Python Lab 2 SSY316

Matthew Newson, Erik Norlin

November 2023

Logistic regression

Logistic regression was implemented in three different ways for the variables *age* and *nodes detected* with an interception. The three different implementations were: from scratch, using the statsmodels module, and using the sklearn module in python. In Table 1, 2, and 3 we can see that the results are identical which indicates that the model done from scratch is accurately implemented.

Table 1: Model properties of the logistic regression model implemented from scratch

model params.	coefficient	se	lower 95%	upper 95%
intercept	0.88	0.86	-0.81	2.57
age	0.01	0.02	-0.02	0.04
nodes detected	-0.10	0.03	-0.15	-0.05

Table 2: Model properties of the logistic regression model using statsmodels

model params.	coefficient	se	lower 95%	upper 95%
intercept	0.88	0.86	-0.81	2.57
age	0.01	0.02	-0.02	0.04
nodes detected	-0.01	0.03	-0.15	-0.05

Table 3: Model properties of the logistic regression model using sklearn. Information about se, lower 95% and upper 95% were unfortunately not presented by the module.

model params.	coefficient
intercept	0.88
age	0.01
nodes detected	-0.10

Logistic regression and Bayesian logistic regression

A comparison between a logistic regression model and a Bayesian logisitic regression model (BLR) was carried out to investigate how the results differ between the two approaches. Both implementations included the parameters *age*, *nodes detected*, and *year*. In table 4 and 5 we can see that the approaches gave similar results, however not

identical. The BLR model gives overall less spread which can indicate that this model is more certain about its parameters.

Table 4: Model properties of the logistic regression model.

model params.	coefficient	se	lower 95%	upper 95%
intercept	2.98	5.36	-7.52	13.48
age	0.01	0.00	0.01	0.01
nodes detected	-0.10	0.00	-0.10	-0.10
year	-0.03	0.00	-0.04	-0.03

Table 5: Model properties of the BLR model.

model params.	coefficient	se	lower 95%	upper 95%
intercept	2.87	3.20	-3.40	9.14
age	0.01	0.02	-0.02	0.04
nodes detected	-0.10	0.03	-0.15	-0.05
year	-0.03	0.05	-0.13	0.07

Performance evaluation of logistic regression and Bayesian logistic regression

The performances of logistic regression and BLR were evaluated for two different models. The first model included the variables *age* and *nodes detected*, and the second model included the variables *age*, *nodes detected* and *year*. Both models also included an interception. In Table 6 we can observe the performance results for both models, in Figure 1 and 2 we can see the ROC curves for both models, and in Figure 3 and 4 we can see the predicted probabilities of logistic regression and BLR for both models.

a) Comparing model 1 and 2

When evaluating performance of a model, it is generally preferable to have larger log-evidence, smaller BIC, smaller log-loss, and larger AUC. From Table 6 we can see that the log-evidence is better for model 2, BIC is better for model 2, log-loss is better for model 1, and AUC is better for model 1.

Table 6: Performance results for model 1 and 2.

	model 1	model 2
log-evidence	-119.43	-118.30
BIC	-116.79	-119.23
log-loss BLR	58.25	58.82
log-loss LR	58.66	59.44
AUC BLR	0.59	0.58
AUC LR	0.59	0.58

In Figure 1 and 2 we can see that the ROC curves are below the reference line up to when the specificity is about 0.25-0.3 which is not preferable. The ROC curves are above the reference line onwards. The ROC curves for logistic regression and BLR follow each other closely which aligns with that the AUCs for logistic regression and BLR are approximately the same.

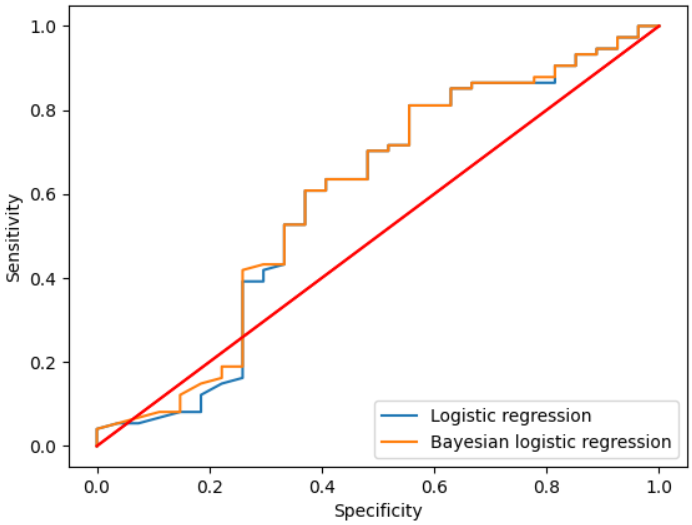

Figure 1: ROC curve of model 1.

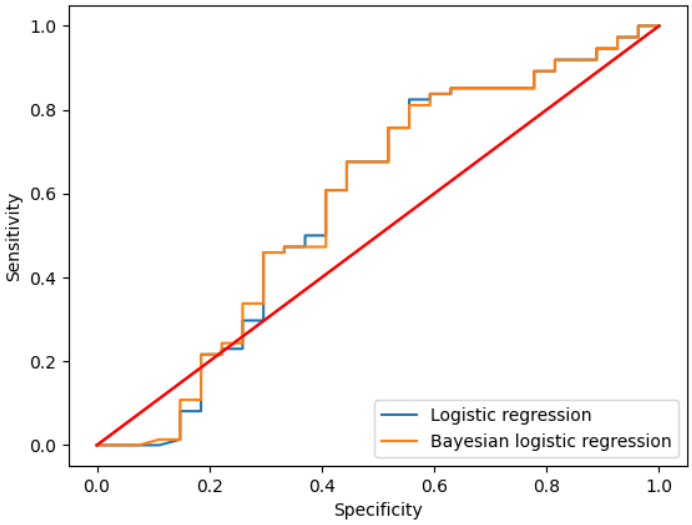

Figure 2: ROC curve of model 2.

In conclusion, no model really outperforms the other, the performances are marginally different and they perform better than the other for different evaluation metrics.

b) Differences in predicted probabilities between logistic regression and Bayesian logistic regression.

We can observe in Figure 3 and 4 that the predicted probabilities for both approaches follow a similar pattern. However, the predicted probabilities for logistic regression are more spread out towards 1 and 0 than the predicted probabilities of the Bayesian models, as if logistic regression seems more certain about its predictions. This could

indicate that the logistic regression is better at prediction in this case, but this property might also contribute to overfitting. We can see that AUCs are the same for both approaches but the log-loss is smaller for the Bayesian models which indicates that the Bayesian models perform better.

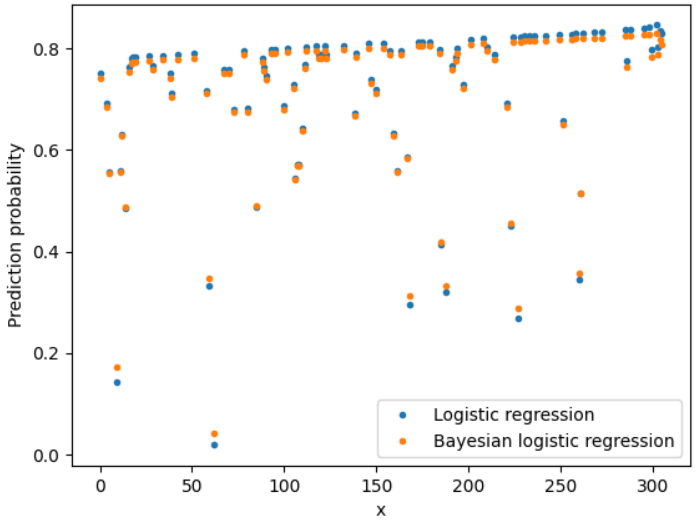

Figure 3: Predicted probabilities of model 1.

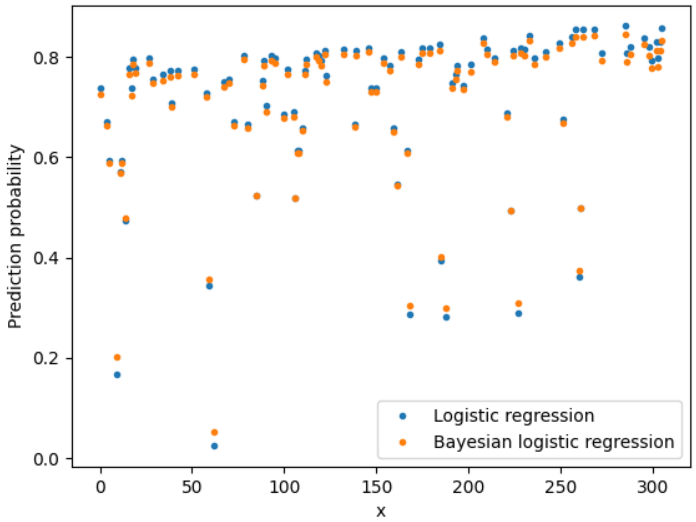

Figure 4: Predicted probabilities of model 2.

Logistic regression and linear discriminant analysis

Two models were fitted, logistic regression and linear discriminant analysis (LDA), to the data set with variables *age*, *nodes detected* and *year* to evaluate their predictive performance. An intercept was also considered for logistic

regression. We can see in Table 7 that logistic regression gives smaller log-loss than LDA, but LDA gives larger AUC. In Figure 5 we can see that the ROC curves does not follow each other as closely as logistic regression and BLR did. Up to around specificity=0.3 the ROC curves are below the reference and above the reference onwards, as before. In Figure 6 we can see that both models display a similar pattern in the predicted probabilities, but LDA shows even more spread than logistic regression, making LDA possibly even more prone to over fitting. From this we could make an educated guess that a Bayesian model would perform better than LDA since it performed better than logistic regression.

Table 7: Performance results for logistic regression and LDA.

	LR	LDA
log-loss	59.43	60.12
AUC	0.58	0.59

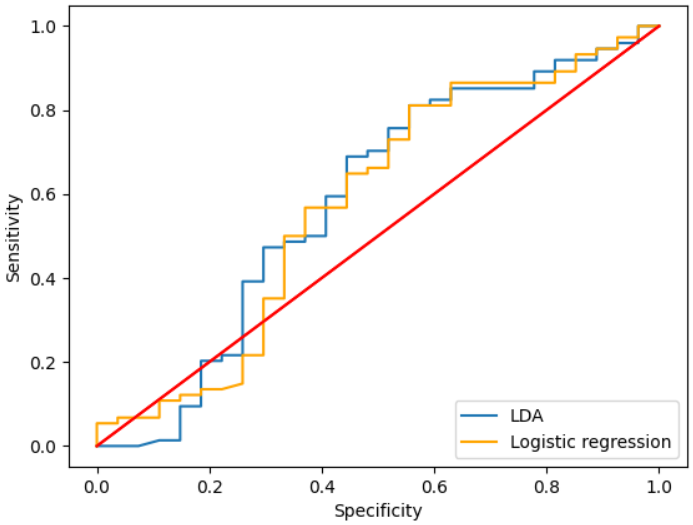

Figure 5: ROC curve of logistic regression and LDA.

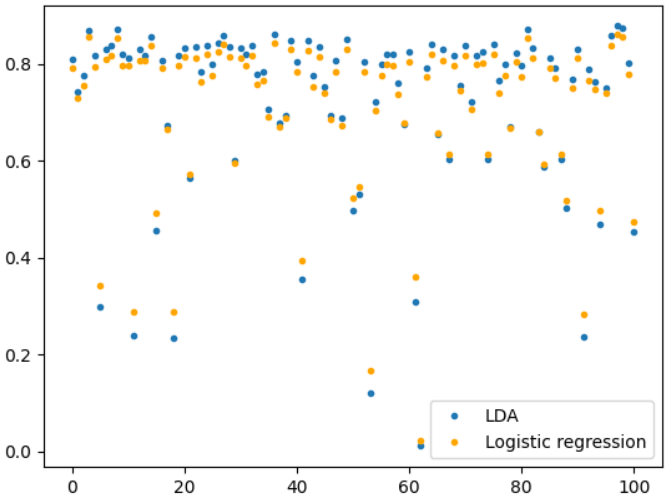

Figure 6: Predicted probabilities of logistic regression and LDA.