HW2: Evaluating Binary Classifiers and Implementing Logistic Regression

Vedant Modi

COMP135: Introduction to Machine Learning

Spring 2025, Tufts University

February 20, 2025

Problem 1

	train	valid	test
num. total examples	390.000	180.000	180.000
num. positive examples	55.000	25.000	25.000
fraction of positive examples	0.141	0.139	0.139

Table 1: Total count, number of positive examples (cancer), and fraction positive out of total. Tabulated per each set.

1a Short answer

The `predict-0-always` classifier reports an accuracy of 0.861 on the test set. This classifier is not good enough for our screening task since deploying this classifier would result in many missed cancer diagnoses (i.e. false negatives). Since we always predict cancer is not present, the model is useless for classifying whether or not a patient has cancer.

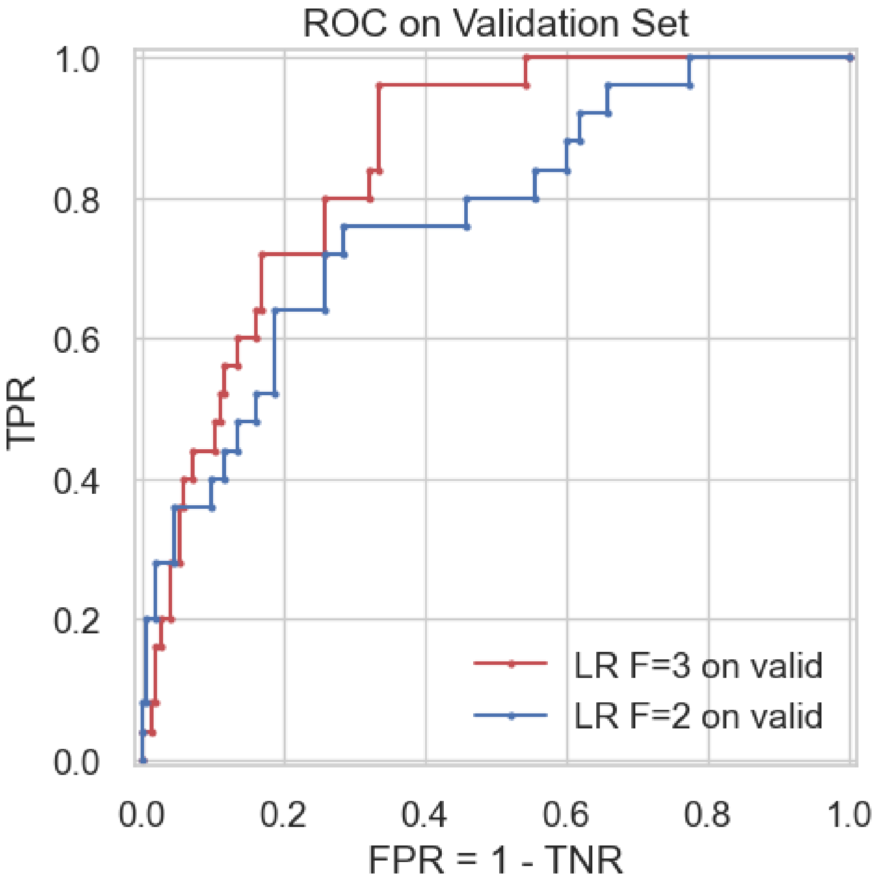

Figure 1: ROC curve with $F = 3, F = 2$

1b Short answer

At very high thresholds (high probability needed for cancer), we see the model with 2 features report a higher ROC curve, and therefore higher performance. For lower thresholds, we see the model with 3 features overtake the model with 2 features in performance.

	F=3 Logistic Regression with default threshold	F=3 Logistic Regression with best threshold
Threshold used, τ	0.500	0.075
Confusion Matrix	Predicted 0 1 True 0 149 6 1 17 8	Predicted 0 1 True 0 92 63 1 1 24
TPR, TNR on test set	TPR = 0.320, TNR = 0.961	TPR = 0.960, TNR = 0.594

Table 2: Test-set performance for various Logistic Regression models

1c Short answer

(i) If every patient in the test set would have had a biopsy in current practice, then all 180 patients would have had a biopsy. For any of the models, if the model predicts 0 (i.e. no cancer), then a biopsy is not considered necessary according to the model.

With the model selecting with the default threshold, we would see 149 + 17 = **166 patients** who have been predicted to not have cancer. So, we do not perform a biopsy on **166 patients** in this situation.

With the model selecting with the best threshold, we would see 92 + 1 = **93 patients** who have been predicted to not have cancer. So, we do not perform a biopsy on **93 patients** in this situation.

(ii) A good patient outcome is where we skipped a biopsy (predicted 0) and did not have cancer (true 0). This is a good outcome since the invasive procedure didn't have to happen, and we predicted that to be true. Suppose we didn't predict that to be true, then the biopsy would be for waste.

For the model selecting with the default threshold, we have that **149 patients** have a good outcome.

For the model selecting with the best threshold, we have that **92 patients** have a good outcome.

1d Short answer

The best model for the doctors' goals is the logistic regression selecting with the best threshold (second model).

The first goal states that we don't want to send away a patient that actually had cancer. To avoid this, we want to reduce the amount of false negatives, and increase the amount of true positives. This corresponds directly to a higher $TPR = \frac{TP}{TP+FN}$ since false negatives decreasing, and true positives increasing both make the TPR increase. Note that the TPR is better for the logistic regression selecting with the best threshold (second model) than the logistic regression selecting with the default threshold (first model). So, the second model achieves the first goal better than the first model.

Per the second goal, an unnecessary biopsy is when a patient is predicted to have cancer but does not end up being sick. This is a false positive in our classification system and a true 0 and predicted 1 in the confusion matrix. The model with the default threshold definitely has less patients with an unnecessary biopsy than the model with the best threshold. So, the default threshold model achieves the second goal for the doctors' better than the best threshold model.

The default threshold model is still not better than the best threshold model. This is because in this situtation, less false negatives is far more important than less false positives. If we have more false negatives, we risk sending those who truly need treatment away. If we have more false positives, it's true that some procedures might be unnecessary, but we'd rather have less false negatives to ensure no one who truly has cancer does not get treated.

For this reason, the second model, the logistic regression selecting with the best threshold, is the best model that achieves the doctors' goal.

1e Short answer

Predicted	0	1
True		
0	511.1	5.6
1	350.0	133.3

Figure 2: Confusion matrix for 1000 samples of similar composition to test set using the Logistic Regression model with the best threshold

Note that the amount of biopsies that would be performed are any patients where we predict 1, since our model suspects they have cancer. Also note that a life threatening mistake is whenever we predict 0 for a patient, but truly have 1 for that patient.

So, in the confusion matrix above, we see that we would perform 5.6 + 133.3 = **138.9 biopsies**, in total. We would make 350.0 total life threatening mistakes (predicted 0, true 1).