F`  `4F`  `5F`  `6F`  `7F`  `8F`  `9F` `10F` Method
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <fct>
## 1 0.958 0.953 0.957 0.956 0.957 0.957 0.954 0.957 0.956 0.955 LDA
## 2 0.878 0.879 0.879 0.882 0.876 0.879 0.877 0.877 0.881 0.881 QDA
```

Here we have the average of LDA is 0.956 and 0.879 for QDA.



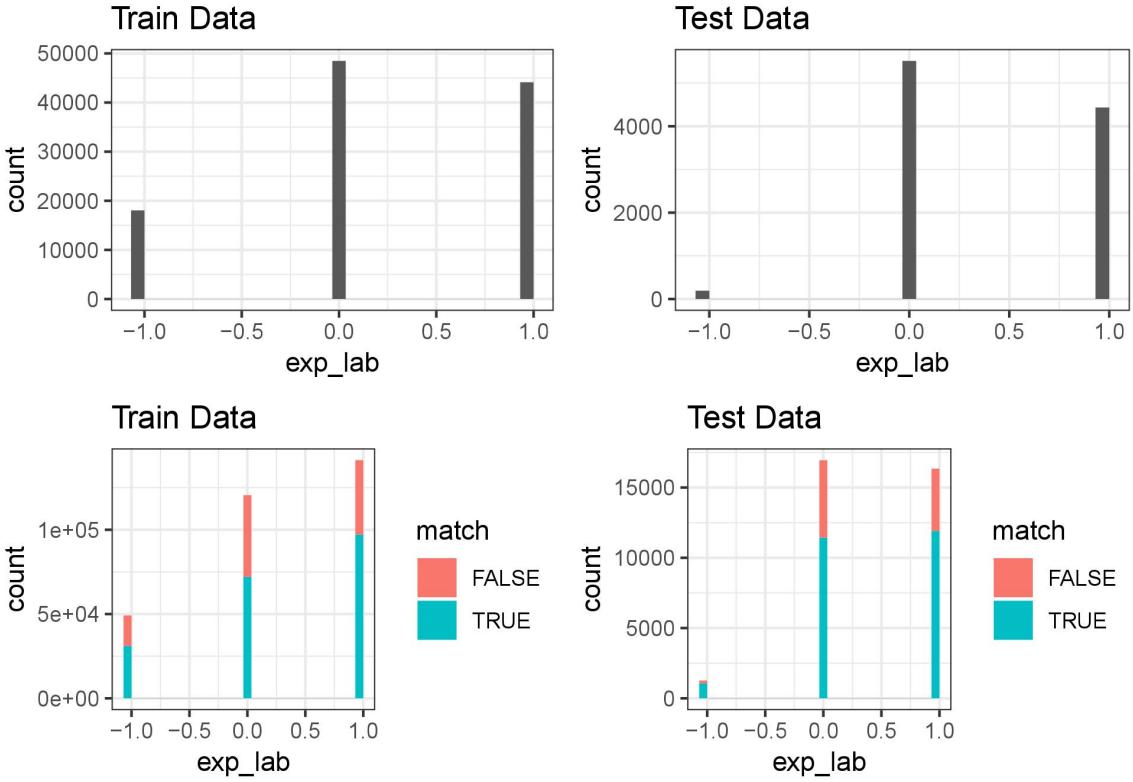
Both LDA and QDA's Kappa statistics were much greater than that for Linear regression classification. Also, both classification methods can handle multiple levels in the response variable. LDA and QDA assign certain likelihood probability to each label for each observation, and then the label that gets the greatest possibility gets to be chosen.

Hence to analyze how LDA and QDA predict labels, the greatest probability assigned to each label was isolated for both models. The difference between the greatest probability used by LDA was subtracted from that used by QDA.

The above procedure were repeated for two cases: when LDA and QDA predictions match and when LDA and QDA classification mismatch. In both cases, we observe that from the visualization above that most of the time the difference in probabiliy is positive, meaning that QDA usually assigns greater probability to the label that ultimately gets to be chosen while LDA usually assigns more comparable probability to the three labels.

Question 4 - B

LDA has the highest corresponding Kappa statistics value, so LDA will be analyzed. Among the misclassified cases for the training set, only about 16% of the labels are -1 (cloud-free). Among the misclassified cases for the test set, only about 2% of the labels are -1 (cloud-free). This tells us that most of the misclassification comes from the other two labels.



Question 4 - C

The baseline accuracy that needs to be beaten is 0.841866. The accuracy on train data is 0.8771873 and the accuracy on test data is 0.963709. From section 4 -B & C, we notice that only small amount of misclassification comes from the case when the label is equal to -1 (cloud free), and so most of the misclassification comes from the other labels. Therefore, it is not an unreasobale idea to cluster both "unlabeled" and "cloud" cases into one so that when the model gets trained, the model has more observations to learn about the case when it is **not** "cloud-free", and thus improve the accuracy.

LDA assumes equality of covariance matrices of the predictor variables X across each all levels of Y. This assumption is relaxed with the QDA model, which assume covariance matrices are not i.i.d. When considering between LDA & QDA its important to know that LDA is a much less flexible classifier than QDA, and so has substantially lower variance. This can potentially lead to improved prediction performance. But there is a trade-off: if LDA's assumption that the the predictor variable share a common variance across each Y response class is badly off, then LDA can suffer from high bias. Roughly speaking, LDA tends to be a better bet than QDA if there are relatively few training observations and so reducing variance is crucial.

Classifier that can assumes different covariance but NOT from predictor variables X are drawn from a multivariate Gaussian (aka normal) distribution.

Question 4 - D

Nothing much changes when we consider the data is i.i.d, it is appears that QDA performs better than the LDA. However, when split the data without assuming i.i.d. LDA performs better than QDA.

Question 4 - E

After doing the analysis, we can see that both LDA and QDA perform better than Linear Regression classification. Additionally, logistic regression had better cross validation accuracy. We also discovered that the results of classification analysis are sensitive to how we split the data. In particular, the i.i.d assumption does affect the result of the analysis. Moreover, in this project, we attempted to implement SVM; however, it took too long to compile. Hence, we decided not to include the results in the project.

Acknowledgement

Contribution by Huy Le:

*Summarized the paper

*Considered different methods to split the data

*formatted the final paper

Contribution by RJ Lee:

*Genereated functions that can produce different visualizations for EDA

*Wrote the CVgeneric function

*trained models on the training set and evaluated on test sets

Information about **Cohen's Kappa statistic** was found from a textbook titled, *Machine Learning with R*- Second Edition by Brett Lantz.