STAT 432 Final Project

Jeff Massman - massman4

11/24/2021

Project Description And Summary

There are two main objectives of this report: to construct models to predict the **histological subtype** and **progesterone (PR) status** of breast cancers, and select a subset of 50 biomarkers from the 1936 available to accurately predict all four outcomes considered (the above two, as well as **Estrogen status** and **HER2 final status**).

For methodology, see below.

At the end of this section, a summary of the findings and conclusions is provided.

This report is organized into several sections:

1. Literature Review

In this section, we will first review the relevant literature and adopt some of the applicable methodology used by authors. There are five different articles that will be examined. Further details can be found in this section. See the end of this report for references.

2. Summary Statistics and Data Processing

This section will be where we provide a brief exposition of the data, including relevant summary statistics, graphics, and any transformations, if necessary. This will also be where we describe, in detail, how we will modify the data to conform to our formulation of the problem.

3. Modelling PR Status

PR Status is a binary variable indicating whether the breast cancer in question contains Progesterone receptors. We label **PR Status** as "Positive," indicating that the observation contains Progesterone receptors, or "negative" otherwise. In this section, we will construct two classification models using two different methods (CART and random forest) and evaluate the effectiveness of these models using the "classification error" on a testing data set, i.e the proportion of misclassifications resultant from the model.

4. Modeling Histological Subtype

In this section, we build a classification model to predict the histological subtype of the breast cancer given the predictors. The modelled response will be binary with categories “infiltrating lobular carcinoma” and “infiltrating ductal carcinoma."" We will again use two different methods for this section (LDA and SVM) with AUC as the criterion. See relevant section for more details.

5. Variable Selection

The goal of this section is to investigate the possibility of selecting a small subset of 50 predictors to accurately classify all four of the considered responses. We will use a three-fold cross-validation with AUC as the metric to evaluate the effectiveness of such an approach. See relevant section for more details.

Conclusion

We found that the CART performed well for classifying PR status, however the poorer performance of the random forest casts the effectiveness of this model into doubt.

For predicting the histological type, we found that the LDA model performed significantly better than the SVM, though the drawback was that we used principal components which obscures the interpretability of the model. Nevertheless, it had a relatively high AUC value.

For the variable selection, we found that a purely data-driven approach, comparing across lasso-selected variables tuned for each response, produced less-than-adequate results, with AUCs hovering around 0.7. On the other hand, while slightly better, the literature-motivated model performed similarly.

Literature Review

In this section, we will summarize the relevant findings of four different articles.

The first article, "Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer", actually provides some background on the data set used in this analysis. In the article, Giovanni Ciriello et al. investigate the problem of classifying the histological subtype of breast cancer subjects. Like this report, the authors primarily focus on the subtypes **invasive lobular carcinoma (ILC)** and **invasive ductal carcinoma (IDC)**. They discovered that mutations targeting genes PTEN, TBX3, and FOXA1 were "ILC enriched features", i.e. that are highly associated with the ILC subtype. Similarly, they found that mutations of the gene GATA3 were indicative of IDC [1].

In "Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes," Joel Parker et al. sought out a 50 variable model to predict breast cancer subtype, which is quite similar to the goal of the variable selection portion of this report. The authors sampled a variety of methods to select their variables, including "the top 'N' t-test statistics for each group, top cluster index scores, and the remaining genes after 'shrinkage' of modified t test statistics." Ultimately, they decided upon the top N method as they found that it produced the smallest cross-validation error. We will borrow some of these ideas in this report [2].

An article published in the Journal of Clinical Oncology in June of 2010 evaluated the current state of ER and PR testing. The article found that the most widely used method, called immunohistiochemistry (IHC) is somewhat problematic, with a 20% miscalculation rate. This article also surveyed a few other attempts by other researchers to classify ER and PR status using different assays of biomarkers [3].

One such article in particular saw the authors identify a 21 biomarker assay in order to predict certain facets of ER, PR, and HER2, with good results. Some -- but not all -- of these markers are contained in our data set; we will incorporate these in our variable selection [4].

A similar article, published in the journal Nature in 2002, introduces another assay to use in predicting clinical outcomes of breast cancer. L. van 't Veer et al. also investigate identifying the ER status of breast cancers. We will borrow some of their findings [5].

Summary Statistics and Data Processing

The data used in this analysis contains 705 observations and 1936 predictor variables, consisting of 860 copy number variations, 249 mutations, 604 gene expressions and 223 protein levels. There are 5 outcome variables, namely `vital.status`, `PR.Status`, `ER.Status`, `HER2.Final.Status`, and `histological.type`. We will discard the response `vital.status` and only consider the other four outcomes. As mentioned before, all four outcomes will be modelled as binary variables. `PR.Status`, `ER.Status`, `HER2.Final.Status` will be modelled as indicator variables and will be encoded as `"Positive"` and `"Negative"`, indicating the presence or absence of the effect in the subject. `histological.type` will have two classes, namely ILC and IDC.

First and foremost, the predictor variables contain no missing values, as is demonstrated with the folloing R code:

```
any(is.na(cancer[,1:1931]))
```

```
## [1] FALSE
```

Therefore, imputation of the data will not be necessary.

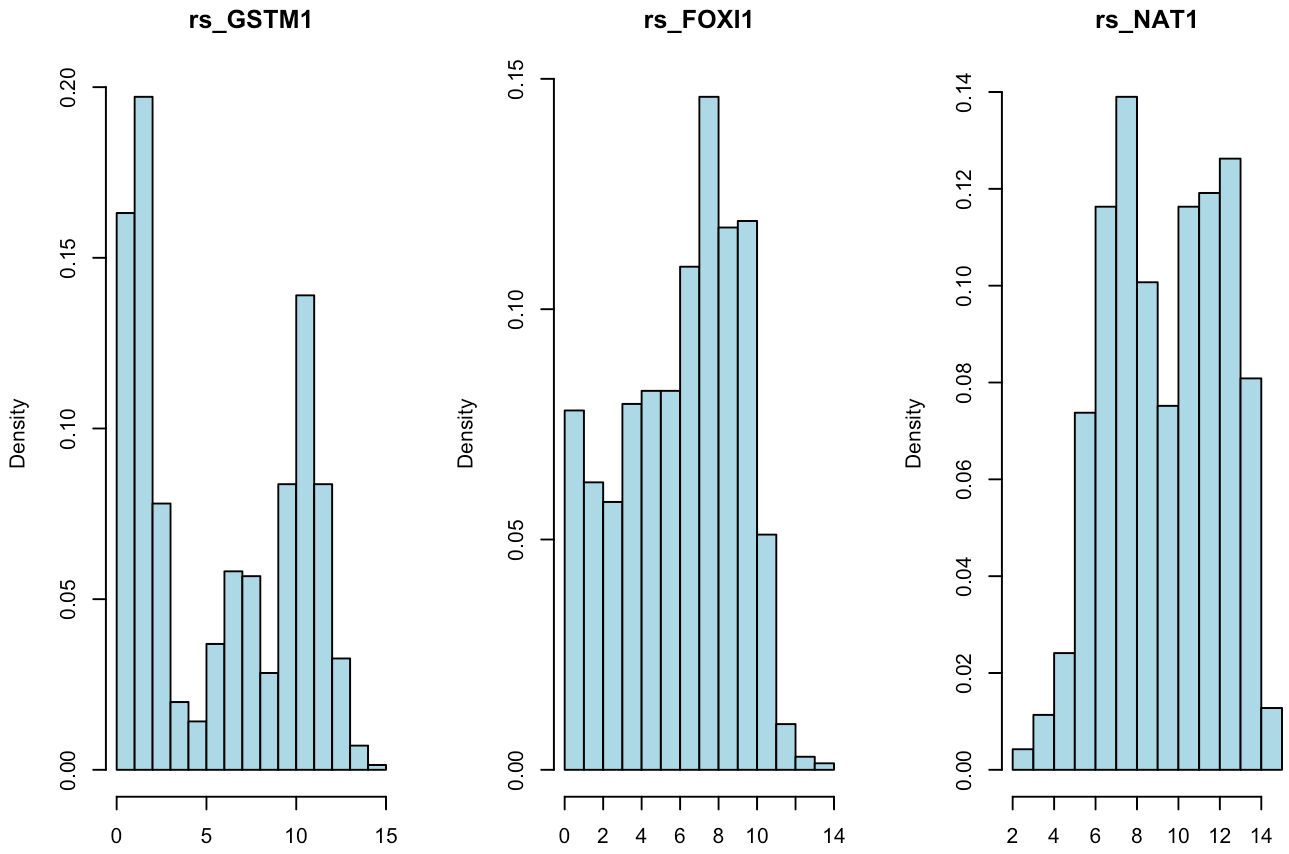
It is a slightly different story for the response variables, however. For `PR.Status`, there is a considerable number of missing values. Observe the following table:

```
##
##                                      Indeterminate
##                         122                     4
##                    Negative            Not Performed
##                         193                    28
## Performed but Not Available                 Positive
##                           5                   353
```

As evidenced by the table, there are a number of useless values in the response. This leaves only 546 total positive and negative values that we can use in the `PR.Status` classification section of this report. The situation is similar for all the other response variables except for `histological.type`, which has no missing values. Therefore, for certain parts of this report, we must use a smaller subset of the entire data set.

That being said, there is a different potential issue with `histological.type`, which is that it is relatively unbalanced. There are 131 ILC and 574 IDC labels. This corresponds to a $18.6\% - 81.4\%$ split, which may prove to be a problem for classification. We will investigate this further in that section of the report.

The first 604 variables (columns) of the data are the gene expressions. These are continuous variables, each with a different distribution. some are multimodal, some are approximately normally distributed, and many of them contain several repeated values of 0. Here are a few example distributions:



Since most of these variables will eventually be discarded in the variable selection section of this report, we need not worry about any transformations.

The next 860 are the copy number variations. These variables are discrete, with four integral values: $-2,\ -1,\ 0,\ 1,$ and 2. Here is an example table for the variable `cn_RIMS2` :

```
##
##  -2  -1   0   1   2
##   1  18 243 313 130
```

In general, there is an imbalance in the distribution. Observe:

```
##
##        -1         0         1
## 0.1383721 0.7465116 0.1151163
```

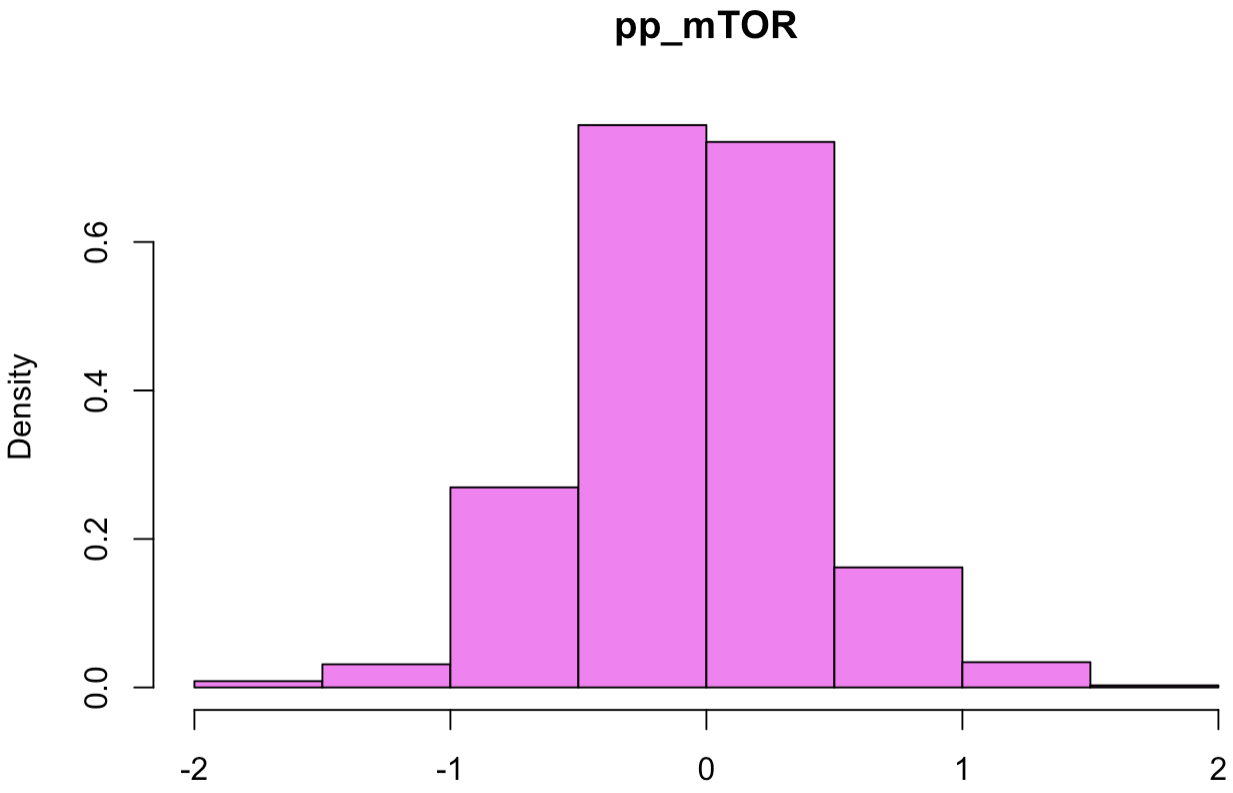
This is a table of the proportion of values which are the most frequent in its variable. For example, in 74.65% of the copy number variation variables, 0 is the most frequently occuring value. -2 and 2 are not the most frequent value in any of the variables.

The next 249 variables are mutations. These variables are binary encoded as 1 and 0, indicating the presence or absence of said mutation respectively. These variables are also imbalanced; in general, 1 is sparse, which makes sense. Here is an example, for `mu_ABCA13` :

```
##
##   0   1
## 679  26
```

As evident in the table, 0 is quite common, while 1 is sparse.

Finally, the remaining 223 variables are the protein levels. These variables are continuous. They follow bell-shaped distributions centered at 0, however, conducting Shapiro-Wilkes tests on a select few of them (with a Bonferroni-corrected significance level) have indicated that they are not normally distributed. For the sake of visualization, here is an example density histogram for the variable `pp_mTOR` :



Further Notes

The data contains some highly correlated predictors. Namely, there exist 1116 pairs of predictors (out of 1873080 possible pairs) that have a correlation coefficient greater than 0.95. None have an equally strong negative correlation. This means that we may have issues with approximate collinearity in our analysis. We will explicitly address this as it comes up in the relevant section(s).

Modelling PR-Status

In this section of the report, we will attempt to build a classification model for PR-Status, analyzing the performance of two different methodologies and deciding which is preferred.

First, as mentioned in the in the prior section, we will have to use a subset of the data that contains only

"positive" and "negative" values of the relevant response. We will further split the data randomly into training and testing sets, with an $80 - 20$ split respectively.

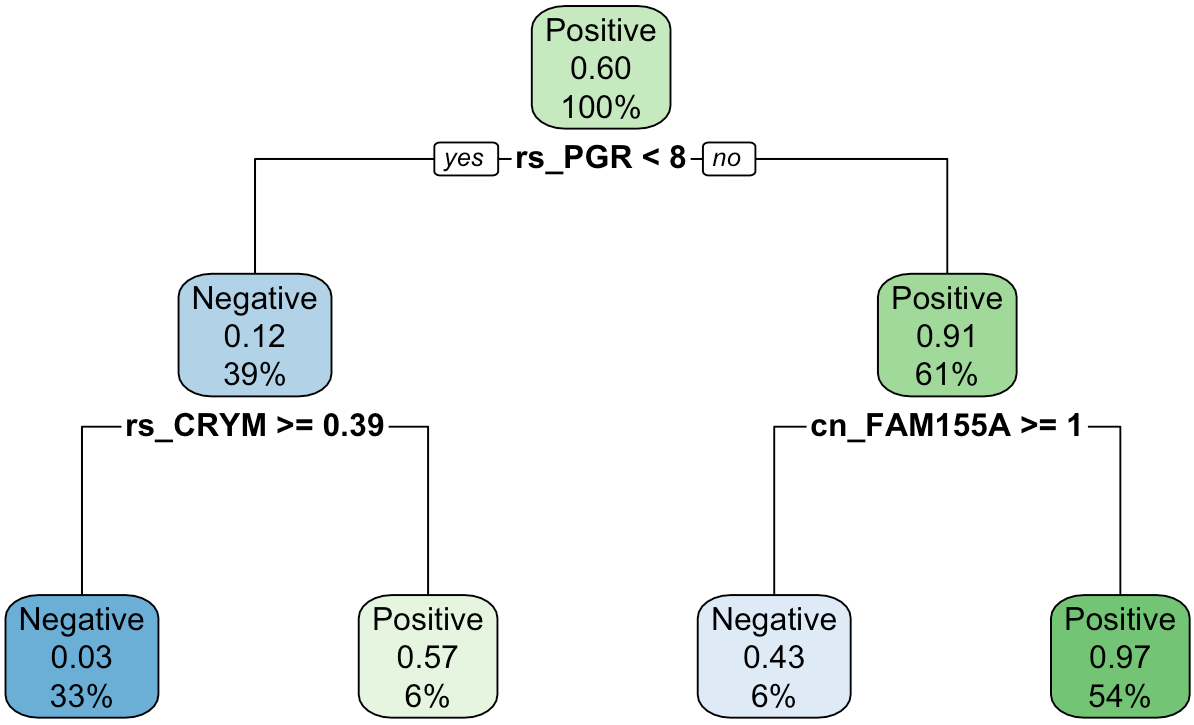
We will first fit a standard Classification Tree (CART) model. The following is the confusion matrix resulting from prediction on the testing data set:

```
##
## yhat       Negative Positive
##   Negative       39        4
##   Positive        5       61
```

The **classification error** associated with this model is 0.0825688 and the accuracy is 0.9174312. This model is decent on its own, however, we may attempt to reduce its complexity by pruning the tree. We will do so using the $1 - \text{SD}$ method. The resulting confusion matrix:

```
##
## yhat2      Negative Positive
##   Negative       39        4
##   Positive        5       61
```

Our new pruned tree maintains the same predictive power as before.



This figure is a small portion of the pruned classification tree, provided to offer some visual intuition as to what the tree is doing. In reality, the tree is too large to reasonably display here.

The CART model is good, but can we do better? We will fit another type of classification model to see if we can improve upon the performance of the previous tree model. In the same branch (pardon the pun) as the classification tree model, we will fit a 1000 tree random forest and analyze the results.

```
##
## pred       Negative Positive
##   Negative       28        3
##   Positive       16       62
```

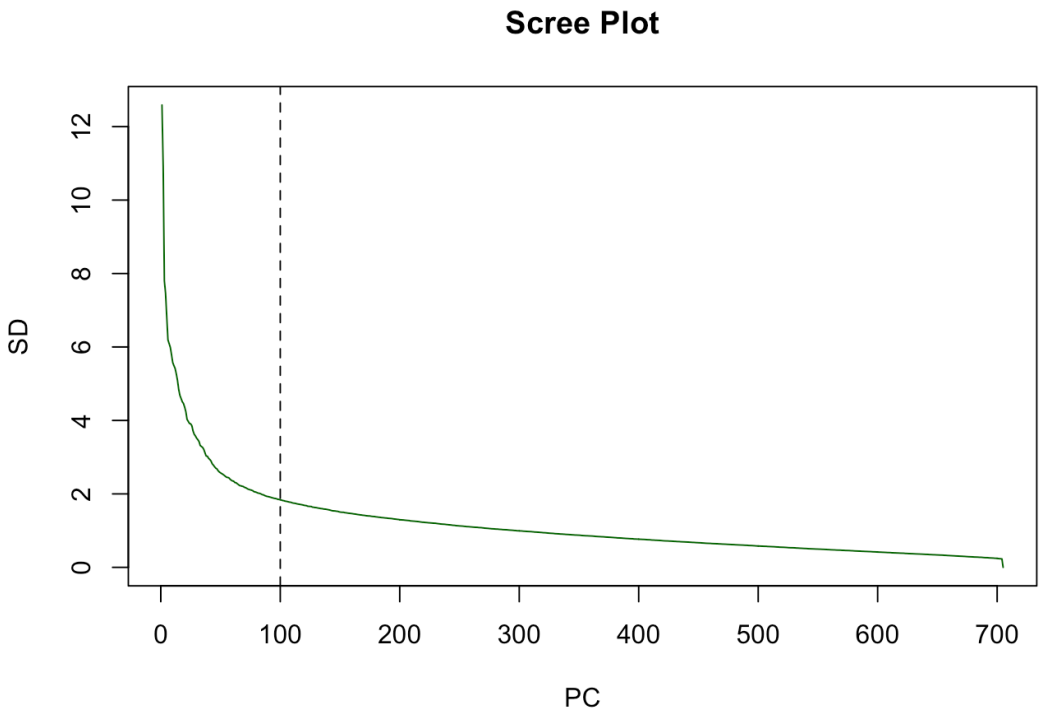
The classification error for the random forest is 0.1743119 which is not that great, and which is significantly higher than the single tree model. Additionally, though the random forest is adept at identifying true positives, it seems to particularly struggle with false positives. This indicates that the random forest seems to prefer predicting into the dominant class, `"Positive"` . Since the random forest is a more stable version of a single classification tree model, this possibly casts the good performance of the prior model into question; its good performance may be due to chance.

Modelling Histological Subtype

In this section, we will build a model to predict the histological subtype of the cancer. We will again use an $80\% - 20\%$ training-testing data split, except this time, since there are no missing values in the response under consideration, we can use the entire data set. The evaluation criterion will be "area under the curve" (AUC).

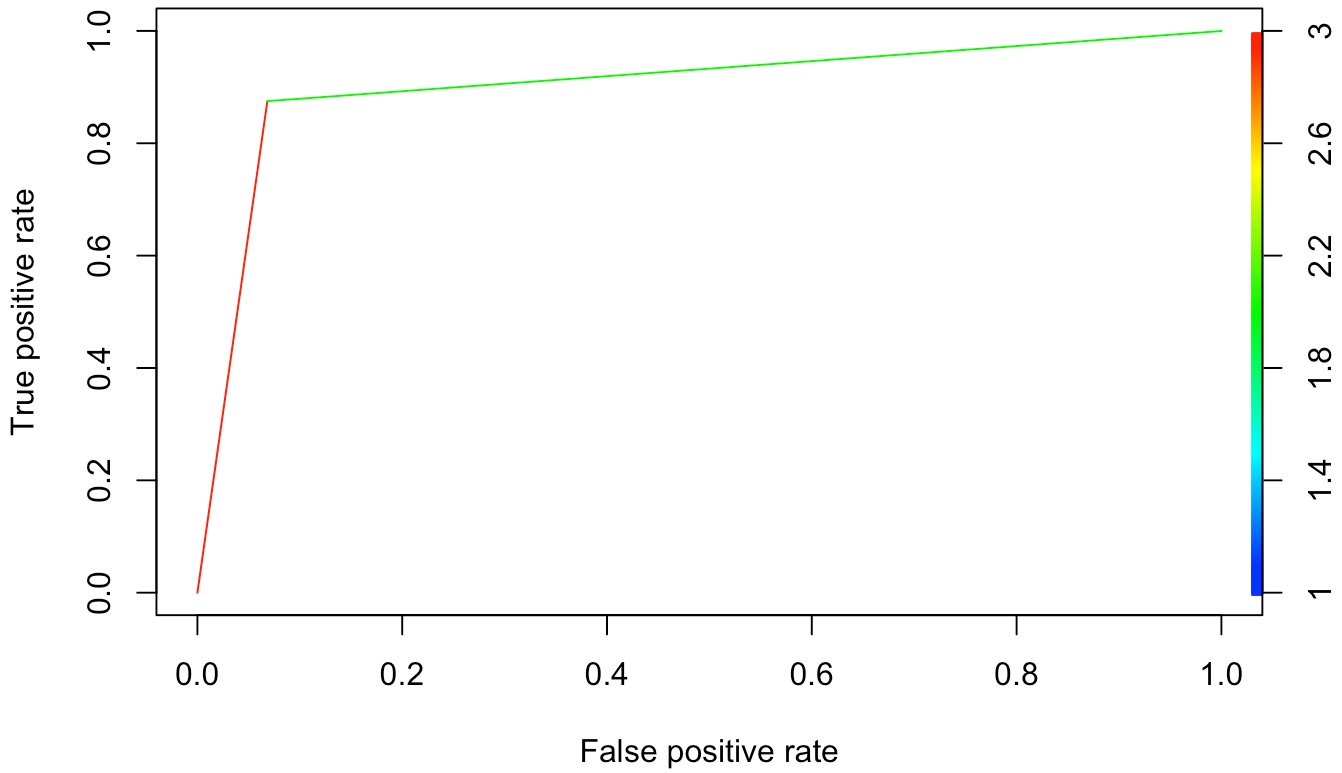
The first method we will use is Linear Discriminant Analysis (LDA). Recall that, as was mentioned in the data summary section, some of the predictors are approximately collinear. This can be an issue for LDA, and we will need to resolve this before proceeding with a proper analysis. We will use principal components to remedy this.

Here is the (standard deviation) Scree Plot:



We will use 100 PCs in our analysis.

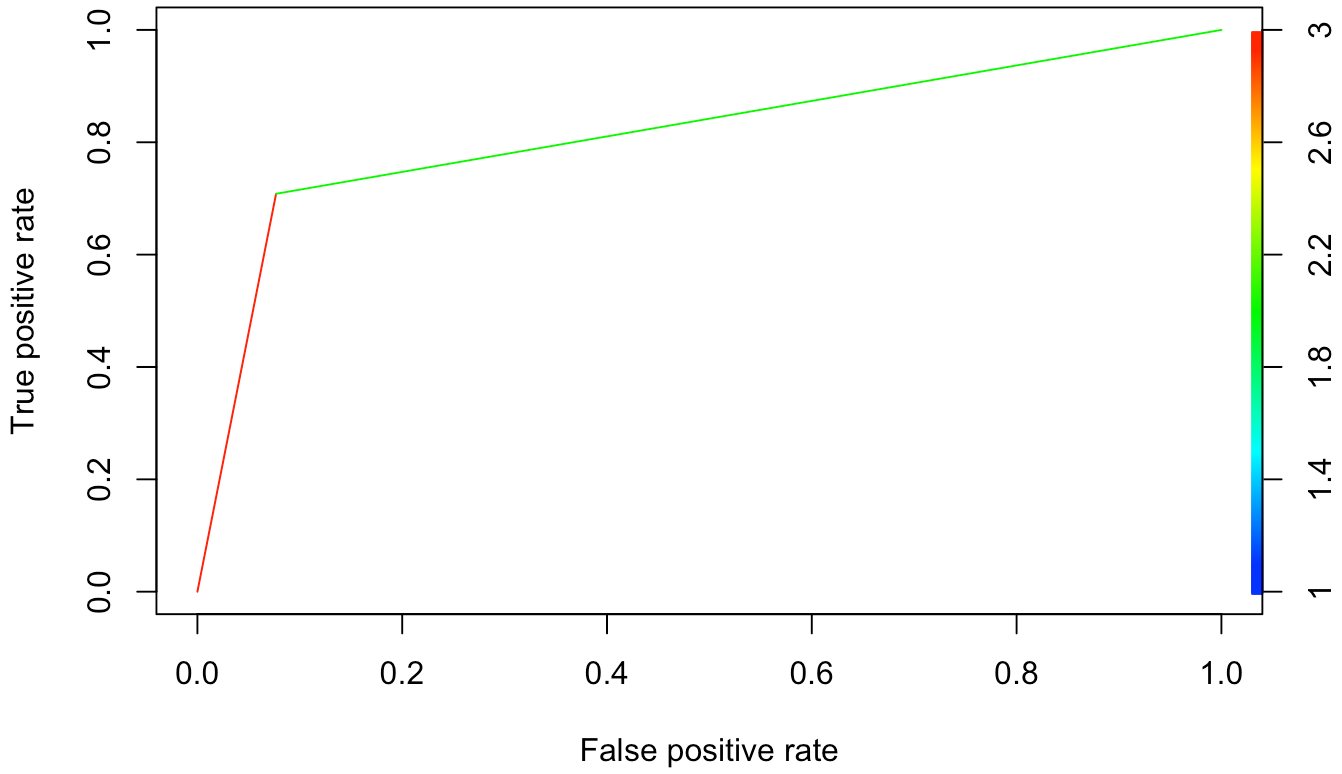
We will further split the PC data into a training and testing set. The performance curve:



The AUC for this model is 0.903312.

This LDA model is pretty good, however not spectacular, and using the PCs may obfuscate the interpretation of the results. Can we do better?

We will now fit a linear Support Vector Machine (SVM) model. Here is the corresponding curve:



The curve is further from the top left than the previous one; the AUC for this model is 0.8157051. Unfortunately, this model performs worse than the LDA with PCs.

Furthermore, the model seems resistant to tuning. Adjusting the `cost` parameter has failed to enhance the model performance. Similarly, changing the kernel (e.g to "radial" for instance) actually worsened the prediction. QDA results in prediction into the dominant class.

Therefore, we conclude this section by reporting that LDA with 100 PCs is the preferred model of the two surveyed.

Variable Selection

In this section, our goal is to select just 50 predictors that are useful in simultaneously--and sufficiently--classifying all four responses, namely `PR.Status`, `ER.Status`, `HER2.Final.Status` and `histological.type`. Selection of predictors will be partially motivated by the literature, as well as by standard statistical methods.

We will proceed with the following methodology for evaluating the models:

1. We will select some set of 50 predictors using the above methods
2. Three-fold cross validation will be conducted on the models, with AUC as the criterion, which will be averaged over the three folds for each model
3. All four average AUCs will then be averaged again and recorded
4. The above four steps will be repeated with a different subset of 50 predictors. AUC values will then be compared to select the optimal set of predictors.

Standard logistic regression will be used to do the actual classifying. The classifying cutoff may change if improved performance is observed.

We begin by using as many of the genes in our data set as possible that were found to be significant by by Joel et al.

This gives us 26 out of the 50 predictors.

We can also include the predictors that Giovanni Ciriello et al. found to be significant in predicting histological subtype. This brings us to 29. Furthermore, we can also include as many of the 21 ER, PR, and HER2 assay markers as are in our data, as provided by Paik, Soonmyung et al.

In all, this brings us to 41 predictors. To obtain the remaining 9, we can use some of the predictors identified in van 't Veer, L. et al.

The result after applying the above procedure:

```
## [1] 0.737842
```

This is not a very good AUC value unfortunately. We will now turn our attention to the data-driven approach.

We will conduct the above enumerated procedure four times, where each set of 50 predictors will be tailored to one of the four responses using lasso. The best of these four will be the representative model of the data-driven approach.

For the lasso, λ is chosen so as to result in 50 nonzero predictors. If we cannot get exactly 50, any extra will be removed.

Also, it should be noted that, as mentioned earlier, some of the responses have missing or incompatible values, so the corresponding observations will have to be removed for that portion of the model-fitting.

After executing the above outlined procedure, we obtain the following results:

Mean 3-CV AUC Across All Classes

	AUC
histological.type	0.7243565
PR.Status	0.7092728
ER.Status	0.7023672
HER2.Final.Status	0.6992398

The 50 predictor set chosen using `histological.type` as the base response had the best performance. However, these models are, frankly speaking, not very good. AUC values in this range are typically undesirable.

Therefore, we conclude that the literature-driven model, though not very good in and of itself, proves to be slightly preferable to the purely data-driven model, with an AUC of 0.737842.

References

1. Ciriello, Giovanni et al. "Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer." *Cell Press* vol. 163,2 (2015): 506-519. doi:10.1016/j.cell.2015.09.033 (doi:10.1016/j.cell.2015.09.033)

2. Parker, Joel S et al. “Supervised risk predictor of breast cancer based on intrinsic subtypes.” *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* vol. 27,8 (2009): 1160-7. doi:10.1200/JCO.2008.18.1370 (doi:10.1200/JCO.2008.18.1370)

3. Hammond, M Elizabeth H et al. “American Society of Clinical Oncology/College Of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer.” *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* vol. 28,16 (2010): 2784-95. doi:10.1200/JCO.2009.25.6529 (doi:10.1200/JCO.2009.25.6529)

4. Paik, Soonmyung et al. “A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer.” *The New England journal of medicine* vol. 351,27 (2004): 2817-26. doi:10.1056/NEJMoa041588 (doi:10.1056/NEJMoa041588)

5. van 't Veer, L., Dai, H., van de Vijver, M. et al. "Gene expression profiling predicts clinical outcome of breast cancer." *Nature* vol. 415, (2002): 530–536. doi: https://doi.org/10.1038/415530a (https://doi.org/10.1038/415530a)