High-dimensional statistics

Project 2: Generalized linear models and classification

Corentin Merle s162662
Jad Akkawi s216641

Data Science and Engineering, bloc 2
Liege University
School of Engineering
Academic year 2022-2023

1 Data & Preliminaries for the supervised classification

For this project we chose to stick with our Project 1 `Air Quality Data Set` from Saverio De Vito, ENEA-National Agency for New Technologies, Energy and Sustainable Economic Development [1].

The purpose of the data is to calibrate an electronic nose's response, i.e. a multi-headed sensor's response using ground truth measurements of air pollutants from a measuring station near the location of the e-nose. In this context we decided to classify whether the air contains a significant amount of Nitrate Oxides (NOx) using the "true" and "sensor" measurements of other air pollutants. In order to specify a dichotomizing threshold for "High NOx" and "Low NOx" we referred to `ToxFAQs™ for Nitrogen Oxides` [2] that states: "The EPA [Environmental Protection Agency] has established that the average concentration of nitrogen dioxide in ambient air in a calendar year should not exceed 0.053 parts of nitrogen dioxide per million parts of air (0.053 ppm)." The article concerns all nitrate oxides so we decided to double the normal average value 0.053 ppm and set our threshold to 0.1 ppm or equivalently 100 ppb (to conform with dataset units).

We thus obtain an unbalanced binary variable with 174 Fails (Low NOx) and 326 Successes (High NOx). Refer to Appendix A for a complete list of dataset variables and there statistical summaries.

Our variables belong to different orders of magnitude therefore a standardization was done before starting our investigation to avoid the vanishing gradient problem during the training phase of the logistic regression exercise.

We proceeded by plotting the Boxplots of various variables while splitting according to the newly derived NOx classes. On Figure 1 to 10 you can observe Boxplots of NOx binary level (FALSE=Low, TRUE=High) with respect to variables.

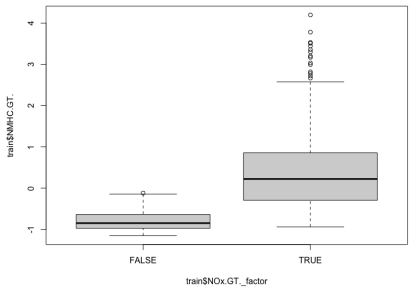

Figure 1: NOx w.r.t NMHC.GT

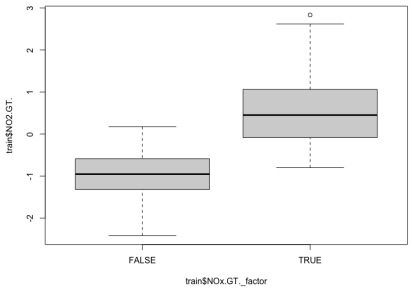

Figure 2: NOx w.r.t NO2.GT

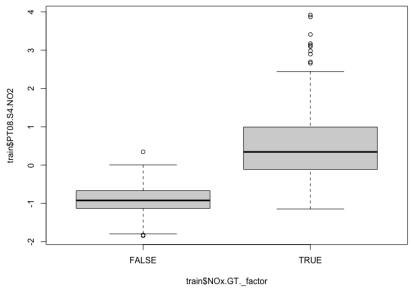

Figure 3: NOx w.r.t S4.NO2

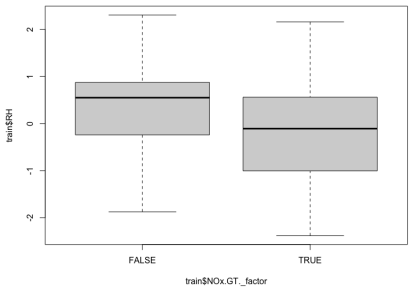

Figure 4: NOx w.r.t RH

[1] https://archive.ics.uci.edu/ml/datasets/Air+Quality

[2] https://wwwn.cdc.gov/TSP/ToxFAQs/ToxFAQsDetails.aspx?faqid=396&toxid=69

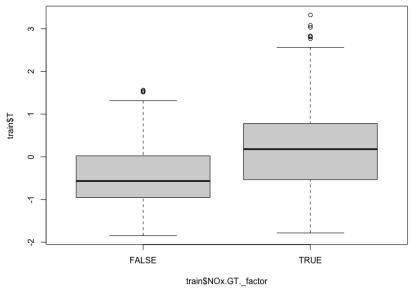

Figure 5: NOx w.r.t T

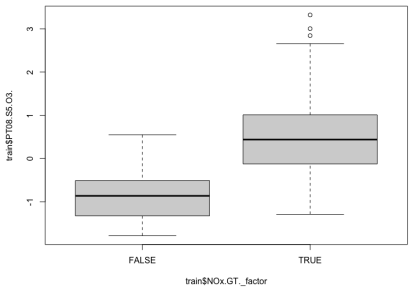

Figure 6: NOx w.r.t S5.O3

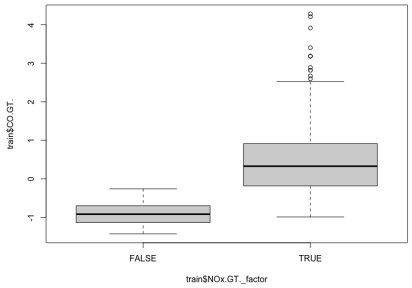

Figure 7: NOx w.r.t CO.GT

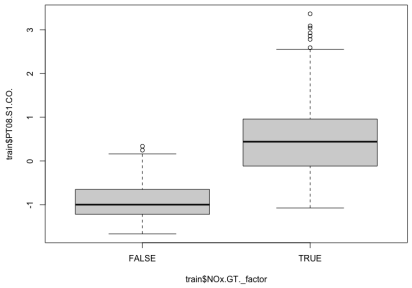

Figure 8: NOx w.r.t S1.CO

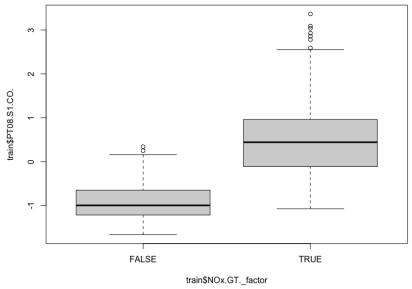

Figure 9: NOx w.r.t C6H6.GT

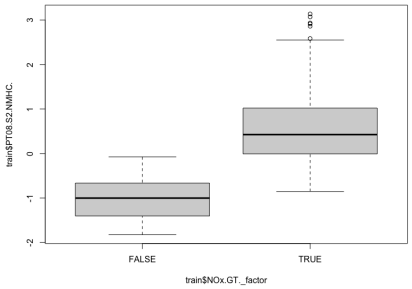

Figure 10: NOx w.r.t S2.NMHC

It's clear that a lot of our variables can be assessed as explanatory for NOx concentration level. By analysing a bit our variables we can see that pollutants presence in the air is linked to the ambient air temperature. Moreover the sources of pollution are linked, meaning they operate at the same time, as factories and trucks. Also same source could be releasing different toxic molecules like vehicle exhaust at rush hours.

We apply an 80/20 % random split of the data, and in the following sections will fit logistic regression models on the train data as well as do a linear discriminant analysis. After threshold selection using ROC-Curve for both methods, we compare performances using the test data.

2 Logistic Regression

In regards to logistic regression, a normal approach would be to start with a full model, assess its in-sample performance.

We calculate a p-value of the omnibus test corresponding to an $H0$ hypothesis corresponding to all parameters equal to 0 and find out that it's of the order of $2.2 * 10^{-16}$ meaning the null hypothesis can be certainly rejected. In summary we get a 6.38% error rate and the following full model confusion matrix(using a 0.5 threshold).

Confusion Matrix		
	FALSE	TRUE
FALSE	121	13
TRUE	12	246

Then we apply some model complexity penalty criterion like the AIC in order to reduce the number of variables. We could start with the full model and then compare the AIC of models with one less variable at each iteration until we find the variable that could be removed without increasing the AIC and this is the variable taken out of the model at each iteration. The algorithm stops when removing any of the remaining variables does not increase the AIC anymore. The function $stepAIC$ with a 'backward' direction does the job. The algorithm has succeeded in removing 5 of the 10 explanatory variables and we were left with the following model:

$NOx.GT.\_factor \sim NMHC.GT. + NO2.GT. + PT08.S4.NO2. + T + RH$

with the lowest AIC of 120.88 while keeping the same confusion matrix and misclassification percentage as the full model.

See in figure 11 the fitted classification and how the number of exact fit is high and the uncertain zone is very narrow (both for full and reduced model, the difference is negligible)

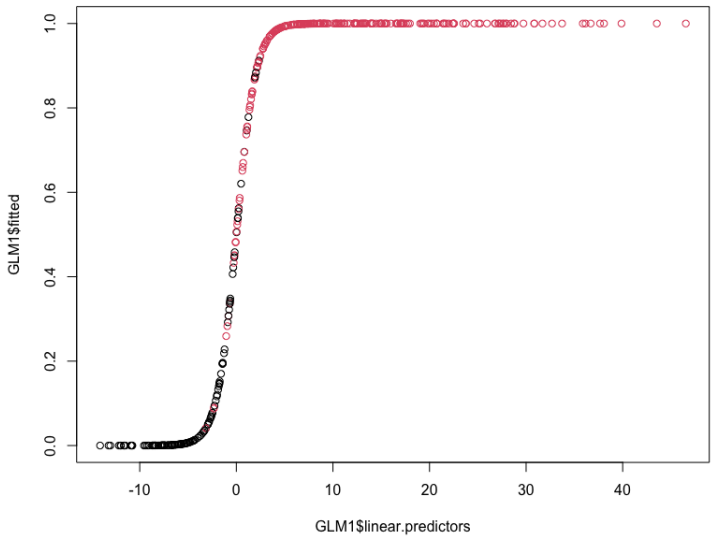

Figure 11: fitted probabilities of train data for Reduced model

Here are the coefficients of each explanatory variable:

Intercept	NMHC.GT.	NO2.GT.	PT08.S4.NO2.	T	RH
5.3261	2.7656	3.2684	5.6415	-2.580653	-2.0977

Figure 12: Coefficients of the optimized model variables

We can see that the other pollutants sensors affect positively the NOx likelihood while Temperature and humidity affect negatively. The importance factor is of same magnitude for all parameters.

Now in order to insure stability of the model coefficients, we inspect the co-linearity of the variables in figure 13 and we calculate the VIFs of our reduced model's variables.

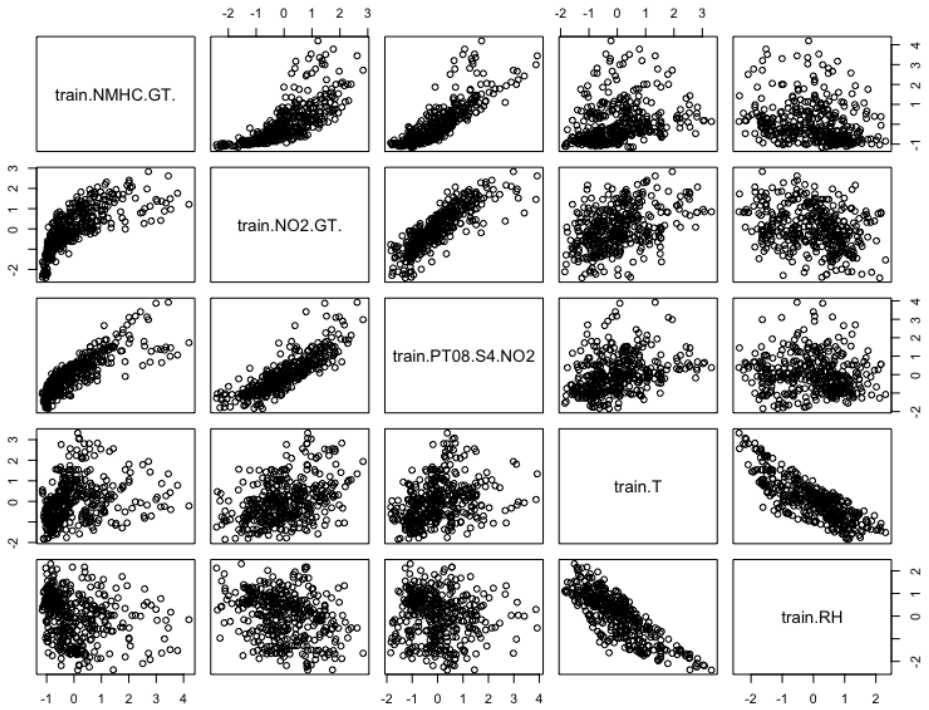

Figure 13: Pairs plot of the reduced model variables

NMHC.GT.	NO2.GT.	PT08.S4.NO2.	T	RH
2.044586	1.027888	4.754754	9.853827	10.907034

Figure 14: Vif values for the optimized model variables

Our Vifs range from 1 to 10.9 so we could consider that they are within the allowable limits that don't inflate the variances of the coefficients too much.

Also it is clear from the pairs plot that the data is mostly scattered randomly and that only a few traces of covariance are apparent in some cases, some case of non-linearity is apparent.

Nevertheless, we have applied a PCA in order to make sure of our model. The code is in the R file, we used the first 4 PCs but we will not publish the results because they are exactly similar to that of the normal logistic regression, same train confusion matrix and error rate.

3 Linear Discriminant Analysis

Since we have a binary classification problem, we only have one canonical variable and that is corresponding to the most discriminating direction. We start by making the model with the entirety of the variable space on the training data and end up with the following results:

Discriminating power = 0.677

	Variable	LD1
	CO.GT.	-0.001417411
	PT08.S1.CO.	0.294018409
	NMHC.GT.	-0.046250353
	C6H6.GT.	-3.947205820
Discriminating Direction:	PT08.S2.NMHC.	5.328638718
	NO2.GT.	0.159894230
	PT08.S4.NO2.	-0.021353357
	PT08.S5.O3.	-0.240639605
	T	-0.056398606
	RH	0.004531233

We can see that NMHC and C6H6 are the main components. In general, discriminating power is quite high and promising.

In figure 16 we can visualize the 1D-Scores, the within variance of the "High NOx" Class is smaller than that of the "Low NOx" class. Also the fringe between both subpopulations is visible but the two subpupolations are distinguishable:

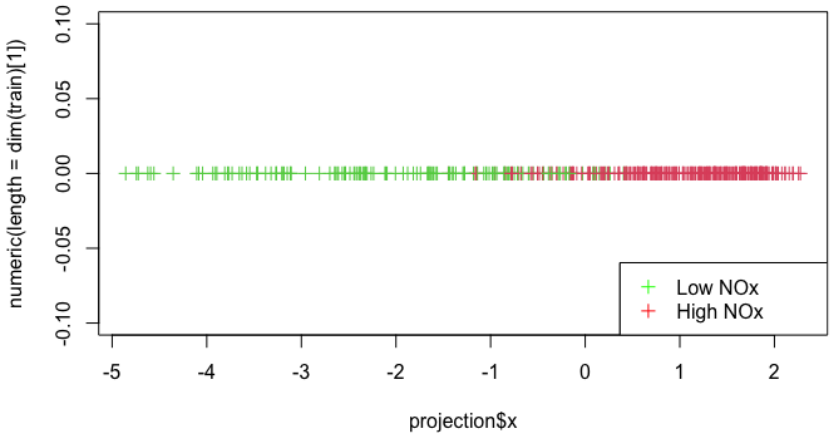

Figure 15: 1-D Scores of full model

We proceed by simplifying the expression of the canonical variable. Indeed some variables did not look “discriminant” when doing the exploratory analysis of Section 3 and they will probably leave the discriminating power intact when removing them one by one in the "leave one variable out" process.

In summary, we did the leave-one-out process many times to be sure, but in the final model, we removed most of the variables in three iterations, first we removed CO.GT., T and RH, then we removed NMHC.GT. and PT08.S4.NO2 and finally we kept "C6H6.GT." and "PT08.S2.NMHC." and removed the rest. The chosen variables of the model are different that those of logistic regression model. From the boxplots, it’s evident that removing RH and T and PT08.S5.O3. would not increase discriminating power since they were already weak discriminant. However, out of the 10 quantitative variables of the full model, only 2 ("C6H6.GT." and "PT08.S2.NMHC.") were kept in order to not decrease the discriminating power. Although many other variables can be regarded as discriminating, our model chose to keep only those 2 maybe because of the redundancy of the information found in the other variables and possible colinearity. This can be managed by applying a dimension reduction via a PCA to eliminate redundancy and colinearity, but due to time constraint we did not go further in our LDA modelling.

We can plot in 2-D those 2 variables with color disctinction of the classes:

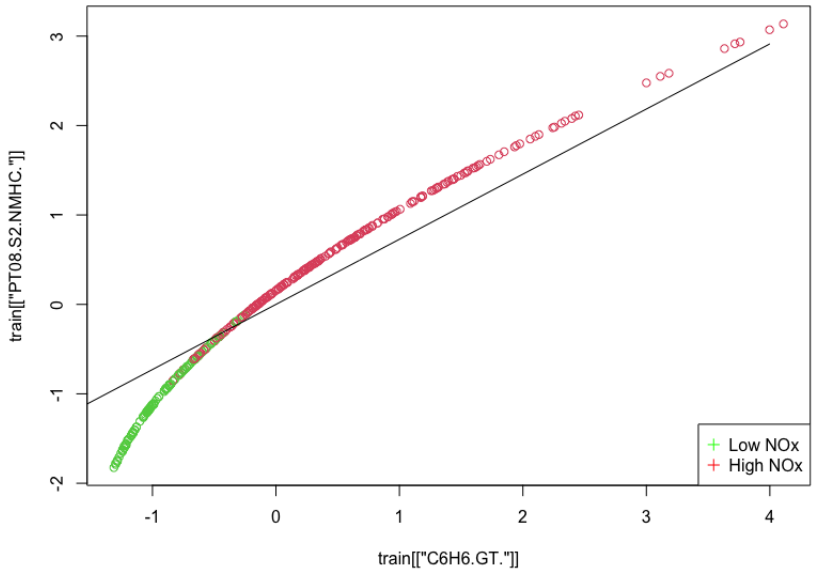

Figure 16: Discriminating Variables with best direction

After reaching this plot in the report, we realised that in fact it's clear why those 2 variables dominate the rest, its because of this perfect relationship between ground truth values and sensor output, which is highly suspecting that this data has been manipulated. Nevertheless it's our data and we will commit to it.

4 ROC CURVE

Using the training data we calculate and plot the ROC curve the reduced models of both methods.

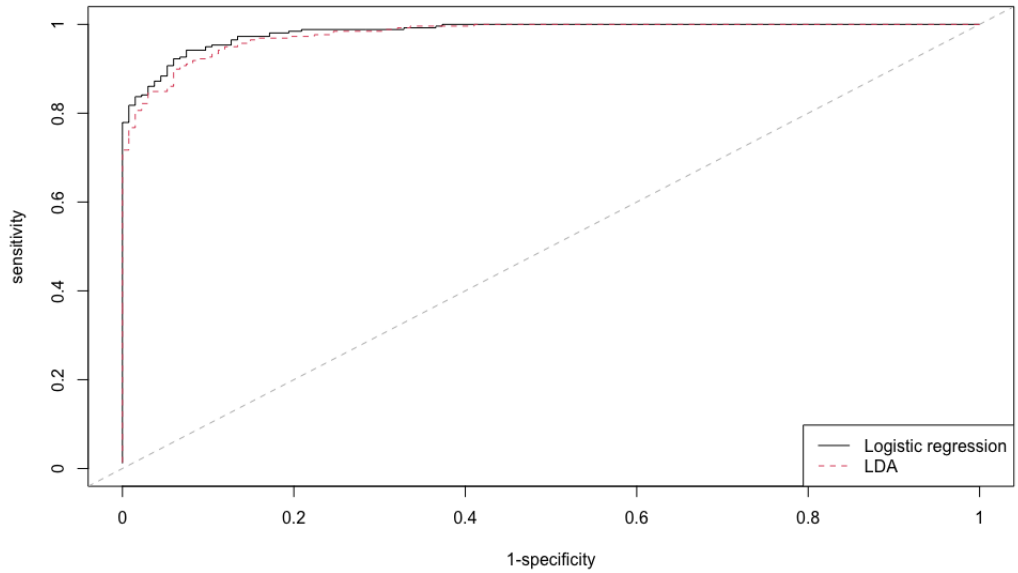

Figure 17: ROC Curve of both final logistic regression and LDA models

We can see that even though each method has chosen different explanatory variables for fitting. The in-sample performance of the models is quite similar and they both show a big area-under-curve. The shape of the ROC CURVE is so ideal due to the number of exact fits in both models. However due to the different choice of parameters we have different predicted probabilities for the observations that lie in the uncertainty area (in between exact fit regions).And this is clear in the following figures:

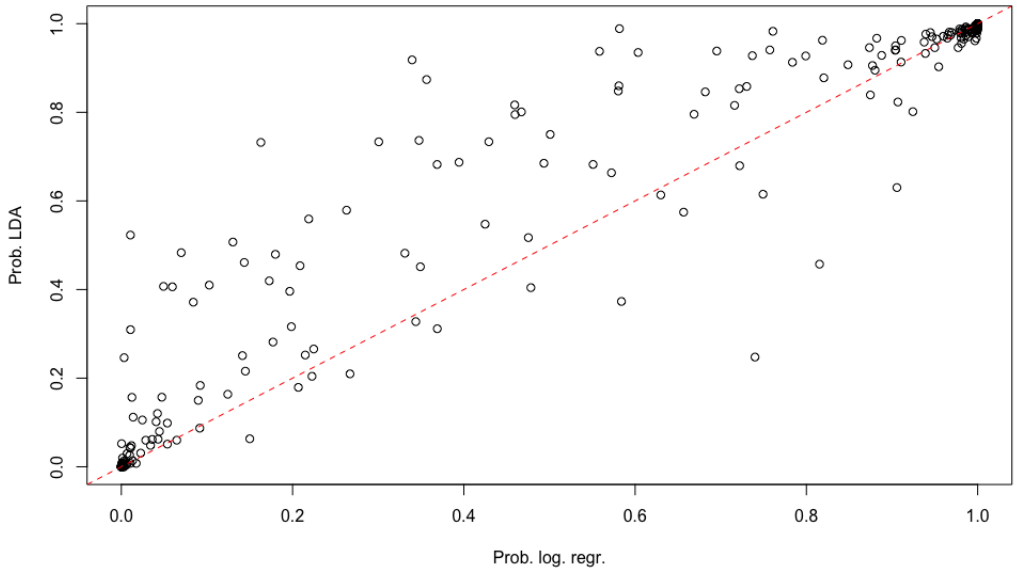

Figure 18: Probability comparison

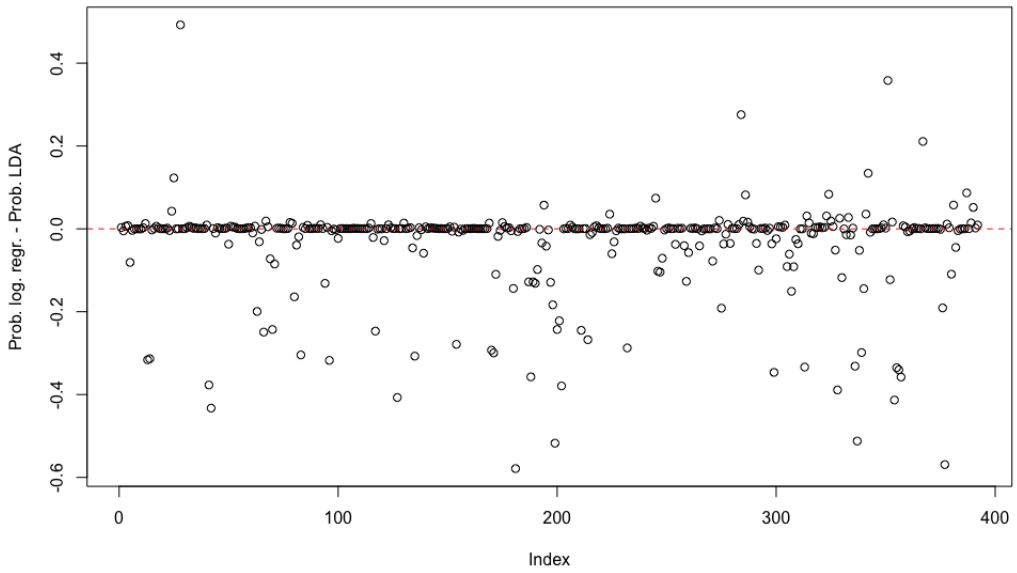

Figure 19: Probability comparison Residuals

Using a 0.5 threshold we can have a look at confusion matrix comparing classifications of both methods:

Confusion Matrix		
	FALSE	TRUE
FALSE	117	18
TRUE	3	254

In order to find the appropriate threshold for each classification technique, we use the minimal distance

between specificity and sensitivity criterion and the Youden's criterion and choose one of them or set a midpoint between them.

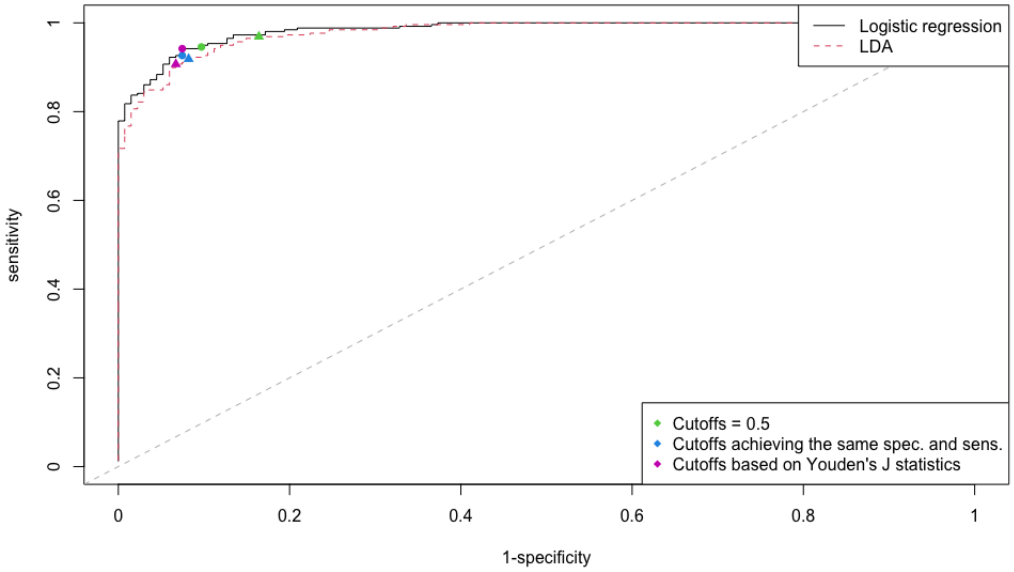

Figure 20: Proposed Cutoffs

In our case, we want to be safe and send out warnings of high pollution even if there is not much pollution in reality. So we are interested in predicting more often "High NOx" (SUCCESS) even if we have a lot of false positives and less false negatives. In classification terms this is translated as accepting a lower specificity for a higher sensitivity. And that's the purple dot(Youden) for logistic regression ($\pi_{logReg} = 0.58$) and the blue triangle(spec $\approx$ sens) for LDA ($\pi_{LDA} = 0.8$).

5 Out-of sample Performance

5.1 Logistic Regression

Test Confusion Matrix		
	FALSE	TRUE
FALSE	38	2
TRUE	4	64

Figure 21: Logistic regression test data confusion matrix

Misclassification error rate = 5.56%
Sensitivity = 0.941
Specificity = 0.95

5.2 Linear Discriminant Analysis:

Test Confusion Matrix		
	FALSE	TRUE
FALSE	39	1
TRUE	6	62

Figure 22: Logistic regression test data confusion matrix

Misclassification error rate = 6.48%
Sensitivity = 0.911
Specificity = 0.975
Out-of sample Errors are quite similar although as explained we prefer the logistic regression since it has lower error rate and higher sensitivity.

Appendix

A Dataset

A.1 Variables list

Our dataset used for the classification problem is (500 observations):

- `CO.GT.` True hourly averaged concentration CO in mg/m^3 (reference analyzer)
- `PT08.S1.CO` (tin oxide) hourly averaged sensor response (nominally CO targeted)
- `NMHC.GT.` True hourly averaged overall Non Metanic HydroCarbons concentration in $\mu g/m^3$ (reference analyzer)
- `C6H6.GT.` True hourly averaged Benzene concentration in $\mu g/m^3$ (reference analyzer)
- `PT08.S2.NMHC.` (titania) hourly averaged sensor response (nominally $NMHC$ targeted)
- `NOx.GT._factor` True hourly averaged NO_x concentration in Parts per billion ppb (Quantitative → binary)
- `NO2.GT.` True hourly averaged NO_2 concentration in $\mu g/m^3$ (reference analyzer)
- `PT08.S4.NO2.` (tungsten oxide) hourly averaged sensor response (nominally NO_2 targeted)
- `PT08.S5.O3.` (indium oxide) hourly averaged sensor response (nominally O_3 targeted)
- `T` in $°C$
- `RH` Relative Humidity (%)

A.2 Statistical summary before normalisation

CO.GT.	PT08.S1.CO.	NMHC.GT.	C6H6.GT.	PT08.S2.NMHC.
Min. :0.300	Min. : 818	Min. : 1.022	Min. : 0.600	Min. : 457.0
1st Qu.:1.200	1st Qu.:1024	1st Qu.: 61.000	1st Qu.: 4.500	1st Qu.: 736.8
Median :2.000	Median :1196	Median :131.500	Median : 8.600	Median : 924.5
Mean :2.252	Mean :1223	Mean :172.732	Mean : 9.943	Mean : 935.4
3rd Qu.:2.900	3rd Qu.:1386	3rd Qu.:237.000	3rd Qu.:13.800	3rd Qu.:1109.5
Max. :8.100	Max. :2040	Max. :797.000	Max. :39.200	Max. :1754.0

NO2.GT.	PT08.S4.NO2.	PT08.S5.O3.	T	RH
Min. : 20.0	Min. :1050	Min. : 341.0	Min. : 6.10	Min. :14.90
1st Qu.: 80.0	1st Qu.:1357	1st Qu.: 761.8	1st Qu.:11.00	1st Qu.:38.40
Median : 99.0	Median :1540	Median : 991.5	Median :14.05	Median :51.95
Mean :101.2	Mean :1571	Mean :1028.3	Mean :14.40	Mean :50.04
3rd Qu.:124.2	3rd Qu.:1730	3rd Qu.:1279.8	3rd Qu.:16.90	3rd Qu.:60.95
Max. :194.0	Max. :2679	Max. :2359.0	Max. :29.30	Max. :83.20

NOx.GT. (Binary)
FALSE:174
TRUE :326

Figure 23: Order of magnitude of the different explanatory variables