Erin Sarlak

HW2: Evaluating Binary Classifiers and Implementing Logistic Regression

Problem 1: Binarry Classifier for Cancer-Risk Screening

Table 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

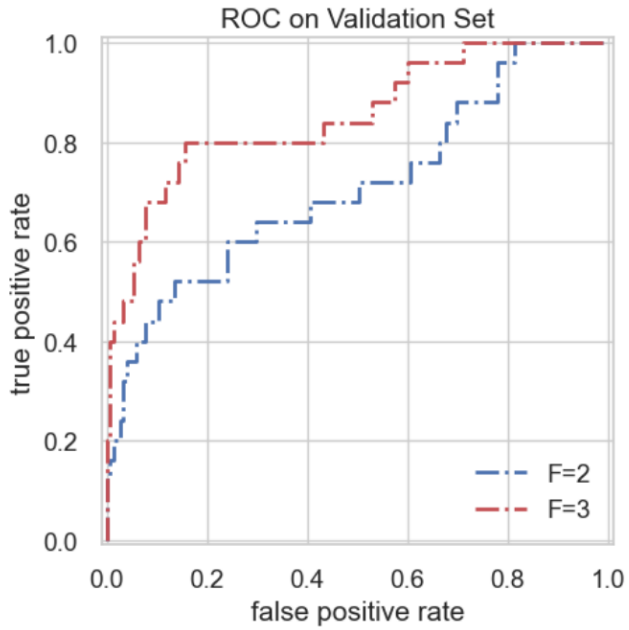
Short Answer 1a:

Accuracy = (True Positives + True Negatives) / Total Number of Instances

= (0 + 155) / 180 = 0.861

Therefore, I calculated the accuracy of “predict-0-always” classifier on the test set to be 0.861, which relatively a high correct classification rate. However, this classifier cannot be considered good enough because the reason that it did well is the imbalance in the data points, which happens when the vast majority of the data points belong to a particular class, in this case, y=0. This implies that this classifier would do a very poor job on classification for a set of datapoints that is balanced, because it is essentially not processing the likelihood of their corresponding classes (predicting), but just assigning 0.

Figure 1:



Short Answer 1b:

The ROC curve for 3-parameter model is consistently higher on the graph across all thresholds. This indicates that the 3-parameter model is better at identifying positive cases for cancer-risk patients regardless of which threshold is chosen. For this problem, a false negative prediction is very dangerous because failing to identify positive cancer patient can cause a patient to not to receive the necessary treatment. For this reason, the model to be used for this problem should prioritize a very high value for the true positive rate in its predictions. I recommend 3-parameter model because it is more successful than 2-parameter model in this aspect.

Figure 2:

	default thr 0.500	
Predicted	0	1
True		
0	152	3
1	15	10
TPR	0.4000	
PPV	0.7692	

	Max TPR s.t. PPV > 0.98 thr 0.6311	
Predicted	0	1
True		
0	155	0
1	20	5
TPR	0.2000	
PPV	1.0000	

	Max PPV s.t. TPR > 0.98 thr 0.0296	
Predicted	0	1
True		
0	57	98
1	0	25
TPR	1.0000	
PPV	0.1852	

Short Answer 1c:

- Using the default threshold of 0.500, the model correctly classified 152 out 155 false patients as false. This means that the vast majority of all the patients that who didn't need the biopsy were saved from an unneeded one.
- For the threshold that maximizes TPR and with a PPV > 0.98, the model correctly identified all 155 patients as false, and therefore no patient received an unnecessary biopsy.
- When the threshold with maximum PPV for TPR > 0.98, the model correctly identified only 57 patients to be false out of all 155. For this reason, only 57 people were saved from an unnecessary biopsy and the other 98 patients endured an unneeded biopsy.

Erin Sarlak

Short Answer 1d:

In the medical context, it is very important to correctly identify true patients who need treatment for their survival. However, when viewed from a hospital's perspective, it's also crucial to minimize the number of unnecessary biopsies in order to reduce the discomfort of the patients and financial costs of performing a biopsy. Nevertheless, the primary concern is ensuring patient survival, and this takes precedence over the other considerations such as financial costs and patient's discomfort due to a biopsy. Therefore, I am for selecting the threshold that maximizes PPV while ensuring that the TPR remains higher than 0.98 for this classification model. Th reason is that the model when using this threshold correctly identifies all patients requiring biopsy, which maximizes their chances of survival. However, it's worth to note that the model at this threshold lead to 98 patient to be incorrectly identified as true, which results in them receiving unnecessary biopsies. PPV value quantifies the proportion of patients with the condition among all those identified as true positives, and tells us what fraction of current biopsies were necessary. The PPV is relatively low at 0.1852, but this threshold at least guarantees that all patients who require a biopsy receives one.