An Analysis of Classification Models Predicting Pancreatic Cancer Via Urinary Biomarker Panel

Jonah Silverman
Muhlenberg College

Introduction

- Pancreatic cancer has a 5 year survival rate as low as $5 - 10\%$[4].
- Lack of early diagnosis contributes to pancreatic cancer's poor prognosis. Currently there are no clinically useful screens for individuals not at heightened risk.
- We set out to build a classifier which is able to differentiate between pancreatic cancer positive patients and healthy controls, based on a panel of Urinary Biomarkers.

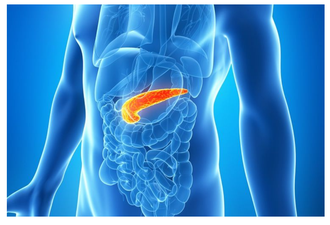

Fig. 1: Pancreas Position in Body

http://www.shutterstock.com/gallery-65904p1.html

The Data Set

This data was collected from multiple health care centers around the world, and has predictor variables of TFF1, LYVE1, REG1B, and creatinine levels, as well as age, sex and pancreatic cancer status.[1]

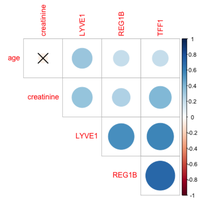

Fig. 2: Correlation Matrix of Predictors

Performance Metrics

The predictive models we build will be evaluated in regards to two metrics.

Classification Accuracy: = $\frac{\#\text{True Positive} + \#\text{True Negative}}{n}$

AUC, or Area under Receiver Operating Characteristic curve (ROC). ROC curves are generated by plotting the true and false positive rates across various classification thresholds. Higher AUC values indicate a better predictive model.

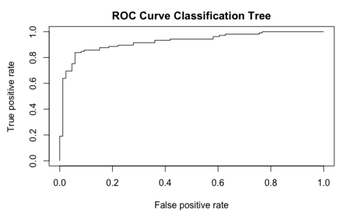

Fig. 3: ROC curve generated from a single classification tree on this data set.

Methods

Tree Based Methods[2]

- Classification trees function by preforming a series of uni-variate splits on the data. All data points which was binned the same way based on the splits fall into the same terminal leaf of the tree.
- Splits are made to decrease the impurity of the tree, with impurity referring to how homogeneous each terminal node is in terms of the classification variable of interest.
- The impurity metric used here is the gini index.
  $Gini = p_1(1 - p_1) + p_2(1 - p_2)$
  With $p_{1,2}$ referring to the respective proportions of class 1 and 2.
- Derived from this basic tree we employ bagging, random forests, and gradient boosting in order to create more powerful models.

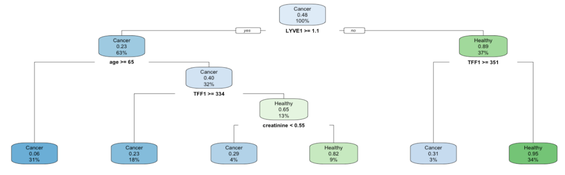

Fig. 4: Classification Tree Example

Support Vector Machine[5]

- Working with p predictors, our data exists in a p-dimensional feature space. Support vector machines generate a p-1 dimensional hyper plane which aims to separate the two classes present in the data set.
- Kernel functions map the predictors into a higher dimensional space where the hyperplane can more effectively divide the two classes by allowing for non-linear decision boundaries.
- The two kernel functions used in this project are:
  The radial kernel: $K < x, x' >= e^{-\gamma||x-x'||^2}$
  The polynomial kernel: $K(x, x') = \gamma(1 + < x, x' >)^d$

General Additive Models[6]

- A General Additive model refers to a function of the form:
  $g(y) = B_0 + f_1(x_1) + f_2(x_2) + ... + f_p(x_p)$
  Where $f_{1,...,p}$ are smoothing functions, in this case B-Splines.
- Splines relate predictors to the logit transformed binary response variable by using piecewise regression along the variable domain, with the constraint that the overall curve must be smooth.

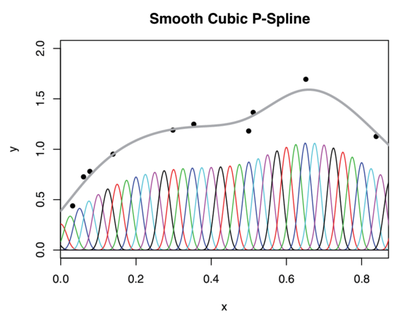

Fig. 5: Example of P-Spline Basis Function

Results

We applied these various metrics to our data set and compared the optimal models based on optimal accuracy and Area Under ROC curve.

Model	Accuracy (95% Conf.)	AUC (95% Conf.)
Classification Tree	0.829 (0.804 - 0.854)	0.888 (0.858 - 0.910)
Bagged Tree	0.868 (0.854 - 0.882)	0.940 (0.933 - 0.947)
Random Forest	0.877 (0.864 - 0.890)	0.941 (0.933 - 0.949)
Boosted Tree	0.870 (0.863 -0.880)	0.945 (0.938 - 0.952)
SVM (Radial Kernel)	0.876 (0.859 - 0.892)	0.930 (0.919 - 0.941
SVM (Polynomial Kernel)	0.872 (0.853 - 0.891)	0.926 (0.913 - 0.939)
GAM	0.869 (0.856 - 0.882)	0.939 (0.929 - 0.949)

Judging from these results we see that the accuracy and ROC confidence intervals are overlapping for many of the models. The Support Vector Machine using a Radial Kernel and Random Forests produced the highest classification accuracy, while a Boosted Tree was able to produce a model with the largest AUC.

Variable Importance

While we cannot preform traditional statistical inference to determine the significance of each of these variables, we can analyze the model to examine which predictors it is relying on most to make accurate predictions. The difference between standard model performance and the performance of the model with one predictor randomized in the test set gives a measure of importance for the randomized predictor. Below we highlight the variable importance plots for the SVM (Radial) and Random Forest.

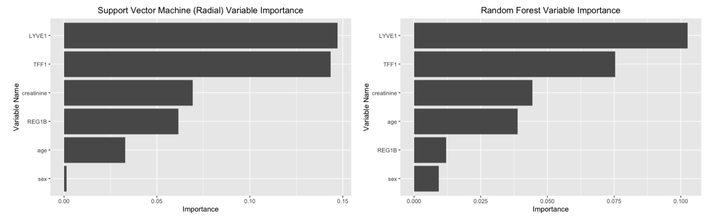

Fig. 6: Variable Importance Plots for SVM and Random Forest

Next Steps

Looking to the future we are excited about trying other machine learning techniques, specifically Artificial, Conventional, and Recurrent Neural Networks.

Acknowledgements

I would like to thank my Advisor, Dr. James Russell, and the Muhlenberg Math and CS Department, and the Muhlenberg College Deans office for supporting these efforts.

References

1.) Blyuss, O., Zaikin, A., Cherepanova, V. et al. Development of PancRISK, a urine biomarker-based risk score for stratified screening of pancreatic cancer patients. Br J Cancer 122, 692–696 (2020).
2.) Breiman, L., Friedman, J. H., Olshen, R. A., amp; Stone, C. J. (1984). Classification and Regression Trees. Chapman amp; Hall/CRC.
3.) Eilers, Paul Marx, Brian Durbán, María. (2015). Twenty years of P-splines. SORT (Statistics and Operations Research Transactions). 39. 149-186.
4.) Key statistics for pancreatic cancer. American Cancer Society. (n.d.). Retrieved December 11, 2021.
5.) Schölkopf, B., amp; Smola, A. J. (2018). Learning with kernels support vector machines, regularization, optimization, and beyond. MIT Press.

6.) Wood, S. (2017). Generalized additive models: An introduction with R (2nd ed.). CRC Press/Taylor amp; Francis Group.