FAULTY STEEL PLATES

Classification

Starting from a dataset containing the different properties of faulted steel plates we classify the type of faults; we also want to verify the association between the steel type used in the plates and the type of damage.

Statistical Learning Project by Simone Bosio and Nicolò Braghini. Academic Year 2024/2025, Professor Marica Manisera

OBJECTIVES

- Classify from a dataset of faulty steel plates with different characteristics (both measures and physical properties) the steel plates with a major or minor fault type.
- Studying the relationship between the different independent and the response variable in order to identify the most relevant ones to include in our models.
- Compare the different models in order to identify the most useful models for the task and analyze their performance.

DATASET

Dataset Contents:

This dataset comes from research by Semeion, Research Center of Sciences of Communication. The variables describe the geometric shape of the plates and the faults and some physical properties like reflections and steel type.

1941 observations, 9 predictors, 1 dummy response variable (1 = major faults, 0 = minor faults).

Source: Kaggle

MODELING

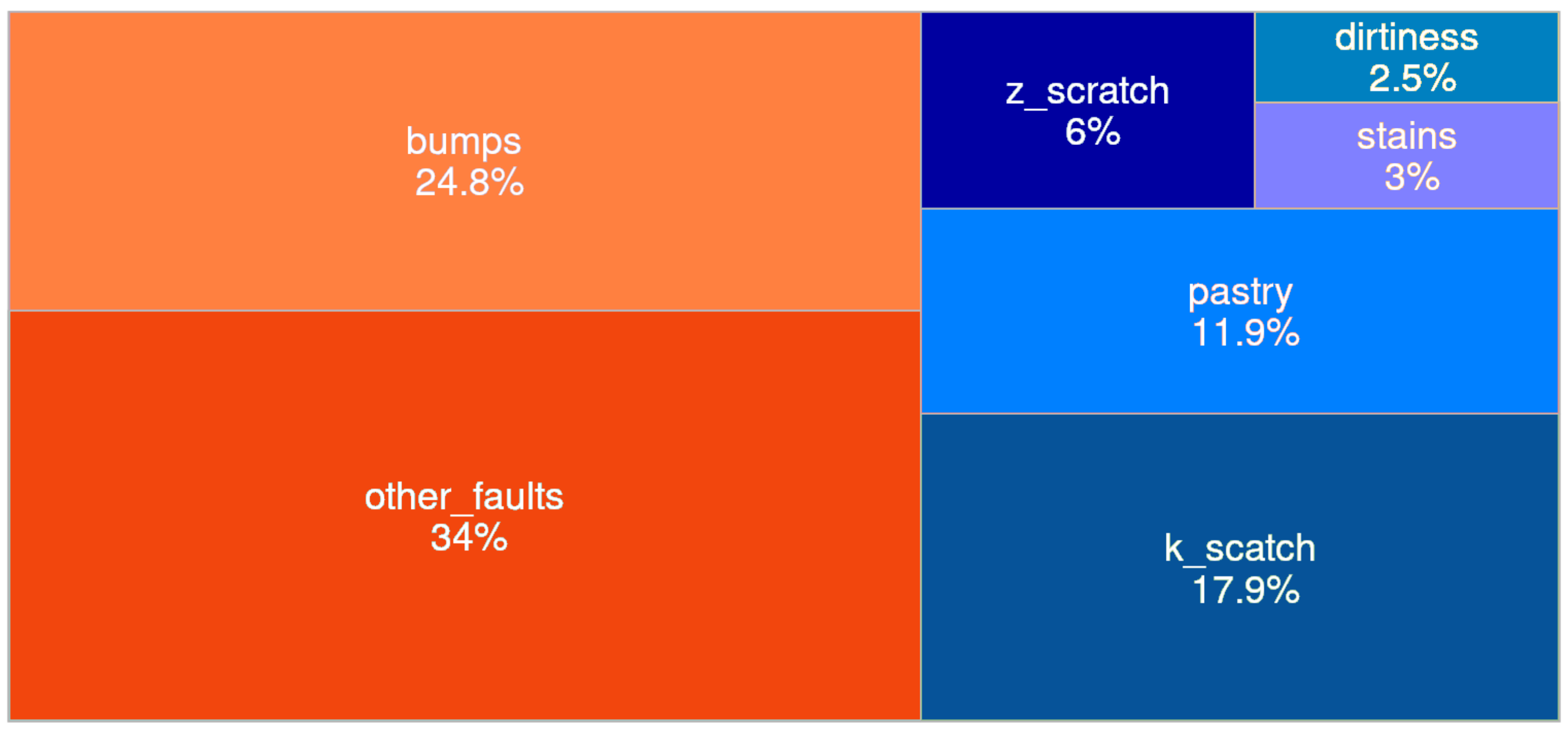
The Classification Models used are:

- Logistic Regression Model (2 variables)
- Logistic Regression Model (4 variables selected through the Stepwise Procedure)
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- Naïve Bayes (NB)
- K-Nearest Neighborhood (KNN)

The Model Diagnostic Tools used are:

- Confusion Matrices
- ROC curves and AUC

FAULT TYPE



Red = Major Fault

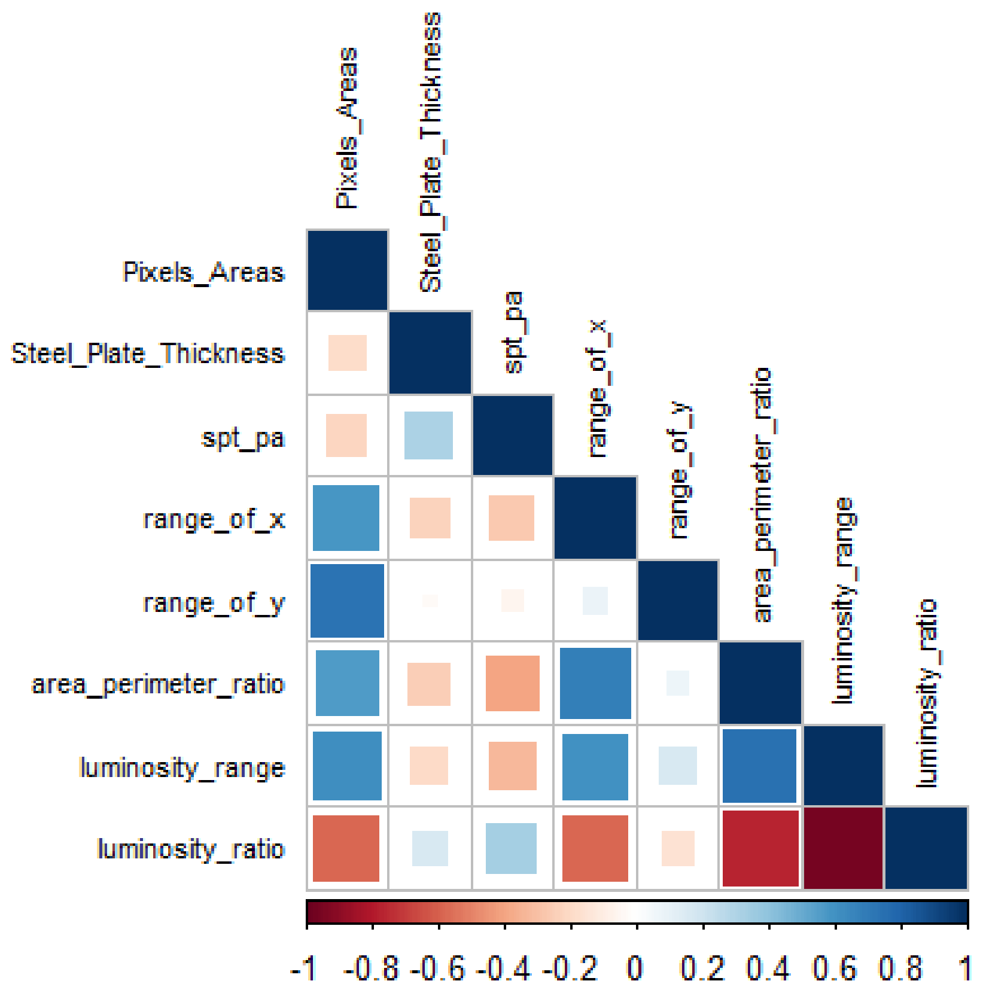
Blue = Minor Fault

In the Dataset we had 7 different kinds of faults, for the sake of simplicity we divided them based on the severity of the faults:

"Scratch", "Dirtiness", "Stains" and "Pastry" were classified as "Minor Faults" because they're all superficial kinds of faults.

"Bumps" and "Other Faults" were classified as "Major Faults" because they modify the structure of the plates.

CORRELATIONS



Before choosing the variables to be used in our models we looked at the correlations. To avoid the multicollinearity issues, we avoided using the variables which are highly correlated to each other in the same models. This is just a preliminary step, then we also calculated the VIF of the chosen models to avoid multicollinearity between 3 or more variables.
From this graph we excluded the qualitative variables "Steel Type" and "Fault Type".

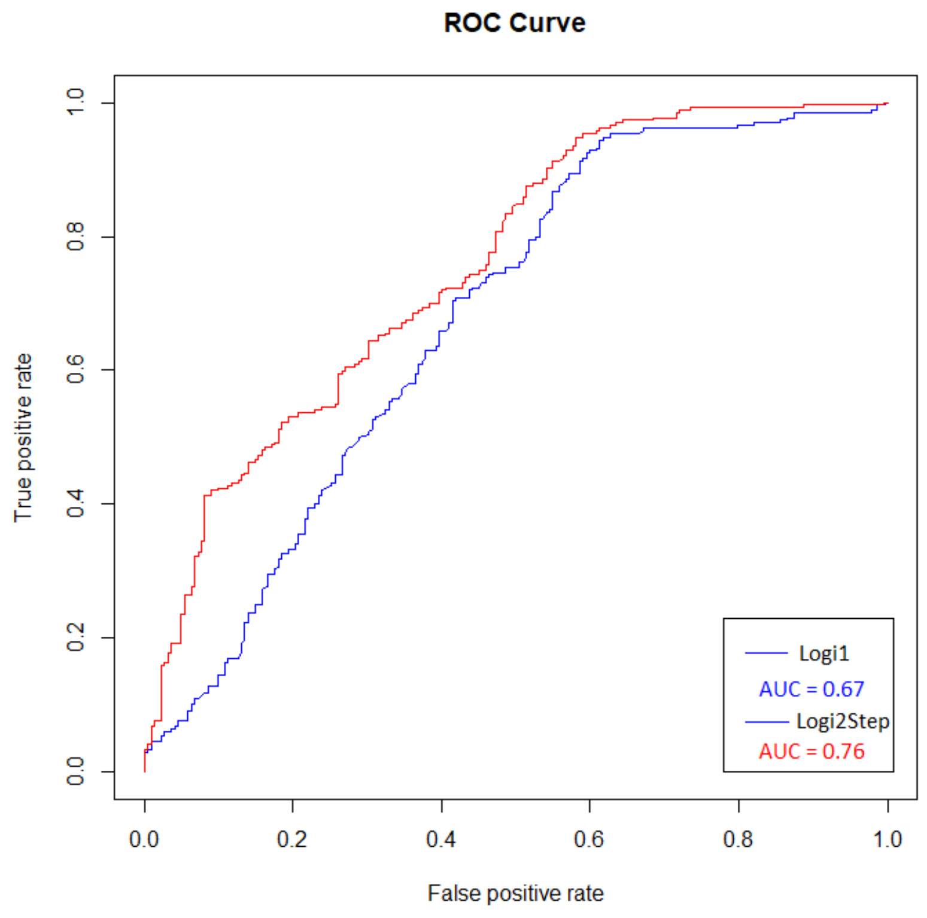
LOGISTIC REGRESSION MODELS

MODELS SUMMARY

	Model1	Model2 (improved via back. stepwise)
(Intercept)	-2.3673***	-0.3251*
	(0.2070)	(0.1373)
luminosity_ratio	3.4439***	
	(0.2929)	
steel_type	0.9022***	0.2713*
	(0.1178)	(0.1279)
Pixels_Areas		-0.0008***
		(0.0001)
Steel_Plate_Thickness		0.0109***
		(0.0015)
range_of_x		0.0136***
		(0.0022)
Num.Obs.	1455	1455

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

ROC CURVES



CONFUSION MATRICES

Logistic 1 (th = 0,5)	0	1	Sum
0	100	122	222
1	42	222	264
Sum	142	344	486

Accuracy	66,26%
Sensitivity	84,09%
Specificity	45,05%
PPV	64,53%
NPV	70,42%

Logistic 2 (th = 0,5)	0	1	Sum
0	93	129	222
1	17	247	264
Sum	110	376	486

Accuracy	**69,96%**
Sensitivity	93,56%
Specificity	41,89%
PPV	65,69%
NPV	84,55%

KNN (th = 0,5, K = 100)	0	1	Sum
0	137	85	222
1	57	207	264
Sum	194	292	486

Accuracy	70,78%
Sensitivity	78,41%
Specificity	61,71%
PPV	70,89%
NPV	70,62%

GENERATIVE MODELS

CONFUSION MATRICES

LDA (th = 0,5)	0	1	Sum
0	91	131	222
1	15	249	264
Sum	106	380	486

Accuracy	69,96%
Sensitivity	94,32%
Specificity	40,99%
PPV	65,53%
NPV	85,85%

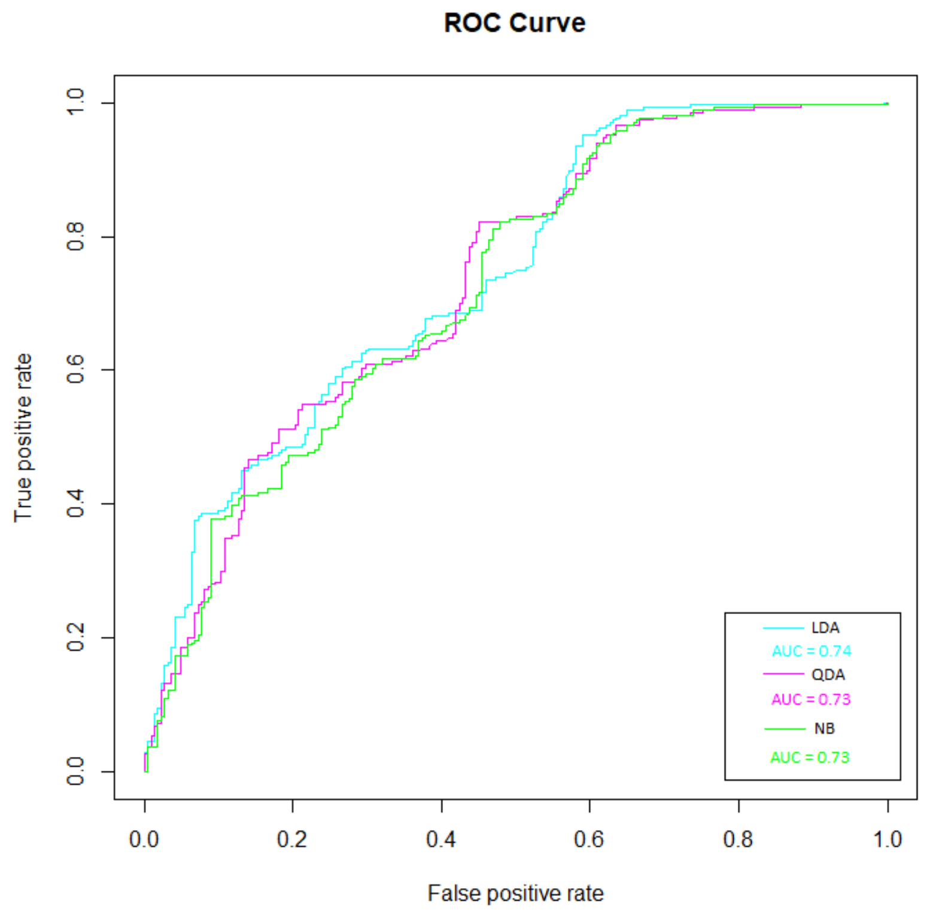
QDA (th = 0,55)	0	1	Sum
0	85	137	222
1	14	250	264
Sum	99	387	486

Accuracy	68,93%
Sensitivity	94,70%
Specificity	38,29%
PPV	64,60%
NPV	85,86%

NB (th = 0,6)	0	1	Sum
0	83	139	222
1	13	251	264
Sum	96	390	486

Accuracy	68,72%
Sensitivity	95,08%
Specificity	37,39%
PPV	64,36%
NPV	86,46%

ROC CURVES



CONCLUSIONS

- All the models have similar performances, so, following the **parsimony principle** we chose as the best one the **Second Logistic Regression.**
- All the predictors used in this model are **statistically significant.**
- We found that using steel type A300 instead of A400 is associated with an increase of the probability of incurring in a major fault.
- The main index is the **accuracy** rather than sensitivity or specificity, since there are no specific downsides in false negative and false positive classifications.
- Since all the models we considered, including the non-parametric KNN model, lead to similar result one further step in this classification would be to use unsupervised statistical learning methods.