Using Machine Learning Algorithms in R for Breast Cancer Predictions

James Frye

Repository link : https://github.com/JFaustus/Breast-Cancer-Machine-Learning

Abstract

Breast Cancer is a common cause of mortality in females and has been the case since ancient greece. Being able to consistently take tests, and identify whether a patient is at risk of having breast cancer can significantly reduce the chances of it being fatal. In fact, finding it early can increase your survival rating to 99%, which means having accurate tests and consistent testing is extremely important. [1] This study into breast cancer uses 4 models of classification and supervised learning in order to find an accurate model that can help identify the cancer quickly. The models I will be using in this study are the following, Support Vector Machine (SVM), Random Forest, Logistic Regression, and the generalized logistic regression net. These four models are pivotal in correctly identifying whether a tumor will be "malignant", or "benign" from the dataset. Using these datasets I will show and prove that based on the dataset I have acquired that the accuracy for the model will be able to hit at 98% accuracy. The dataset is based on the Wisconsin Breast Cancer dataset, and can be found on Kaggle. https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset. The main objective of this paper is to predict and diagnose breast cancer, using these models, and find the most effective one using a confusion matrix, ROC, F1, and other metrics. This will be done using R, Rstudio, and the caret package.

1. Introduction

According to breastcancer.org breast cancer accounts for 12.5% of all new annual cancer cases worldwide, making it the most common cancer in the world. Each year about 30% of all newly diagnosed cancers in women are breast cancer. Approximately 13% of U.S women are going to develop invasive breast cancer in the course of their life.[2] This will only continue to be the case and increase as the population increases and more women are subject to this type of cancer. This reinforces the idea that between how many cases happen a year, that it's important to get tested and have an accurate model for prevention. This goes hand in hand with the previous idea that if it is detected early then women will have a 99% chance to survive.
Having the chance to use data science and machine learning in order to detect cases of breast cancer will make this solution much easier if the models used are at the right accuracy. Hospitals in general have access to a ton of useful data and datasets large enough that looking through it wouldn't be able to tell you much. With the rise of big data sets and machine learning this data will become more useful over time as better models are made, and other important data is discovered that can identify as predictor variables. As I said previously this is the case for the model I have worked on for this study as it has achieved a 98.3% accuracy overall in the GLM net model. Being able to have a model with a high accuracy as this can help significantly in the diagnosis of breast cancer, but also other diseases and help prevent and treat these diseases given a strong enough model. The models I chose to use in this study are among the most common classification types when dealing with categorical variables such as identifying something as being malignant or benign. The rest of the sections in this document will be about the implementation and accuracy of said models and how I have arrived at the different metrics for them. Section 2 will discuss the different models and their accuracy compared to others on the subject, section 3 will discuss the differences between the models and lastly section 4 will be the conclusion of the models.

2. Papers with similar models

Looking at different studies involving breast cancer you will see similar models being used comparatively to the ones I will be using in this study. Among them the common models seem to be, support vector machines (SVM) and Logistic Regression which are used in every paper I have read on the subject so far. Mohammed Amine Naji used a support vector machine in order to find an accuracy of 97.2% using python and its Scikit-learn library. [3]

Another author named Maryam Panahiazar, Nolan Chen, Dmytro Lituiev, and Dexter Hadley used a mammogram data set used the same models, and ended up with its AUC being calculated for each outcome (Right Positive = 0.989, Left Positive = 0.994, Negative = 0.987, Bilateral Positive = 0.983, and Weighted Avg. = 0.989). This accuracy was quite high for a dataset of over 10000 pathology reports. [4]. This was done using a logistic regression model or glm model.

This is an extremely common model used any time there are categorical variables that put out a binary result of 0 or 1. Since they used a cancer risk score of 0 or 1 this model seems to be common sense to implement for this type of data.

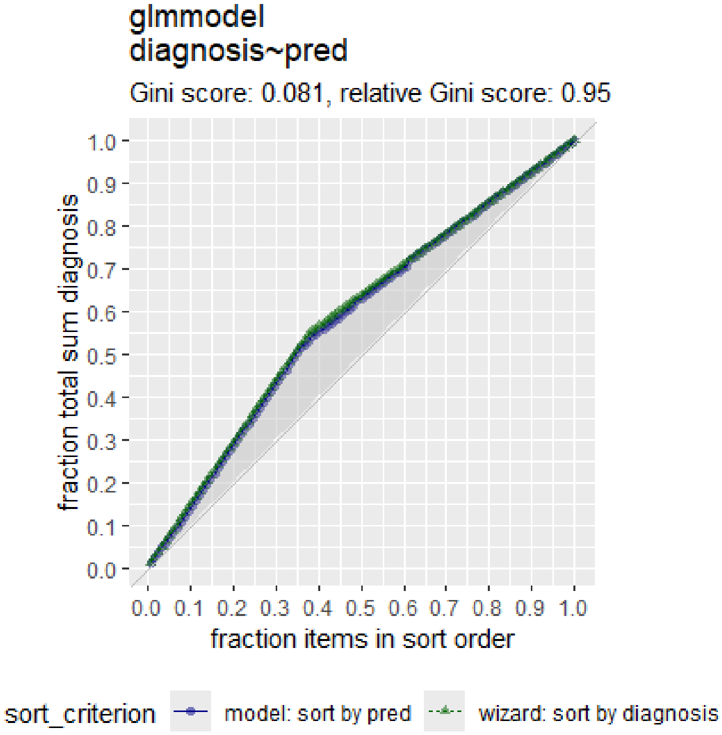
Muhammad Umer decided to use Convoluted features with ensemble machine learning while looking into breast cancer diagnosis. His model achieved an accuracy rating of 100%. He compared his model to 7 others all of varying accuracies ranging from 95%-99%. Using a data set of 45% malignant, and 55% benign, with 32 attributes was able to use the neural network model to achieve his accuracy. [5]

3. Use of Models

To start this off I used the dataset Breast Cancer from Kaggle located here https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset. This shows a dataset with 32 variables, and 569 observations. This data was split with the total coming out to be 357(62.7%) for benign, and 212(37.2%) for malignant. With this data I first started off with a logistic regression model three different ways in order to make sure the data was accurate. First off was using the tidymodels, and tidyverse packages to create a recipe to make the model. This ended up giving me a confusion matrix with an accuracy of 95.6% overall. Afterwards came the normal model made using glm, but the difference here being that the model is made to use its prediction data and pick anything that got factored to over 50% and group it up according to the data. This resulted in an accuracy of 95.6% which is the exact same. The confusion matrix for this model was also the same.

Now something to keep in mind is that with a classification problem such as this the important metrics to keep in mind are the accuracy but also the sensitivity, specificity, Precision, Recall, and F1 values. Each value specifies something useful to learn with the precision and recall values being important for F1. The F1 score determines the average of the precision and recall scores where a score of 1 means all the predictions are correct. [6] This is determined by taking in information on all of the errors made in a confusion matrix. The F1 value for this model is 96.5% which is a fantastic score. The graph below is useful in figuring out how closely the data relates and is another metric in figuring out the accuracy. Its close to 0 so it is also fairly accurate and the lines intersect each other extremely well. The next model was done in caret using the glm model as well and had the exact same outputs. All metrics I've used indicate that this model is very accurate.

Gain Curve plot for GLM

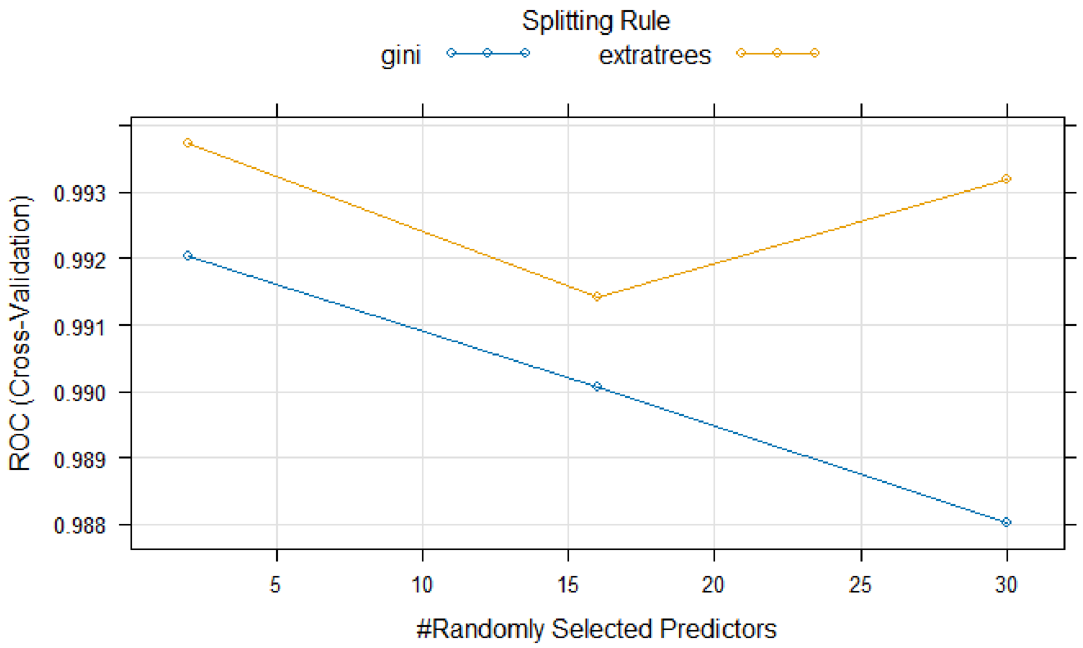


The next model used was the random forest model. What this does is it takes in the dataset, puts them into different groups and has each "expert" vote on a final decision. These subsets of data each contribute to the overall accuracy at the end along with its predictions. [7] My model has a different confusion matrix compared to the glm model and was able to have an accuracy of 99.1% with an f1 value of 99.3%. The graph below shows the model being plotted at 3 different points being, 2, 16, and 30. The parameters ended up being the best as you can see from 16 having the highest ROC and then taking the average you'll get the accuracy overall. This model's accuracy is extremely high for it, and shows it in the data as well with high metrics across the board.

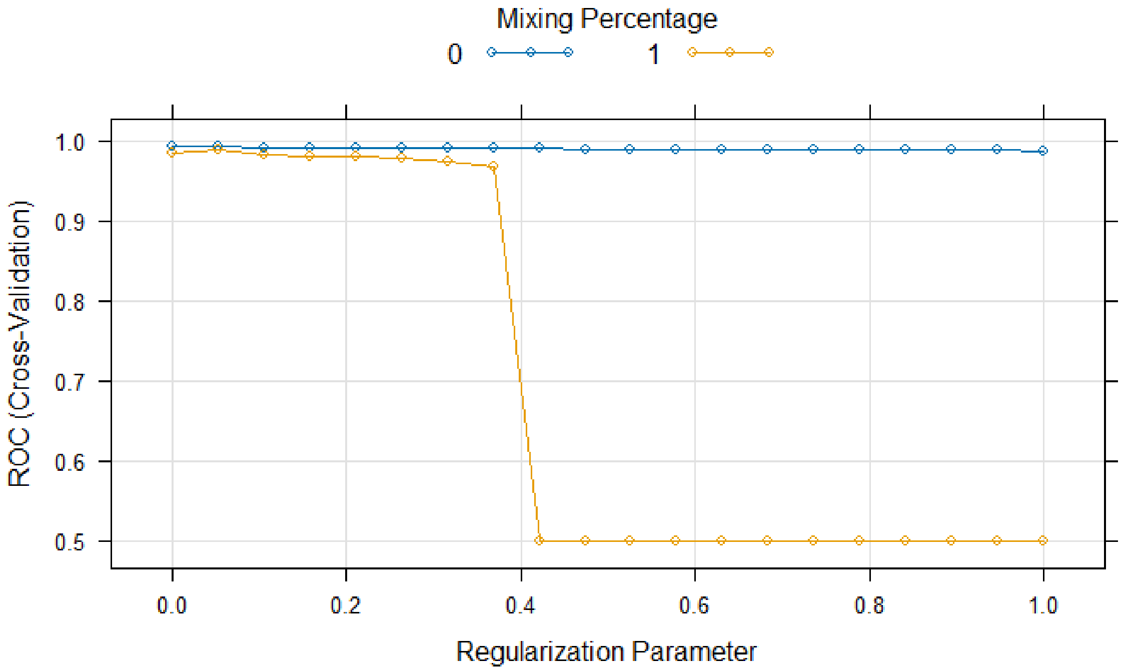
The next is the SVM or support vector machine model, which takes data points from the dataset, and outputs a hyperplane which is simply a line. This line separates each of these data points into groups either blue or red. Using this hyperplane it can easily categorize the different data points and find out the accuracy. [8] My SVM model was able to hit an accuracy of 95.6% and an F1 value of 96.5% which is pretty much on the money for the same accuracy as the glm model. Which has a similar confusion matrix as well.

This model I used was called the glmnet model, and it's a package that fits generalized linear and similar models via a penalized maximum likelihood. It uses regularization as a method to prevent overfitting which also uses the lasso, ridge, and elastic net regression to do this. [9] Using this model I was able to achieve an accuracy of 98.3% and an F1 score of 98.6%.

Random Forest Graph



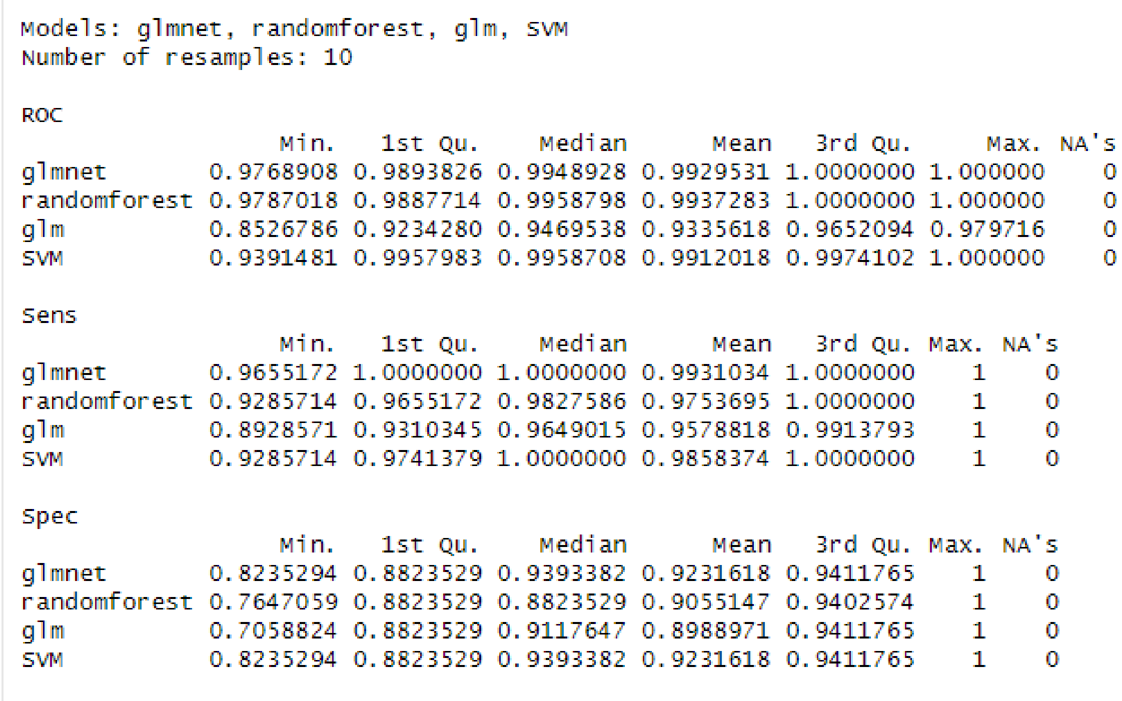
GLMnet graph



4. Conclusion.

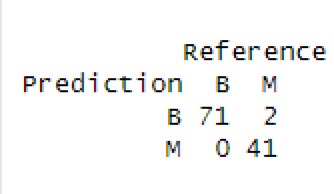
On average the random forest model predicts a better accuracy but can dip under at times, but in terms of having a consistent accuracy the glmnet model ended up on top. It has a high accuracy to go along with high metrics across the board. There are plenty of ways to get this modeled in other ways to increase accuracy, or even feature clean up on the data set itself. Because this model is smaller than some others the accuracy of the models is a bit easier to work with. I hope I can work with bigger datasets later in order to find ways to work out better models when given these big datasets. I aim to make a tool that will be able to be developed to closely and accurately show that a patient has breast cancer.

Comparison table



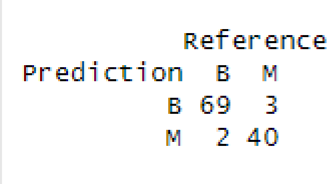
```
Models: glmnet, randomforest, glm, SVM
Number of resamples: 10

ROC
                  Min.   1st Qu.    Median      Mean   3rd Qu.     Max. NA's
glmnet       0.9768908 0.9893826 0.9948928 0.9929531 1.0000000 1.000000    0
randomforest 0.9787018 0.9887714 0.9958798 0.9937283 1.0000000 1.000000    0
glm          0.8526786 0.9234280 0.9469538 0.9335618 0.9652094 0.979716    0
SVM          0.9391481 0.9957983 0.9958708 0.9912018 0.9974102 1.000000    0

Sens
                  Min.   1st Qu.    Median      Mean   3rd Qu. Max. NA's
glmnet       0.9655172 1.0000000 1.0000000 0.9931034 1.0000000    1    0
randomforest 0.9285714 0.9655172 0.9827586 0.9753695 1.0000000    1    0
glm          0.8928571 0.9310345 0.9649015 0.9578818 0.9913793    1    0
SVM          0.9285714 0.9741379 1.0000000 0.9858374 1.0000000    1    0

Spec
                  Min.   1st Qu.    Median      Mean   3rd Qu. Max. NA's
glmnet       0.8235294 0.8823529 0.9393382 0.9231618 0.9411765    1    0
randomforest 0.7647059 0.8823529 0.8823529 0.9055147 0.9402574    1    0
glm          0.7058824 0.8823529 0.9117647 0.8988971 0.9411765    1    0
SVM          0.8235294 0.8823529 0.9393382 0.9231618 0.9411765    1    0
```

Confusion matrix for GLMnet, and random forest



```
          Reference
Prediction  B  M
         B 71  2
         M  0 41
```

Confusion for SVM and GLM



```
          Reference
Prediction  B  M
         B 69  3
         M  2 40
```

References

[1] Health, C. (2022, October 5). *Why it's so important to get regular breast cancer screenings*. health. https://health.ucdavis.edu/blog/cultivating-health/why-its-so-important-to-get-regular-breast-cancer-screenings/2022/10#:~:text=It's%20important%20to%20get%20regular,relative%20survival%20rate%20is%2099%25.

[2] *Breast cancer facts and statistics 2024*. (n.d.). https://www.breastcancer.org/facts-statistics

[3] Naji, Mohammed Amine. "Machine Learning Algorithms For Breast Cancer Prediction and Diagnosis." *University of Mons, mons*, Procedia, 2021.

[4] Panahiazar, M., Chen, N., Lituiev, D., & Hadley, D. (2022). Empowering study of breast cancer data with application of artificial intelligence technology: promises, challenges, and use cases. *Clinical & experimental metastasis*, *39*(1), 249–254. https://doi.org/10.1007/s10585-021-10125-8

[5] Umer, M., Naveed, M., Alrowais, F., Ishaq, A., Hejaili, A. A., Alsubai, S., Eshmawi, A. A., Mohamed, A., & Ashraf, I. (2022). "Breast Cancer Detection Using Convoluted Features and Ensemble Machine Learning Algorithm". *Cancers*, *14*(23), 6015. https://doi.org/10.3390/cancers14236015

[6] Sharma, Natashsa. "Understanding and Applying F1 Score: Ai Evaluation Essentials with Hands-on Coding Example." *Arize AI*, 8 Apr. 2024, arize.com/blog-course/f1-score/.

[7] Shafi, Adam. "Random Forest Classification with Scikit-Learn." *DataCamp*, DataCamp, 24 Feb. 2023, www.datacamp.com/tutorial/random-forests-classifier-python.

[8] Le, James. "Support Vector Machines in R Tutorial." *DataCamp*, DataCamp, 22 Aug. 2018, www.datacamp.com/tutorial/support-vector-machines-r.

[9] Hastie, Trevor, et al. "An Introduction to \`glmnet\`." *An Introduction to \`glmnet\` • Glmnet*, 27 Mar. 2023, glmnet.stanford.edu/articles/glmnet.html#:~:text=Glmnet%20is%20a%20package%20that,for%20the%20regularization%20parameter%20lambda.