Homework 2

Matthew Carey

Problem 1:

The following table looks at the total sample size, as well as the number of positive subjects and their ratio to the total set for the training, validation, and test data sets:

```
                                  train  valid   test
num. total examples               390.0  180.0  180.0
num. positive examples            195.0   90.0   90.0
fraction of positive examples       0.5    0.5    0.5
```

a) Given this data set, if we create a baseline model predictor that predicts a value of 0 (does not have cancer) for every single entry, we will obtain an accuracy metric of 0.861. This clearly demonstrates that accuracy is not the only metric worth considering, as even though this model seems fairly accurate, this predictor will miss every single positive case, leaving people undiagnosed and not seeking treatment.

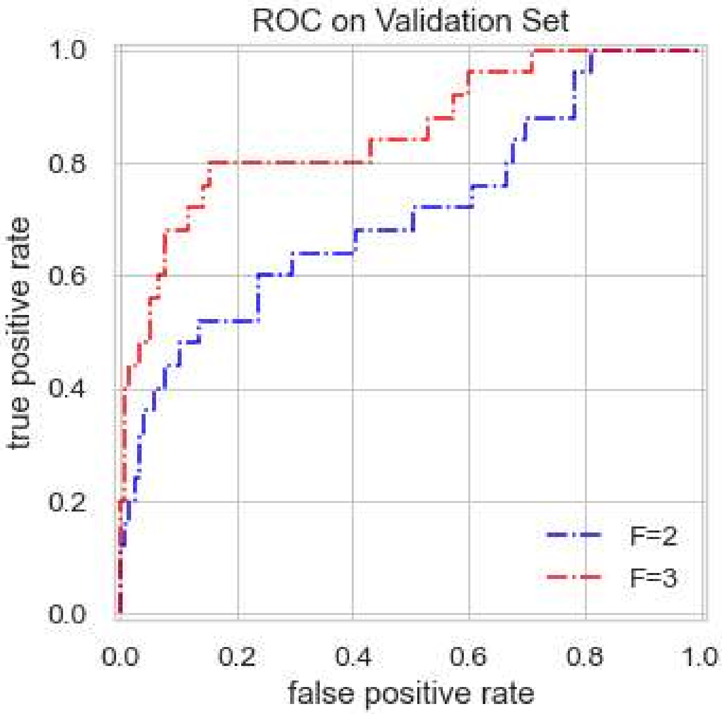
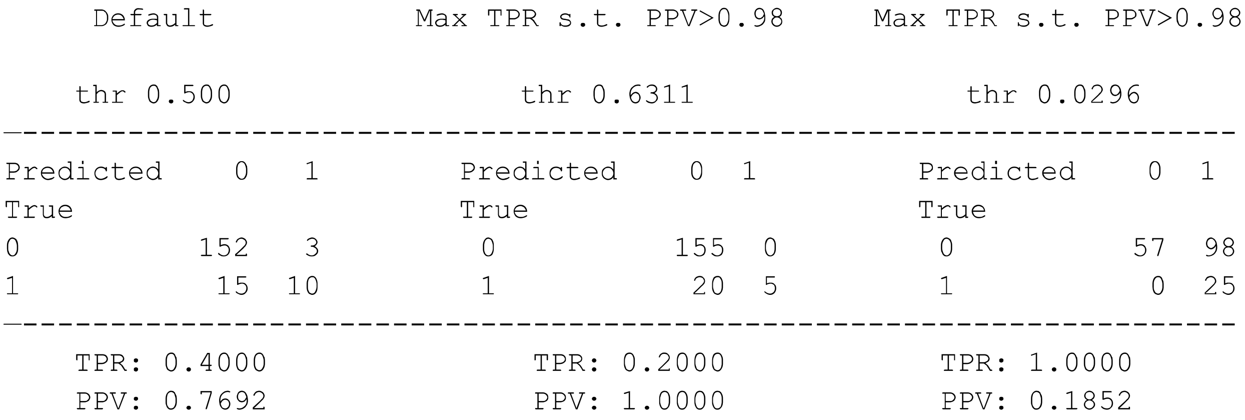


Figure 1

b) In this ROC curve, we can see that the 3 feature dataset outperforms the 2 feature data set across every threshold. We can see this because the 3 feature data set has a greater area under the ROC curve. Thus, I would recommend that the 3 feature data set is used.



```
    Default             Max TPR s.t. PPV>0.98     Max TPR s.t. PPV>0.98

   thr 0.500                 thr 0.6311                thr 0.0296
----------------------------------------------------------------------
Predicted    0   1         Predicted    0  1         Predicted    0  1
True                       True                      True
0           152   3         0          155  0         0           57  98
1            15  10         1           20  5         1            0  25
----------------------------------------------------------------------
    TPR: 0.4000                TPR: 0.2000              TPR: 1.0000
    PPV: 0.7692                PPV: 1.0000              PPV: 0.1852
```

Figure 2

c) Currently, every patient has a biopsy. However, if we apply this model, those who are correctly predicted to be negative will be saved from having an unnecessary biopsy. Thus, with a threshold of 0.5, 152 people will not have to have an unnecessary biopsy, but 15 will go undiagnosed. With a threshold of 0.6311, 155 people will be saved from an unnecessary biopsy, but 20 will go undiagnosed. With a threshold of 0.0296, only 57 people will be saved from an unnecessary biopsy, but no one will go undiagnosed.

d) With the stated goal of avoiding life-threatening mistakes whenever possible, while also eliminating unnecessary biopsies, the best thresholding strategy would be using a threshold of 0.0296. In the other strategies, there is a significant amount of life threatening mistakes where people with cancer are mistakenly called negative and will not get treatment. With a threshold of 0.0296, this hospital will still be able to save 31.6% of people from getting unnecessary biopsies, while ensuring that no one goes undiagnosed. Although it does not decrease unnecessary biopsies as much as other models, it still decreases it from current practice, and does not result in any missed diagnoses, which is far more important to prevent compared to an unnecessary biopsy.