Predicting Breast Cancer Survival Using Treatment and Patient Factors

William Chen | wchen808@stanford.edu
Henry Wang | hwang9@stanford.edu

1. Introduction

Breast cancer is the leading type of cancer in women worldwide, accounting for about 25% of all cancer diagnoses. Though this fraction is lower in developed countries, breast cancer is still the second most commonly diagnosed cancer in the United States. In addition, breast cancer still claims the second highest death toll of all cancer in women, behind lung cancer[1]. With this project, we attempt to shed some light on risk factors and effective treatments of breast cancer.

2. Background and Preparations

Breast cancer is especially difficult to characterize because there are many risk factors, subtypes, and treatment methods. In order to explore potential correlations between the many features of breast cancer patients and patient survival, we obtained a large data set of ~10,000 patients from the Stanford Medical School. The data set contained categorical patient factor features (e.g. socioeconomic status, degree of education, stage of disease), binary treatment features (e.g. specific medications, chemotherapy, radiation treatment), and binary outcome features (e.g. 1, 3 and 5 year survival) for each of the patients. We chose 5-year survival as our outcome variable of interest and noticed that the proportion of patients who were deceased within five years of diagnosis ("positive" test cases) was quite low (~10%). To mitigate the issue of class balance, we randomly selected negatives so that our total proportion of positives was 40%. Our final data set consisted of roughly 500 features for a total of 2500 patients, 1000 of whom did not survive past 5 years of diagnosis.

Our goal was to perform several different analyses on many input variables to improve management of breast cancer by identifying at-risk patients. We started with logistic regression and SVM classifiers to measure the effects of various treatment and patient factors on survival. We then implemented basic recursive partitioning, random forest, and gradient-boosted classification tree algorithms to improve the accuracy of our predictions. Our hypothesis was that through these methods and our input variables, we could accurately predict 5-year survival of breast cancer patients.

3. Initial Analysis of Drugs/Treatments

We began our analysis of the data by first estimating the predictive power of the entire feature set, which consists of 468 commonly used drugs that appeared in at least 1% of our patients, along with 5 other treatments of interest (chemotherapy, hormone therapy, etc.). To effectively analyze such a large set of data, we decided to use logistic regression for its general accuracy on

binary outcome variables (survivability in 5 years) and followed it up with support vector machines for its flexibility to choose the kernel that best matched our features' behaviors. Since these two algorithms fit general data sets well without making too many assumptions on the data (e.g. no assumptions about the existence of correlations), they seem to be the most appropriate for an initial estimate. Using the glmnet package in R[2], we were able to perform logistic regression analysis on our data set and generate the ROC curve shown in Figure 1. For SVM, we made use of the 'e1071' R package[3] to generate results of 10-fold cross validation shown in Table 1.

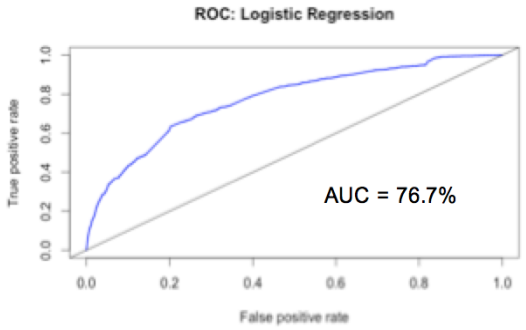

Figure 1. Receiver Operating Characteristic (ROC) curve of logistic regression on complete data set.

10-Fold Cross Validation (SVM)	
PPV	0.611
TPV	0.753
FPV	0.688
F1	0.674

Table 1. SVM analysis on complete data set.

Fortunately, we can see that the full feature set is able to give promising results from both algorithms, but the positive predictive value as seen from the SVM analysis is a bit lower than we would like. One explanation for this could be the existence of noisy features (drugs that do not significantly contribute to a patient's well-being), something we will try to eliminate. In addition, it is unrealistic for doctors to actually test and collect a patient's reaction to almost 500 different drugs. In our following analyses, we will attempt to isolate the features that are most indicative of patient survival to increase effectiveness while maintaining high predictive power.

4. Analysis of Feature Importance (Through Decision Tree Classifiers)

Our next supervised learning approach was to implement classification trees. Based on the nature of training and testing decision tree classifiers, we split our cohort of 2500 patients into two equally sized sets with equal proportions of "positives" – patients who did not survive past 5 years of diagnosis. One set was used to train our predictive model and the other to validate it.

Decision tree classifiers aim to identify subgroups of the data set, represented as conjunctions of certain input features, that generally share similar class labels. To do so, decision trees recursively partition the data in a binary fashion based on a condition on one of the input variables (e.g. 'cancer stage' > 2). The specific split is selected as that which maximizes the homogeneity of the nodes (i.e. minimizes the sum of the Gini Diversity Indices of the resulting

nodes). Splitting ends when either the terminal nodes are too small or when subsequent splits would not improve the purity of the node as measured by the Gini Diversity Index.[4] Through this heuristic, decision tree classifiers are often able to achieve excellent performance.

We had several motivations for implementing classification trees. One main advantage of classification trees over logistic regression and SVM classifiers is that classification trees are non-parametric. In particular, they do not assume that features are linearly related, which makes them more robust to different inputs. Moreover, classification trees reduce the feature space by inherently selecting and weighting features – those features with the greatest predictive power are the factors on which the tree is split, with higher priority features comprising the primary splits. A third benefit is that classification trees are not heavily impacted by outliers, as those outliers can represented as separate nodes.[5]

Recursively Partitioning Decision Tree

Our first attempt at using a decision tree utilized the 'rpart' R package.[6] Our approach implemented a single recursively partitioning decision tree as described above. However, to prevent overfitting, we pruned the resulting tree such that only partitions that would improve relative error by a fixed threshold would be included. The ROC curve for our classification tree method is displayed in Figure 2, with the weights of the most important features in Table 2.

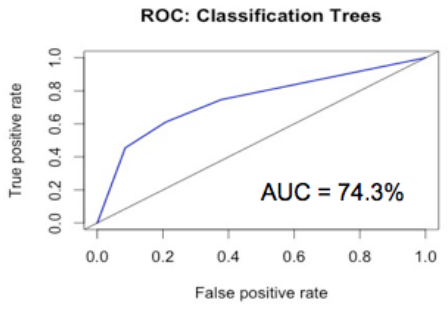

Figure 2. Receiver Operating Characteristic (ROC) curve of single recursive partitioning tree classifier

Patient/Treatment Factors	Relative Variable Importance
Final Stage	46
Chemotherapy Treatment	20
Tumor Grade	16
Dexamethasone	3
Radiation Treatment	3
Antiemetics and Antinauseants	3
Hormone Treatment	1
Endocrine Therapy	1
Zoledronic Acid	1
Capecitabine	1
Antimetabolites	1
Anastrozole	1

Table 2. Top 12 factors for predicting 5-year survival

Random Forest Model

As our initial classification tree approach appeared quite promising, we continued to explore modifications to our classifier. In particular, we identified random forest as an approach that could enhance the performance by growing multiple classification trees simultaneously and making class predictions based on weighted "votes" of each tree. One particular problem that decision trees often face is overfitting as a result of learning trees with very large heights. In our specific implementation of a decision tree, we chose to prune our tree to prevent overfitting, but

as a result, we sacrificed predictive accuracy. Through the weighted average approach, random forest classifiers address the overfitting issue through a more sophisticated mechanism than pruning, and as a result, they often achieve better classification performance. The ROC curve of the random forest classifier generated using the 'randomForest' R package[7] is shown in Figure 3, and a variable importance plot is shown in Figure 4.

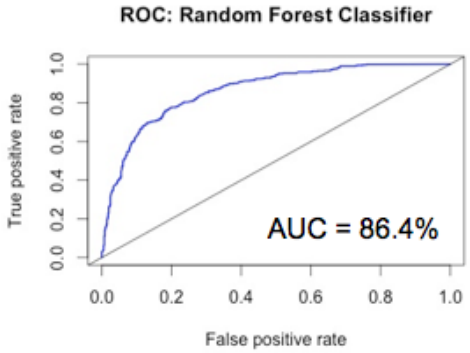

Figure 3. Receiver Operating Characteristic (ROC) curve of random forest classifier

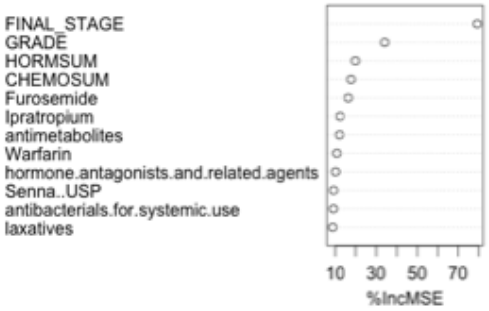

Figure 4. Variable importance plot with respect to incremental increase in prediction accuracy

5. Analysis of Top Features

From the results in the previous section, we can see that there are indeed several key features in our data that have much more predictive power than others. Though the list contains non-treatment controls (final stage and tumor grade), we are still able to compile a list of the drugs/treatments with the highest impact on survival. Our final step will thus be to confirm that these features are indeed enough to make an accurate prediction on patient survival by themselves.

We will do so by making use of Gradient Boosted Decision Trees as a classifier to analyze the predictive power of our high-weighted feature subset. Though the Gradient Boosted Tree (GBT) algorithm is similar to that of Random Forests, the crucial difference is that Random Forests maintain an ensemble of trees in parallel to vote on a process, while Gradient Boosted Trees are used in series to iteratively construct a prediction function.[8] This allows GBT to make predictions much more efficiently than Random Forests can, making it a more suitable algorithm to use.

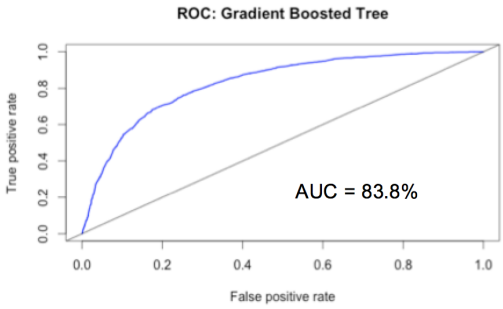

Figure 5. Receiver Operating Characteristic (ROC) curve of gradient boosted tree

Using the 'xgboost' package for R[9], we performed the Gradient Boosted Trees algorithm on the top ten treatments found in the previous section, with the resulting ROC curve shown in Figure 5. With a comparable AUC value to the result from the Random Forest algorithm (and even greater than that of logistic regression), we can see that limiting ourselves to these 10 features is indeed sufficient to make a fairly accurate prediction on a patient.

6. Conclusions and Future Directions

From the various analyses above, we can first conclude that our data set cannot be explained by a simple linear model as we had hoped. More importantly, however, we can conclude that, although the survival of a patient after five years is fairly consistent with the overall set of features given, there is a small group of drugs/treatments that is extremely predictive. In fact, the subset of ten treatments found above is enough to make predictions that are about as accurate as using the entire feature set. This result does not only enable us to build an efficient classifier without random noise, but also allows doctors to make an accurate prediction with a very limited amount of knowledge about the patient.

With these conclusions in mind, we can set our sights towards future analyses that can help us further understand the relationships between our features and the patient's outcome. Specifically, we can further study the relevance of our features by running our Classification Tree and Random Forest algorithms repeatedly while removing the top ten features after every run, allowing us to generate a gradient of feature importance. To further optimize our classifier, we can also run clustering algorithms and Principal Component Analysis between the features to look for strong correlations that can be removed. Our end goal of developing an accurate classifier for 5-year survival could further the management of care for breast cancer patients by identifying those most at risk of adverse events.

References:

1. "U.S. Breast Cancer Statistics. http://www.breastcancer.org/symptoms/understand_bc/statistics
2. Friedman, Jerome H. et al. "Regularization Paths for Generalized Linear Models via Coordinate Descent". http://www.jstatsoft.org/v33/i01/
3. Meyer, David et al. "Package 'e1071'". http://cran.r-project.org/web/packages/e1071/e1071.pdf
4. Breiman L., Friedman J. H., Olshen R. A., and Stone, C. J. (1984) Classification and Regression Trees. Wadsworth.
5. "Classification and Regression Trees (CART) Theory and Applications". http://edoc.hu-berlin.de/master/timofeev-roman-2004-12-20/PDF/timofeev.pdf
6. Therneau, Terry et al. "Package 'rpart'". http://cran.r-project.org/web/packages/rpart/rpart.pdf
7. Breiman, Leo et al. "Package 'randomForest'". http://cran.r-project.org/web/packages/randomForest/randomForest.pdf
8. "Introduction to Boosting Trees for Regression and Classification". http://www.statsoft.com/Textbook/Boosting-Trees-Regression-Classification
9. Chen, Tianqi. "xgboost". http://cran.r-project.org/web/packages/xgboost/index.html