Homework 6

Denis Ostroushko

Model Development

For the purpose of this exercise, we will fit a LASSO regularized regression model and evaluate its predictive power.

For computational simplicity, I will consider a model with no interactions.

It is always good to know the true positive rate and 'rarity' of events in the data, to have a good idea about the model building process and metrics that we need to use to properly assess predictive power of the candidate model.

```
  0   1
102  48
```

The events are not quite rare, but also the outcomes are more skewed towards `no death event` outcomes.

Figure 1 shows how which variables are included in the model as we relax the penalty and how the coefficient change. We expect to have a final model that has a few predictors with large coefficients and some coefficients that are closer to zero.

Additionally, using CV we can pick a parameter λ that will allow us to fit the model that *should* maximize out of sample AUC.

Figure 2 shows the results of CV. As we can see, the value of λ that maximizes AUC is very similar to other considered values. These results suggest that if we pick the lowest value of lambda, and create a more complex model that contains more variables, we still should achieve the same out of sample AUC. We will proceed with the least complicated model, that maximizes OOS AUC.

```
[1] "Log-lambda value to minimize CV AUC: -2.28"
```

```
[1] "Log-lambda value within 1 S.E. of minimizing value: -2"
```

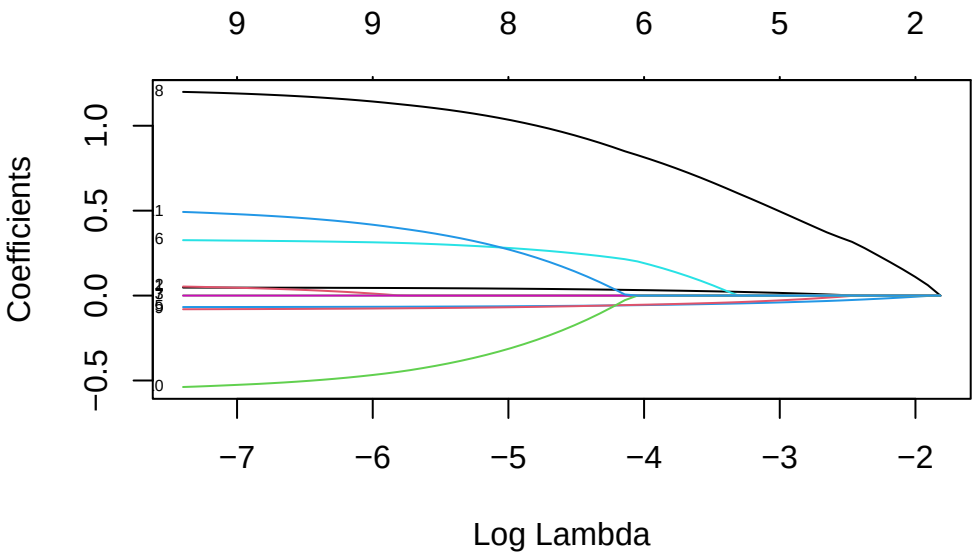

Figure 1: Relationship between Parameter Lambda and the number of variables included in the model

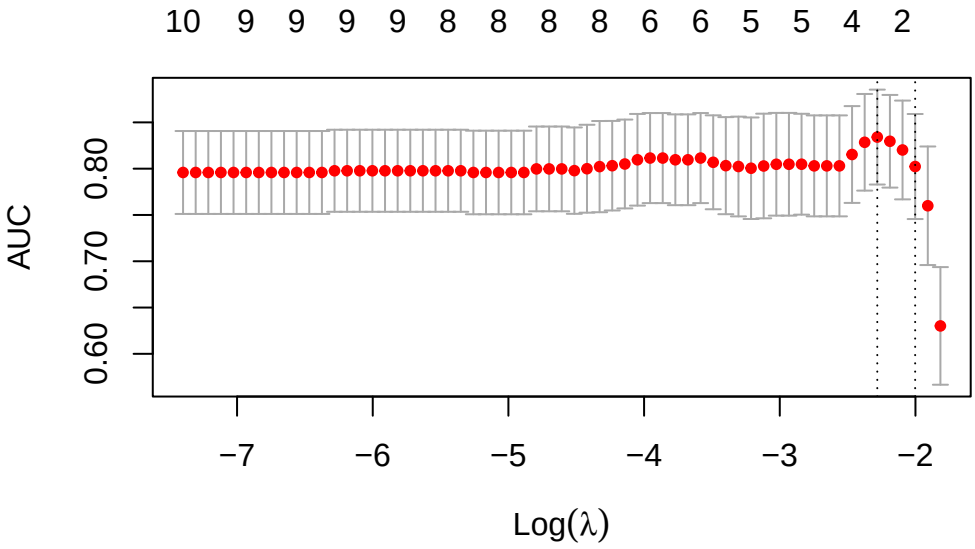

Figure 2: Cross validation of lambda results

Our model will incorporate these variables, their coefficients are also listed below:

(Intercept)	ejection_fraction	serum_creatinine
-0.3752	-0.0191	0.2386

Exercises

(1)

Figure 3 shows In Sample ROC curve. This is a very strongly favorable results, however, we know that In Sample AUC value is always too optimistic. I colored the curve by the cutoff value. We decrease the value of π_0 cutoff as we go along the x-axis from left to right.

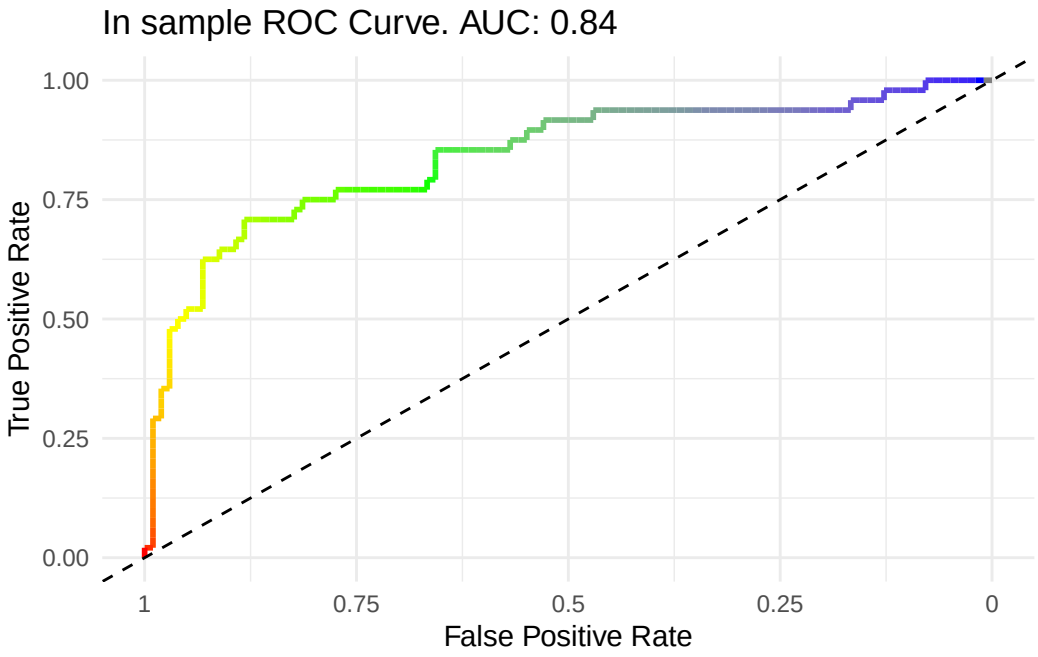

Figure 3: In sample AUC value and ROC curve

(2)

LOOCV does not work for the classification problem because it does not provide adequate information to us. ROC curve summarizes a set of TPR and FPR values. These are proportions, bounded by zero and one. If we use just one observation, then we always will have a value of TPR that is equal to one, and FPR equal to zero, and vice versa. Therefore, we simply would not obtain an ROC curve.

Now we can perform a 10-fold cross validation. This means that at each iterations, approximately, 135 observations will be used to fit the model, and 15 observations will be used to obtain out-of-sample (OOS) AUC value for a set of predictions.

Figure 4 shows the distribution of OOS AUC values for the 10-fold validation. As we can see, there is pretty big variance in the AUC values.

```
      AUC
 Min.   :0.5455
 1st Qu.:0.6461
 Median :0.7906
 Mean   :0.7683
 3rd Qu.:0.8472
 Max.   :1.0000
```

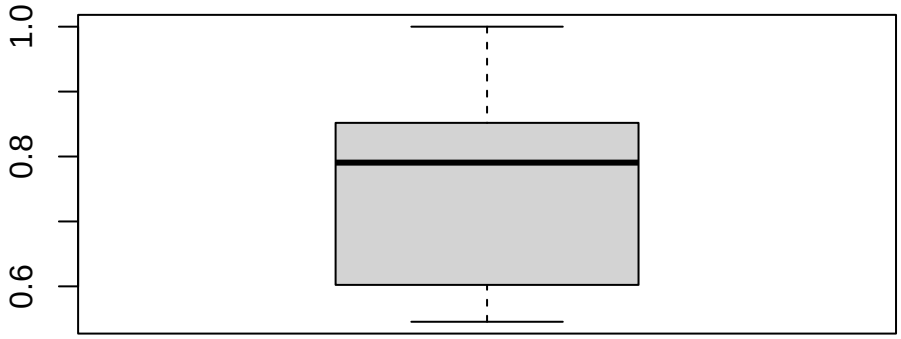

Figure 4: Out-Of-Sample AUC values for the 10-fold validation

(3)

First, we will evaluate the impact of varying K in K-fold cross validation. The values of K I considered are `2, 3, 5, 10, 15`. So, as K gets bigger, the training set gets larger and validation sample gets smaller.

For each value of K I recorded summary of OOS AUC. First, we plot the average OOS AUC for each value of K. Figure 5 shows the results. It seems that the average value fluctuates randomly, but there is no trend. Perhaps, this is just sampling variation.

We can also can investigate how the increase in K affect variance of AUC scores. Figure 6 clearly shows that as K increases, so is the variance in the OOS AUC scores. This is quite interesting, I did not expect this to happen. Perhaps, higher K and corresponding lower validation data set size can produce cases where the event is rare produce values of AUC that are closer to 1, or 0.5, which increase variance.

Impact of Seed on 10 Fold cross validation

Figure 7 and Figure 8 show how mean AUC and variance of AUC scores change when we set a new seed for 10-fold cross validation. There is no strong trend, and the values seem to be fluctuating randomly in a bounded interval.

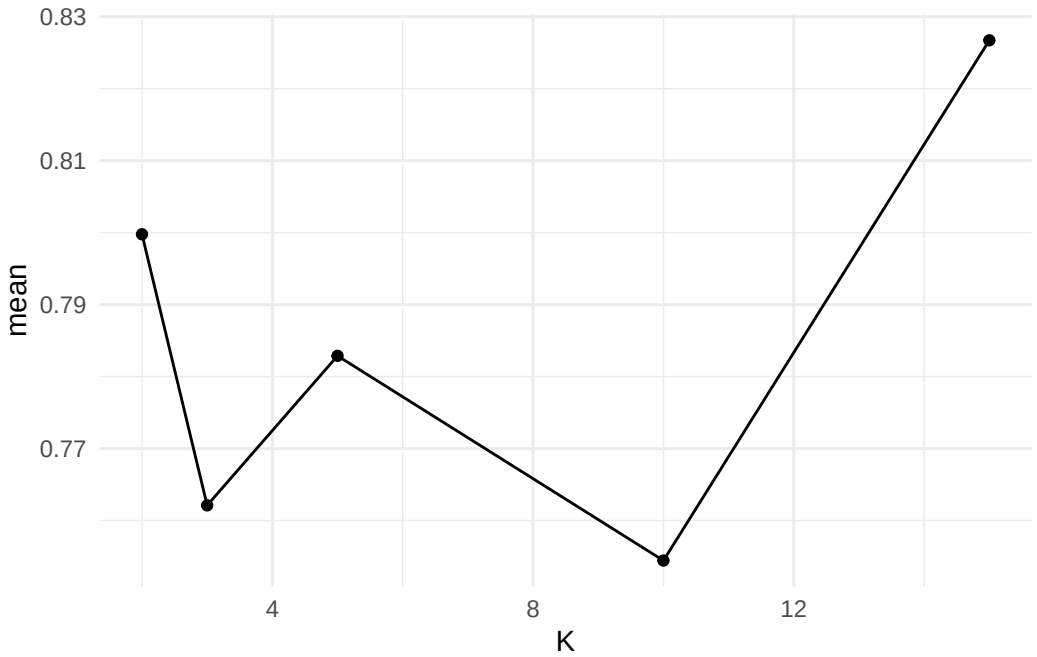

Figure 5: Relationship between Out of Sample AUC and K in K-fold validation

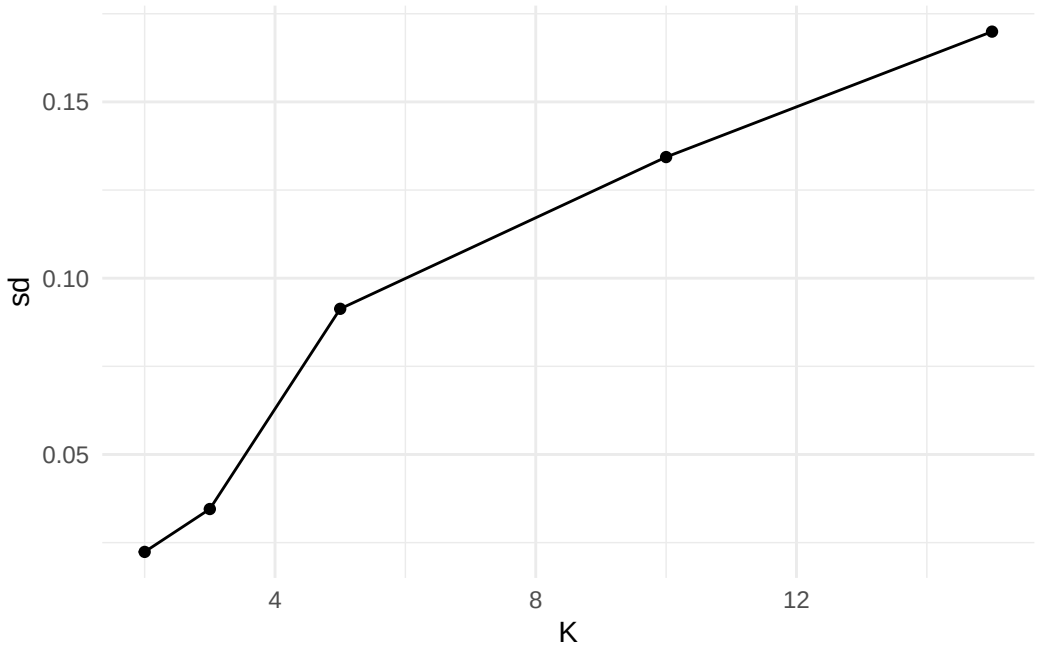

Figure 6: Relationship between variation of Out of Sample AUC and K in K-fold validation

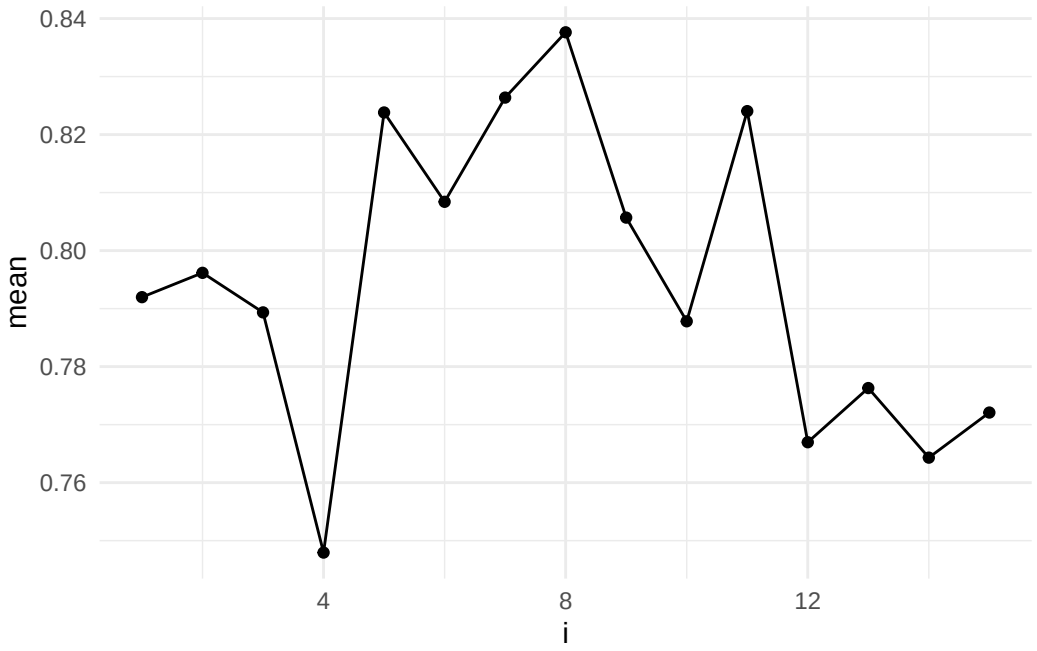

Figure 7: Effect of seed change on mean OOS AUC

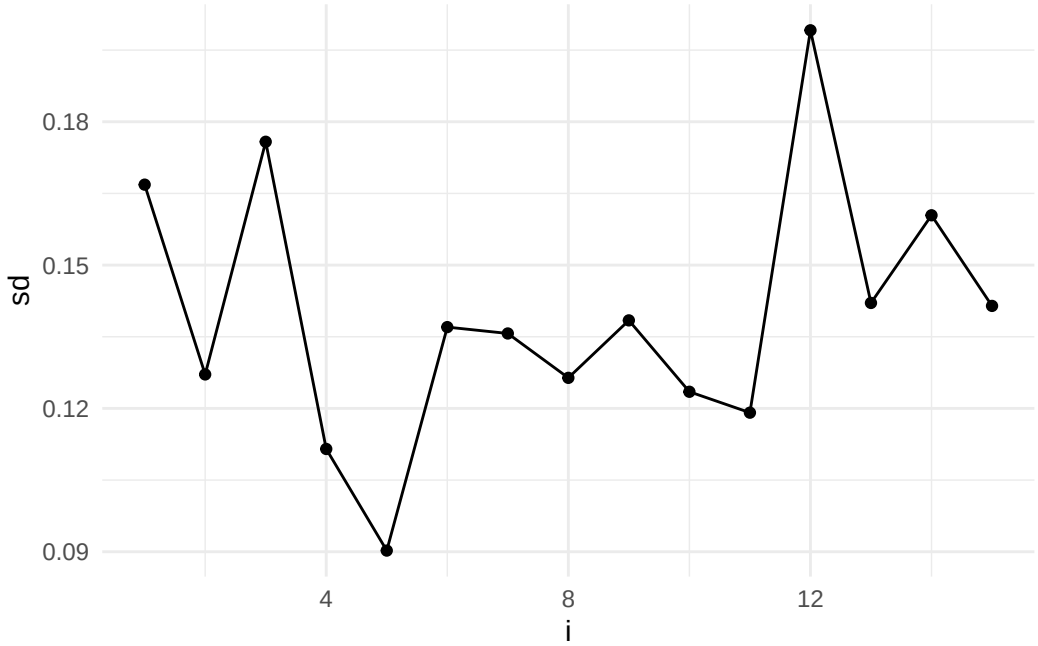

Figure 8: Effect of seed change on variance of OOS AUC

(4)

Figure 9 shows that the AUC score for the validation data set is lower that what we observed for the training data. We expect this kind of behavior, as we know that the model usually over fits to the training data.

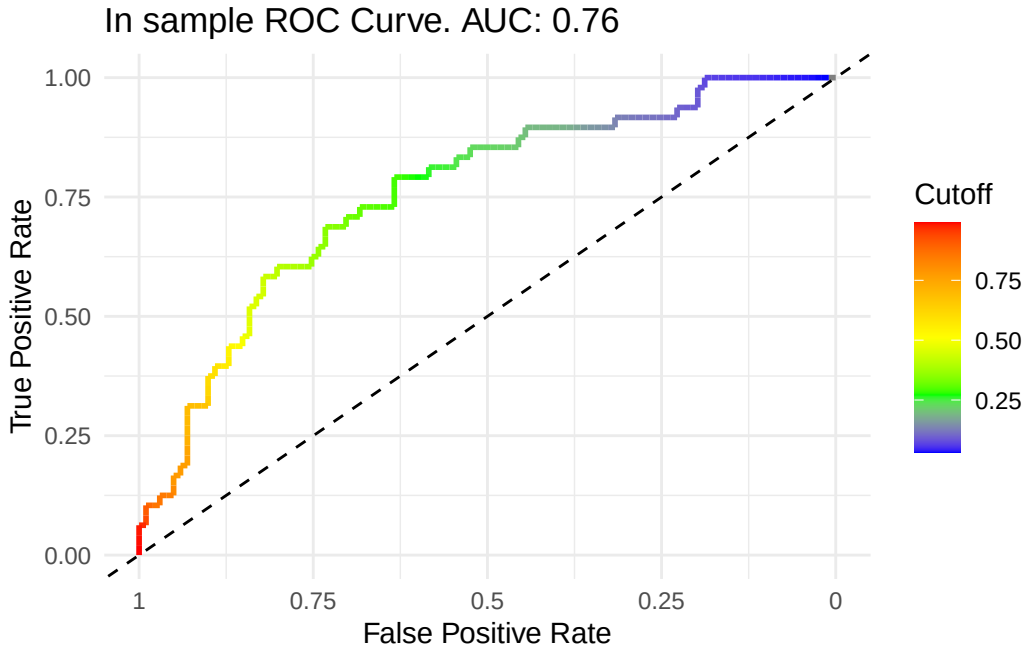

Figure 9: ROC Curve of the validation data

We can also compare the two ROC curves side by side to identify where the main differences occur. Figure 10 shows the two curves. If appears that the main deference occurs when we increase the decrease the cutoff.

Using training data, this decreases in cutoff allowed us to correctly capture more and more true positives. This may suggests that predicted probabilities for the group of positives are towards lower values of the 0-1 range.

This is not the case for the validation data sample, as with the decrease in the cutoff π_0 we see that the curve does not go straight up, and we start to observe the effects of a trade off right away.

AUC for the validation data is lower than the average expected under the 10-fold and other K-fold cross-validation methods, but is still within range of values under sampling variability. In other words, when we varied seed and K, we saw that the average AUC may drop to values as low as 0.75, which is the case with the validation data here. Performance is on the lower end of expected accuracy, given data variability.

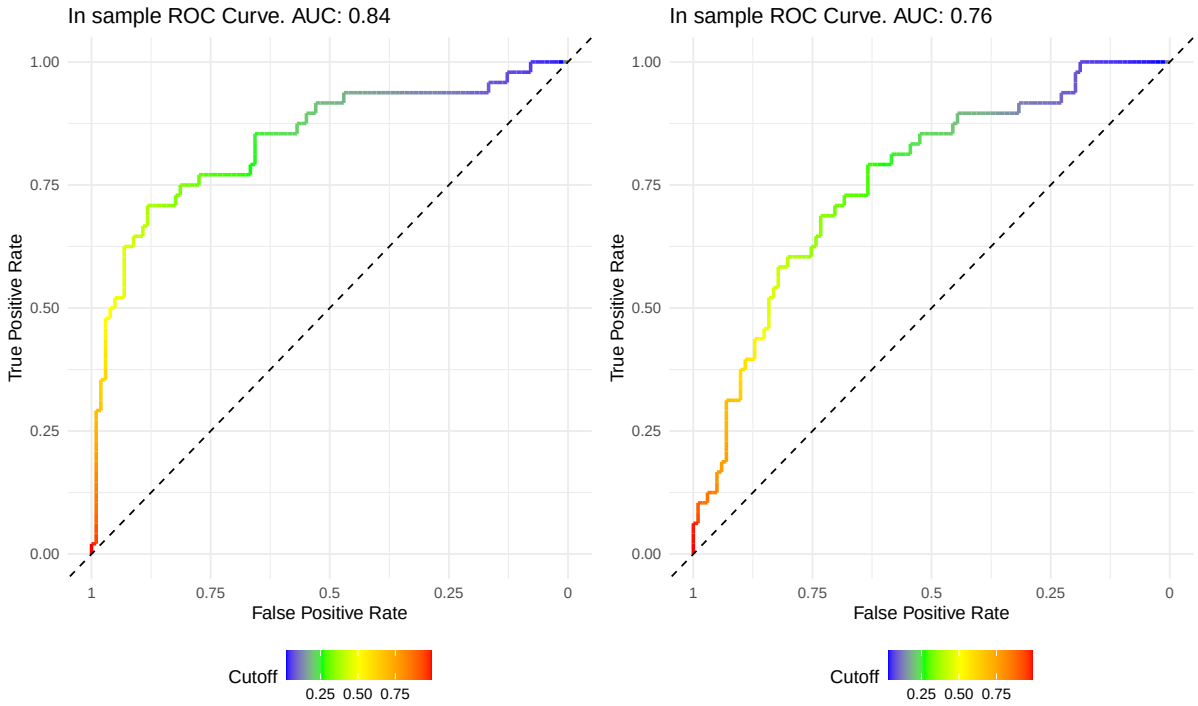

Figure 10: Side by side comparison of In Sample and Out of Sample ROC curves

(5)

Using a LASSO model with two predictors produces a fair predictive performance, but as Figure 11 suggests, the model is extremely poorly calibrated.

For those who are truly at lower risk of a death event, the model tends to predict risks that are lower than those, so the model understates the actual risk of a death event for people with low chances of a death event.

It goes the opposite way for those at the higher end of the risk spectrum, where the model drastically overestimates the risk of mortality.

Given that we also can missclassify patients, using this model we may scare patients and give them news that are far worse than what the reality would be.

Dxy	C (ROC)	R2	D	D:Chi-sq	D:p
0.51258251	0.75629125	0.16116123	0.11580409	18.25480922	NA
U	U:Chi-sq	U:p	Q	Brier	Intercept
0.03196198	6.76233554	0.03400772	0.08384211	0.18792590	-0.38358233
Slope	Emax	E90	Eavg	S:z	S:p
0.62048321	0.30149858	0.16723564	0.04781153	1.69666772	0.08975950

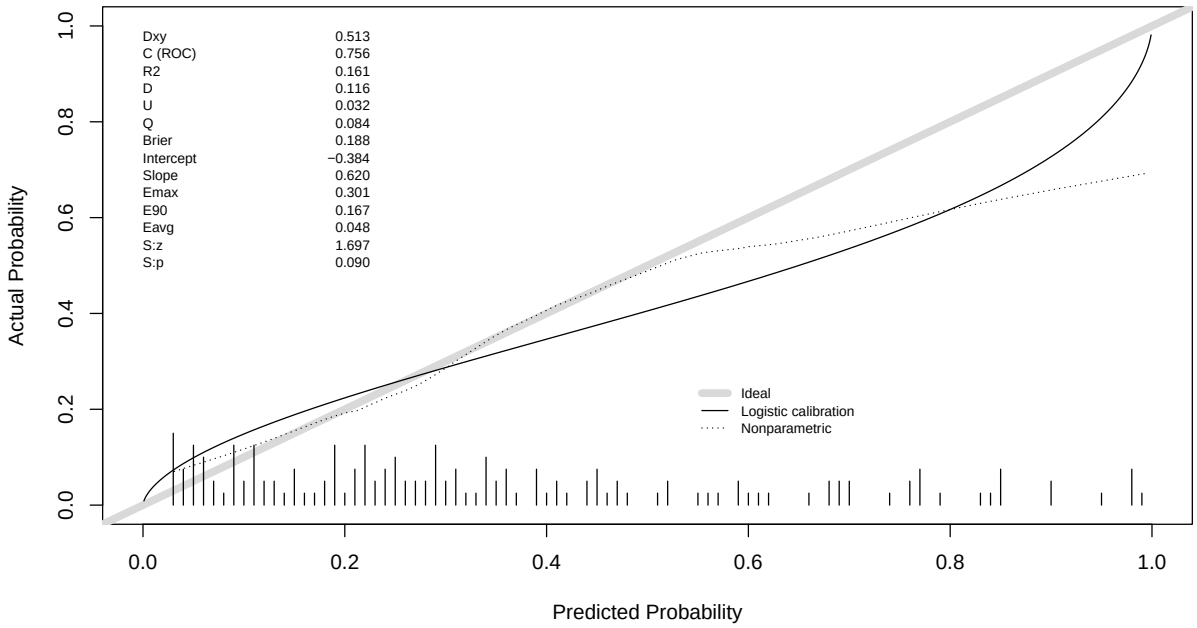

Figure 11: Two-predictor model calibration plot

As an experiment, I relaxed the shrinkage penalty and included more variables by making the penalty parameter smaller. Using $\lambda = \texttt{exp(-4)} = 0.02$, we can see on Figure 12 that the AUC and shape of the curve remain the same. Therefore, we have reason to believe that overall the ordering of probabilities remain approximately the same.

A new calibration plot on Figure 13 suggests that inclusion of more variables fixes calibration issues.

```
       Dxy     C (ROC)          R2           D    D:Chi-sq         D:p
0.51072607  0.75536304  0.21847648  0.16326559 25.32657283          NA
         U    U:Chi-sq         U:p           Q       Brier   Intercept
-0.01130890  0.31497445  0.85428773  0.17457449  0.18177385 -0.12731855
     Slope        Emax         E90        Eavg         S:z         S:p
0.95138508  0.21789128  0.06259897  0.03880975 -0.14323625  0.88610360
```

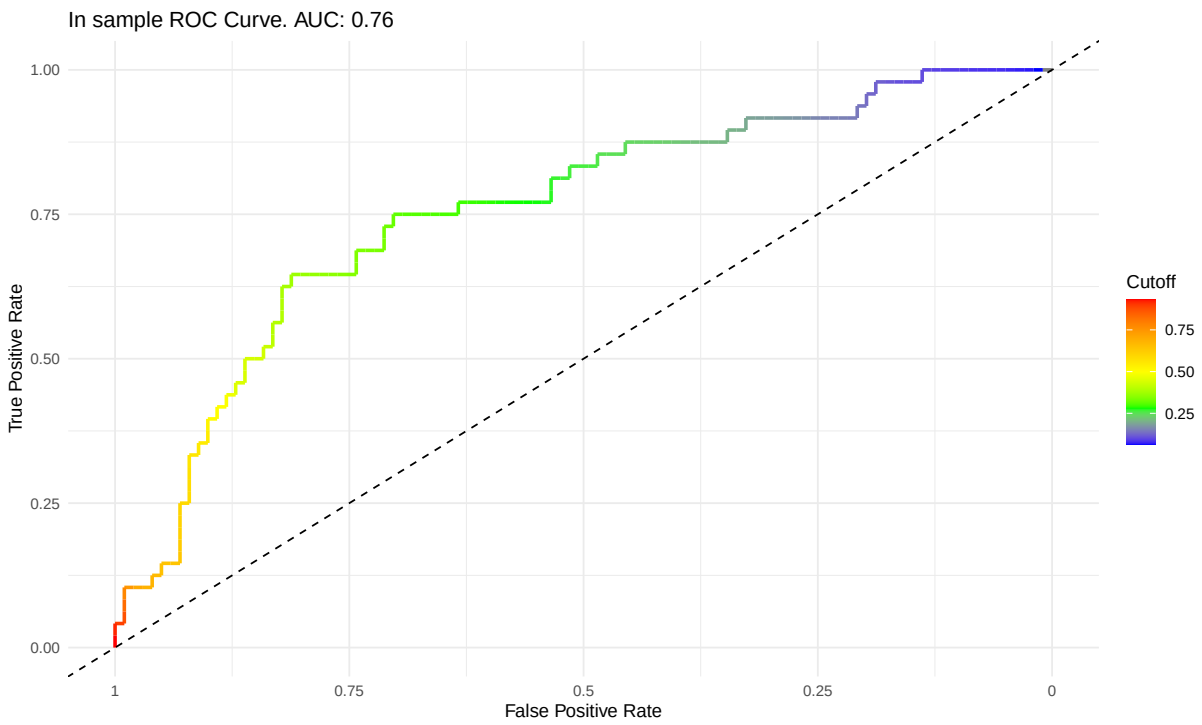

Figure 12: New Model ROC on the validation data

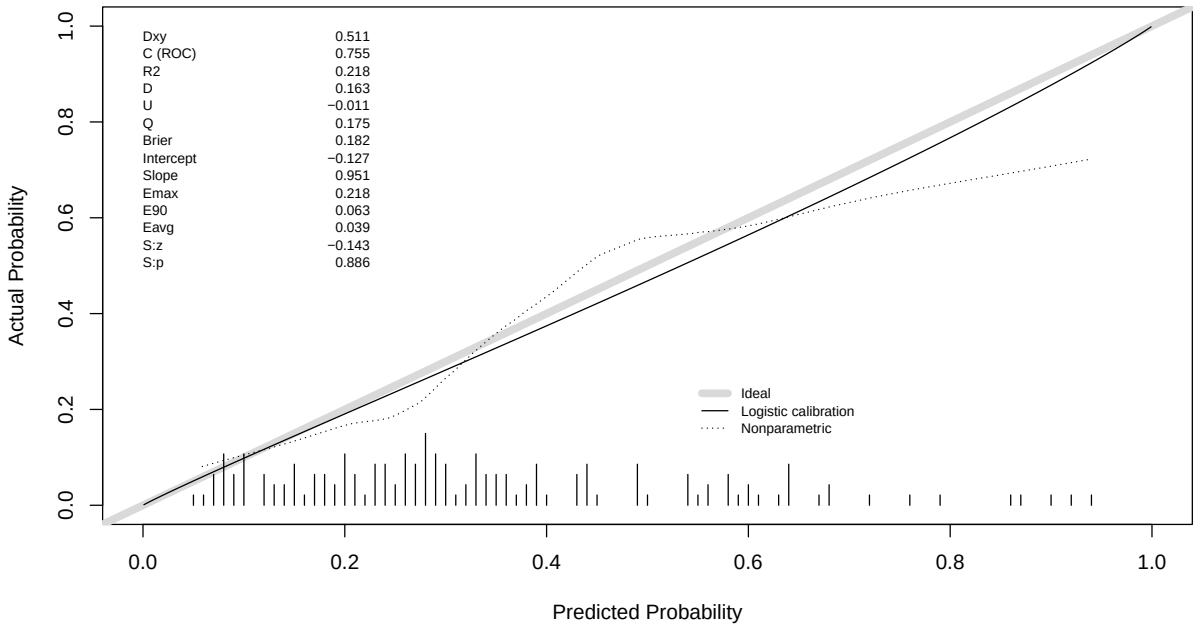

Figure 13: New calibration plot on the validation data