A Comparison of Logistic Regression(LR) and Random Forest(RF) Applied for Breast Cancer Prediction

Asha Guruvayurappan | City University of London

Description and Motivation

Breast cancer is the most commonly diagnosed cancer and the leading cause of cancer death in woman[1]

- Construct and compare two models for a binary classification problem to predict Breast Cancer type.
- Investigate the efficiency of logistic regression and random forest models in predicting whether a cancer is benign or malignant
- Compare the results of similar implementations obtained by **Adel S. Assiri , Saima Nazir and Sergio A. Velastin in Breast Tumor Classification Using an Ensemble Machine Learning Method[2]**

Data Exploration

- The data set Breast Cancer Wisconsin (Diagnostic) is obtained from Kaggle.
- The data set comprises of 569 instances, which are unique patient records. Each instance consist of 31 features that describe the characteristics of the cell nuclei of breast mass and 1 target which classify if the cancer is **benign (b) or malignant (m)**, Figure 1.
- The data set had one column with unique patient IDs and one column with only null values, these columns were dropped as they had no impact on the dataset.
- The target column 'diagnosis' is converted to 0 for benign and 1 for malignant. Figure 1 shows the distribution of target column.
- When analysing the features, by diving into groups with respect to mean, standard error and "worst" or largest (mean of the three largest values), There are several corelated predictors[3]; Figure 2.
- The mean area of the tissue nucleus has a strong positive correlation with mean values of radius and perimeter(r between 1-0.75); Some parameters are moderately positive correlated (r between 0.5-0.75), those are concavity and area, concavity and perimeter etc; Likewise, we see few negative correlations between fractal-dimension with radius, texture values.
- A similar trend is observed in 'worst' value parameters.
- Analysing feature mean to its feature worst; Figure 3, we can easily distinguish between Benign and Malignant from the scatter plot. Most of the benign observations are centred in the lower left quadrant while the malignant observations are centred in the upper right quadrant, also these variable interactions have an almost linear relationship.
- When the mean radius, concavity or area increases the probability of cancer being malignant increases. Figure 3

Feature Engineering and Transformation

- Preliminary model evaluation indicated that no standardization is required, and the data as it is yielded best results.
- All features follow a normal distribution except few. Visualizing data showed few outliers but on analysis they did not skew with the model results and so no reduction is performed in this study.
- Figure 1 shows that the dataset is imbalanced with only 212 Malignant cases and 357 Benign cases, in addition to that the total number of records in the data set is small. To enlarge the size of minority class and the size of overall dataset Synthetic minority over-sampling technique (SMOTE) is used on the training set. SMOTE explanation and steps are outlined in SMOTE for Learning from Imbalanced Data by Fernandez, A.; Garcia, S.; Herrera, F.; Chawla, N.V [4]
- To avoid overfitting of model Gaussian noise is added to training data. [5]

Model Selection

Logistic Regression(LR)

- Logistic regression is the go-to method for a supervised binary classification problem (classifying cancer type to be Benign or Malignant)
- The paper **Breast Tumor Classification Using an Ensemble Machine Learning Method by Adel S. Assiri , Saima Nazir and Sergio A. Velastin[2],** concludes that Logistic Regression is one of the most accurate machine learning models for this problem.

How it works?

- The underlying technique of LR is same as linear regression. Linear regression would only linearly separate data, this will not work for every problem, and will skew the results in presence of outliers.
- Logistic Regression(LR) uses a Sigmoid function, which takes any real values between 0 and 1 forming a ***S*** curve.

Advantages

- Easy to implement and interpret.
- Efficient performance for classification problems
- Logistic regression is less inclined to overfitting a model

Disadvantages

- If number of observations are lesser, LR skews the results.
- The major limitation of LR is the assumption of linearity between the dependent variable and the independent variables

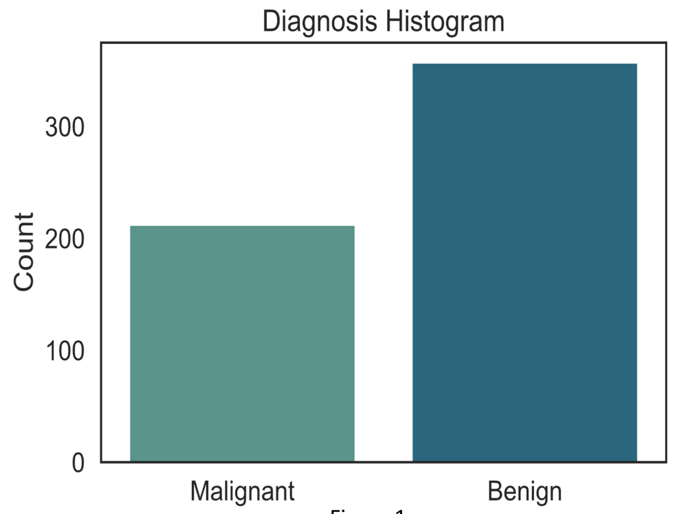

Figure 1

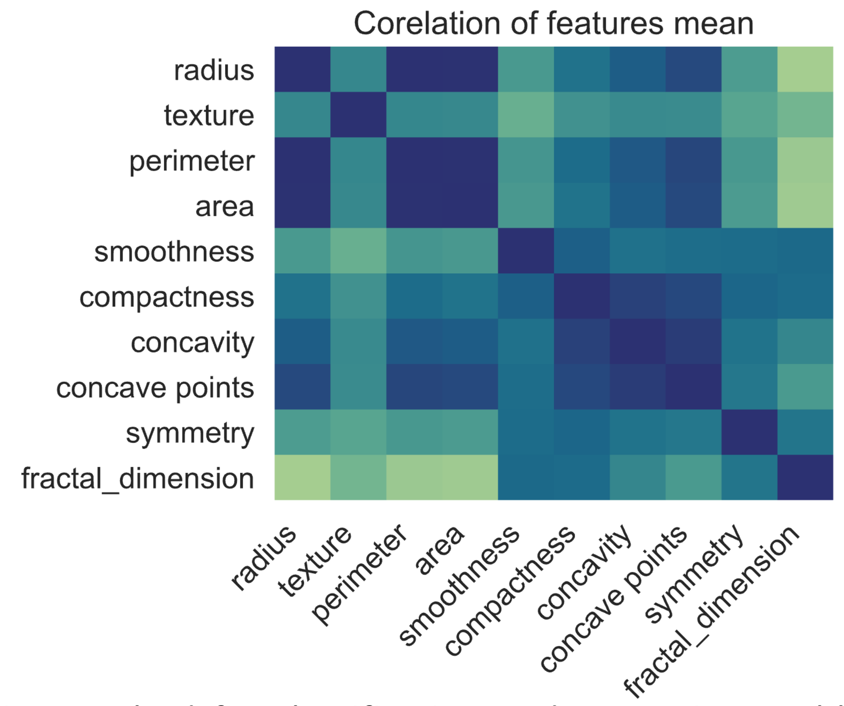

Figure 2

Random Forest (RF)

- RF is a supervised, ensemble learning method for classification and regression problems. It is an extension of bootstrap aggregation(bagging) of decision trees .
- TreeBagger grows the decision trees in the ensemble using bootstrap samples of the data. Also, TreeBagger selects a random subset of predictors to use at each decision split as in the random forest algorithm[6]

How it works?

- Random forest combines multiple Decision Trees to reach a single result. It uses bagging and feature importance when building individual trees and create an uncorrelated forest to obtain an accurate and stable prediction.

Advantages

- RF is a flexible and easy to use machine learning algorithm
- It can be used on both classification and regression problems
- it reduces overfitting in decision trees and helps to improve the accuracy.

Disadvantages

- Training time of RF is much more when compared to any other models as it has to train and combine numerous trees
- It requires much computational power.
- **They are unstable**, meaning that a small change in the data can lead to a large change in the structure of the optimal decision tree

Hypothesis

- Logistic regression will have lower training time when compared to Random Forest.
- Logistic regression is expected to perform better than RF as specified in **Breast Tumor Classification Using an Ensemble Machine Learning Method[2]**
- Feature reduction (using algorithms) should improve accuracy of both the models.

Model Training and Evaluation

- The original dataset is split into 70% - 30% as training – testing sets resulting in 399 training and 170 testing records.
- SMOTE technique is imposed on the training set to balance and increase the training records[4]. SMOTE sampling can benefit feature selection which can be applied to reduce feature dimensions[7]
- Versus the original published study[2], Gaussian Noise is induced onto training set to avoid overfitting of models[5] and to evaluate the accuracy with noise.

Logistic Regression(LR)

- A Binomially distributed model created for LR was Cross validated by 10-fold
- The goal is to represent the target class's posterior probabilities as linear functions of the predictors, ensuring that they sum to one and stay within the range of 0 to 1. Setting a threshold value allows us to determine the expected target class from these probabilities (e.g., 0.5).
- Average accuracy(Cross validated by 10-fold) after performing Lasso Regularization for feature did not have a huge impact on the model when compared to all features.
- Although the AUC for train and test data was high, the percentage of predicted malignant cases were low in tested data.

Random Forest (RF)

- Bootstrap-aggregated (*bagged*) decision trees combine the results of many decision trees, which reduces the effects of overfitting and improves generalization
- Best results were found at the following hyperparameters: 1000 trees and minimum leaf size of 5 by iterating through different combinations of tree and leaf size. Figure 4
- Identifying important predictors above threshold of 0.5 had a slight improvement on the model.
- Feature selection, tree size and leaf size were obtained by validating out-of-bag records in the training data.

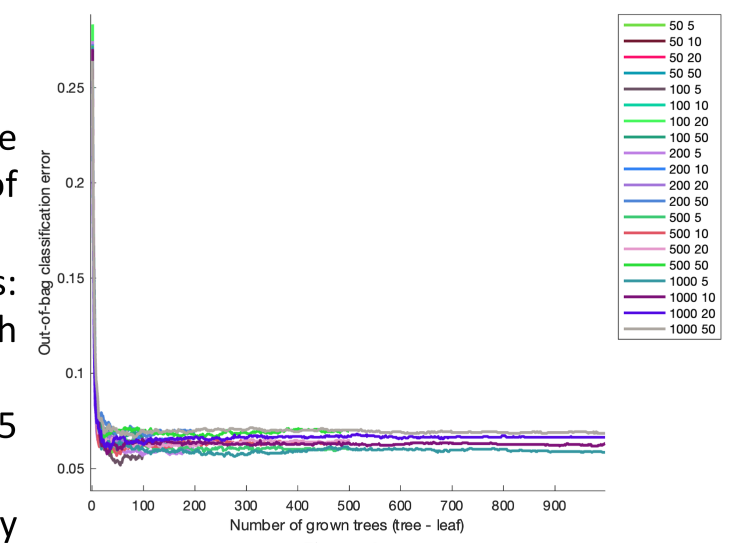

Figure 4

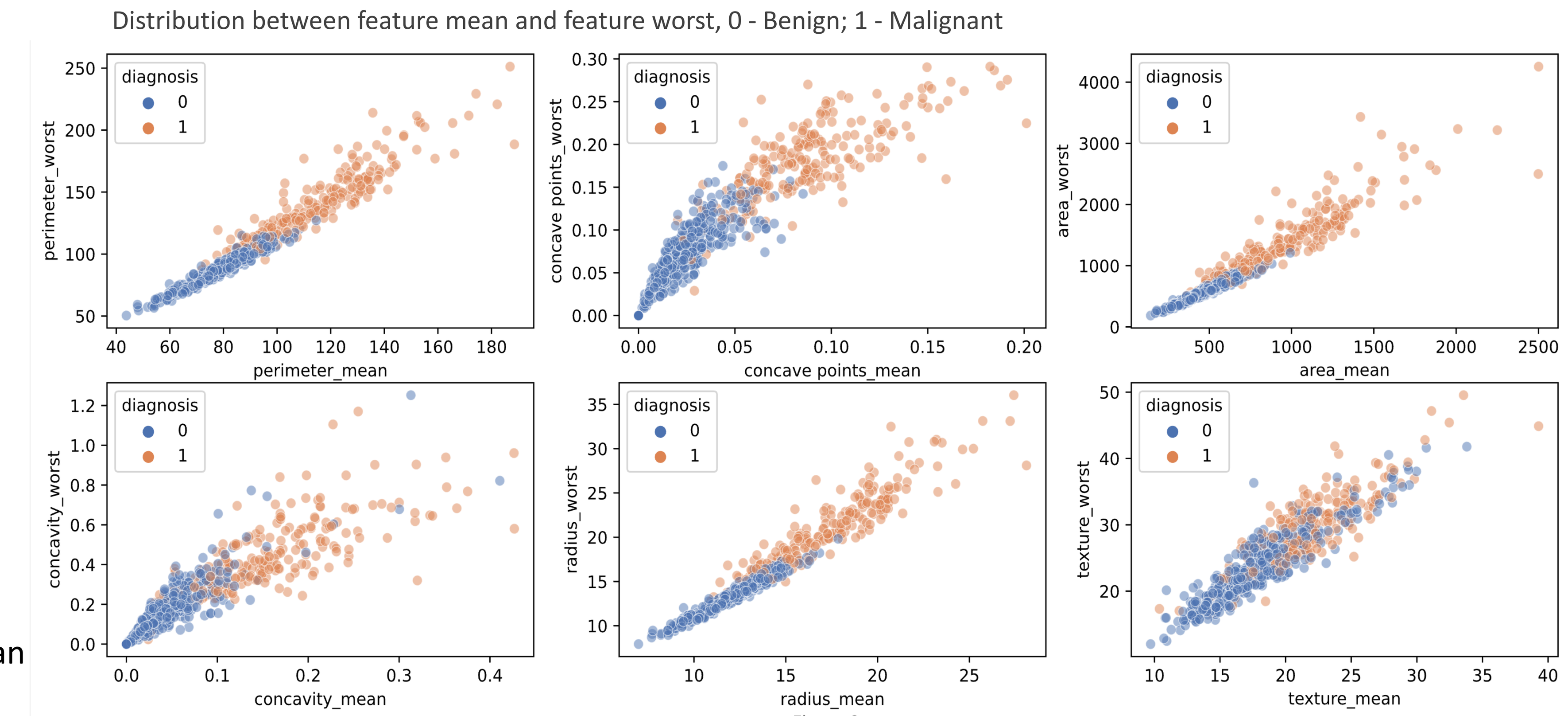

Figure 3

Result Analysis

- Both LR and RF performed well on basic data, but addition of noise reduced the performance of LR.
- Selecting lasso hyperparameters for LR had only a minimal impact on the model, the model with all features gave a high accuracy on cross validation which contradicts the assumption.
- The best tree and leaf size for RF had a great impact model which was identified from out-of-bag errors, also tuning the hyperparameters improved model performance as suggested.
- Random Forest gave 92.35% accuracy on test (unseen) data; Logistic Regression regression obtained 91.76% accuracy which counter our assumption, Figure 7-8
- The original study**[2]** implied that Logistic regression is the best classification model for this dataset with an accuracy of 98.25%, but the results obtained are contradicting to this study; adding Gaussian noise decreased the performance of LR drastically.
- The study suggested that Random Forest obtained 96.49% accuracy[2], the model even with noise over data produced an accuracy much better than LR on test set (92.35%).
- The Recall percentage for predicting Malignant cases are lower in both models. When compared among the two, recall of RF is better than LR for predicting malignant cases.
- LR training time was much lower when compared to RF as expected, but also prediction time for unseen test data was lower in LR.
- Logistic regression and Random Forests largely agree with each other in variable selection. The important variables selected by these two methods were almost the same.
- However, test AUC of LR is marginally higher when compared to RD, this could be due to random split of train-test data.

	Logistic Regression		Random Forest	
	B	M	B	M
Precision	0.89	0.98	0.91	0.94
Recall	0.99	0.79	0.97	0.84
F1 Score	0.93	0.87	0.94	0.89
AUC Test	0.991		0.978	
Training Acc.	94.03		95.61	
Test Acc.	91.76		92.35	
Training Time (Avg.)	0.043s		31.28s	
Testing Time(Avg.)	0.02s		0.77s	

Figure 5

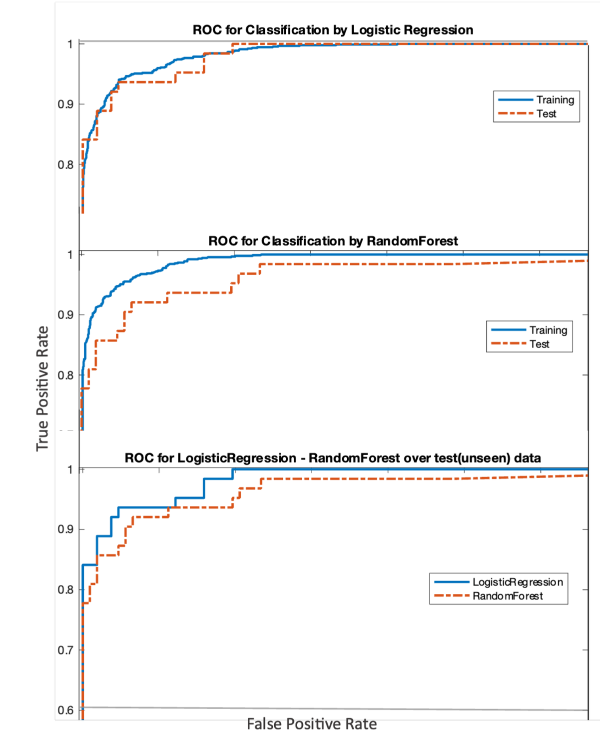

Figure 6

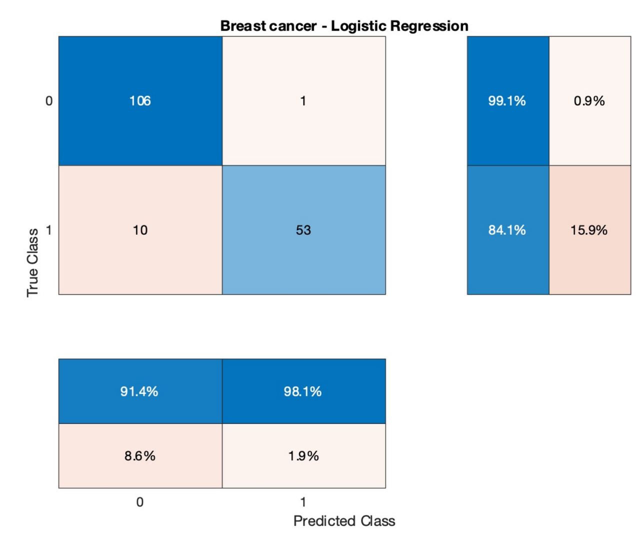

Figure 7

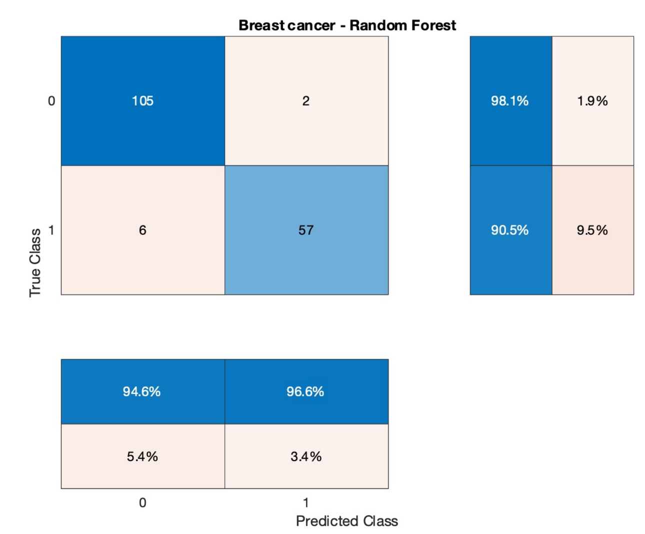

Figure 8

Lessons Learnt

- RF is a more stable model when compared to LR, as noise over data drastically decrease model performance for LR.
- Hyperparameter optimization might not always have a positive impact on model for LR.
- Although the accuracy of overall model is high; recall for malignant cases(model predicting malignant cases to be benign) are low which would have consequence in real world.

Future Directions

- Both training and test data had no missing values, identify how the models would perform in case of presence of missing values. If the performance reduces; what can be done further to improve it.
- Try and train with a large number of records as the current dataset had only 569 records, also try to improve the features of dataset by integrating more patient information like age, sex.
- Try to improve prediction of malignant cases(recall)

Refences

[1] Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries; by Hyuna Sung, PhD; Jacques Ferlay, MSc, ME2; Rebecca L. Siegel, MPH; Mathieu Laversanne, MSc; Isabelle Soerjomataram, MD, MSc, PhD; Ahmedin Jemal, DMV, PhD; Freddie Bray, BSc, MSc, PhD

[2] Breast Tumor Classification Using an Ensemble Machine Learning Method by Adel S. Assiri , Saima Nazir and Sergio A. Velastin ; Published: 29 May 2020

[3] Breast Cancer Classification Using Machine Learning Techniques: A Review ;by Srwa Hasan Abdulla, Ali Makki Sagheer, Hadi Veisi ; Published online: 19 August 2021

[4] Fernandez, A.; Garcia, S.; *Herrera, F.; Chawla, N.V. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary; Published. 2018.*

[5] Adding Noise for Robust Deep Neural Network Models; by Sovit Ranjan Rath ; Published February 3, 2020

[6] MathWorks United Kingdom. Uk.Mathworks.com ; ref : https://uk.mathworks.com/help/stats/treebagger.html

[7] Solanki, Y.; Chakrabarti, P.; Jasinski, M.; Leonowicz, Z.; Bolshev, V.; Vinogradov, A.; Jasinska, E.; Gono, R.; Nami, M. A Hybrid Supervised Machine Learning Classifier System for Breast Cancer Prognosis *Using Feature Selection and Data Imbalance Handling Approaches. Electronics 2021, 10, 699*

[8] Personalized analysis of breast cancer using sample-specific networks; by Ke Zhu, Cong Pian, Qiong Xiang, Xin Liu, Yuanyuan Chen; published on May 15, 2020

[9] Discovering the shades of Feature Selection Methods by Swetha Manoj— April 23, 2021