Evaluating Binary Classifiers and Implementing Logistic Regression

Cooper Golemme

February 21, 2025

1 Binary Classifier for Cancer-Risk Screening

We will begin building our classifier for cancer-risk screening by loading the provided data. Below is some useful information about the training, validation and testing datasets.

	train	valid	test
num. total examples	390	180	180
num. pos. examples	55	25	25
fraction of pos. examples	0.141	0.139	0.139

Table 1: Table showing the total examples, examples where cancer has been identified and the fraction of cancer examples in the training, validation, and testing datasets.

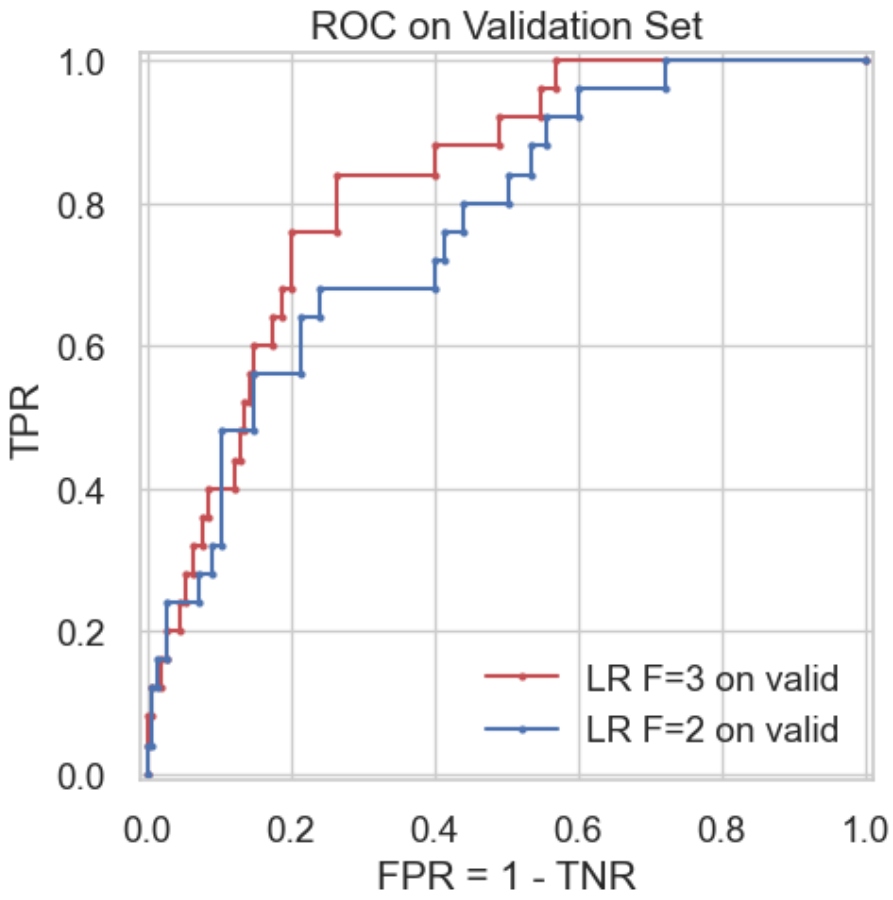
We will continue by offering a baseline prediction for cancer-risk. Our baseline will be always 0, indicating, no cancer.

1.1 Baseline

As is the case when creating any model, we want to create a baseline model to compare our future designs to. In this case, we choose to guess 0 for every data point.

This prediction achieves 86.11% accuracy on the testing data. This makes sense because we guess 0 for every data point and from Table 1, we know 13.9% of the testing set has cancer. The remaining 86.11% do not have cancer and our model correctly predicts that these patients do not.

The baseline model is not good enough because it fails to give biopsies to 100% of the patients who ended up having cancer. The main goal of our classifier is to make sure that people who ended up having cancer will get a biopsy. The secondary goal of our classifier is to reduce the number of biopsies we preform. It is incredibly harmful for a classifier to have 100% of the people who have cancer will be sent home without a biopsy. It is much better for a model to



have a high false positive rate – where the person does not have cancer but we thought they did – as opposed to a high false negative rate. In this case, we give a person who doesn't need it, a minor invasive biopsy, instead of sending home someone who has cancer. For all of these reasons, this baseline model is not good enough. Additionally, this model takes in no data to make its prediction, which is a cause for concern. How could the model provide any insights without taking in any data? Clearly it cannot. A good model will take in the data to generate a meaningful classifier that prioritizes giving people biopsies who could potentially have cancer.

1.2 Logistic Regression

1.2.1 Comparing Models with ROC Analysis

From Figure 1.2.1, we can see that one model does not dominate the other overall in terms of performance across all thresholds. If one model out preformed the other across all thresholds, the lines of the ROC graph would not intersect. From Figure 1.2.1, we see the lines intersect. The lines cross multiple times, indicating that for some threshold values, the 3 feature model is better and for others the 2 feature model is better. Overall, the 3 feature model outperforms the 2 feature model more often across different threshold values, but does not dominate the 2 feature model for every threshold.

1.3 Selecting Decision Threshold

Threshold	0.5	0.0612
Confusion Matrix	146 9 14 11	65 90 0 25
TPR	0.440	1
TNR	0.942	0.419

Table 2: Caption

1.3.1 Results

Under the current strategy, all patients have the biopsy. The results of the biopsy then determine if the patient has cancer or not. The goal for this model is to be able to save some people from having to get a biopsy, when they likely will not have cancer. We know from observing the data that in the test set, there are 180 patients, who got a biopsy and 25 of which had cancer.

For the 0.5 threshold, out of the 180 patients, we told 160 people not to get a biopsy. Out of those 160 people we told not to get a biopsy 146 did not end up having cancer and 14 did. We saved 160 biopsies from being preformed, and in 146 of those cases, it resulted in good patient outcomes in 14 and it resulted in bad patient outcomes.

For the 0.0612 threshold, out of the 180 patients, we told a much lower 65 to not get a biopsy. Out of these 65 people to whom we told not to get a biopsy, all 65 ended up not having cancer. We saved 65 people from getting a needless biopsy, which is less than the 146 needless biopsies we saved with the 0.5 threshold, but importantly, not one of the 65 in this group ended up having cancer, while 14 did in with the 0.5 threshold. Out of these 65 skipped biopsies, all 65 resulted in good patient outcomes and none did not.

1.3.2 Which threshold best meets our goals?

If we examine the goals of this classifier, we want to

1. whenever possible, avoid providing no biopsy to patients that are truly sick

2. also try to eliminate unnecessary biopsies

If we were to choose a threshold strategy that best aligns with this goal from Table 2, we would likely choose the 0.0612 threshold because it prioritizes goal 1 and makes a reasonable effort on goal 2. As outlined in the previous section, on the testing data, we fully accomplished goal 1; we did not withhold biopsy on any patient who actually had cancer. This is the main goal of this classifier. Above all else, we want to make sure people who are actually sick get biopsies. Secondarily, we saved 65 people from getting needless biopsy, which is good for the second goal. While the 0.5 threshold saved more people from getting a needless biopsy – 146 people to be exact – 14 people who had cancer went without a biopsy, which does not align with goal 1. Therefore, if we were to choose which threshold value is better, we would likely say 0.0612 rather than 0.5 because it aligns more closely with the goals of the classifier.

1.3.3 Generalization

If we were to apply the chosen strategy to a new sample of 1000 patients, assuming that the sample was of similar construction to the test set used, we would perform 639 biopsies, compared to the 1000 we would have preformed, and would make 0 life-threatening mistakes. On the testing set, we achieved a 100% true positive rate, which is why, if the new 1000 patients data is similar to the test data, we can expect to make 0 life threatening mistakes.