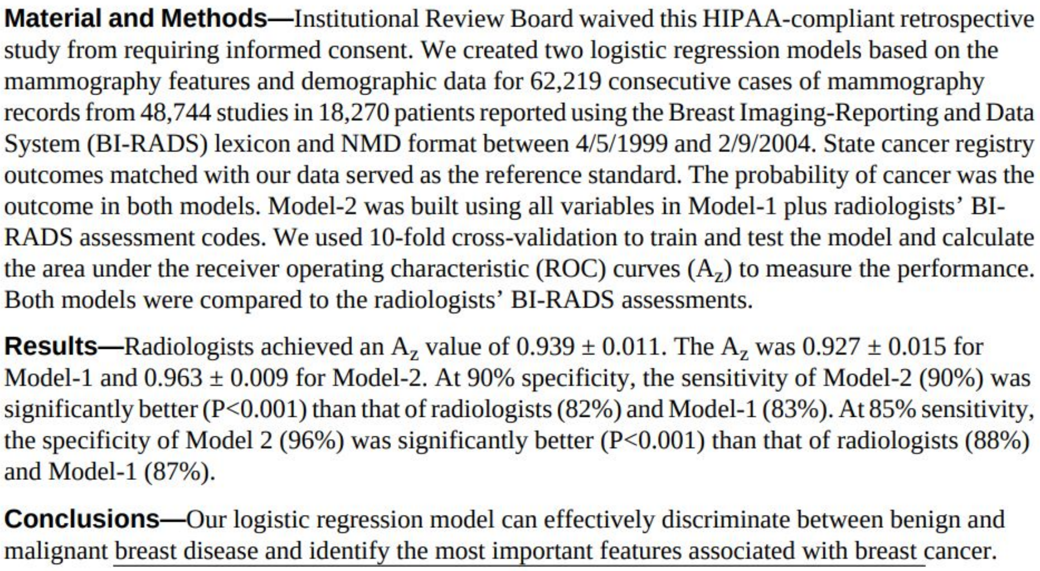
Day 2

Providing universal access to AI education and practice

ML IN SCIENTIFIC ARTICLE

A Logistic Regression Model Based on the National Mammography Database Format to Aid Breast Cancer Diagnosis

Jagpreet Chhatwal, MS, Oguzhan Alagoz, PhD, Mary J. Lindstrom, PhD, Charles E. Kahn Jr., MD, MS, Katherine A. Shaffer, MD, and Elizabeth S. Burnside, MD, MPH, MS



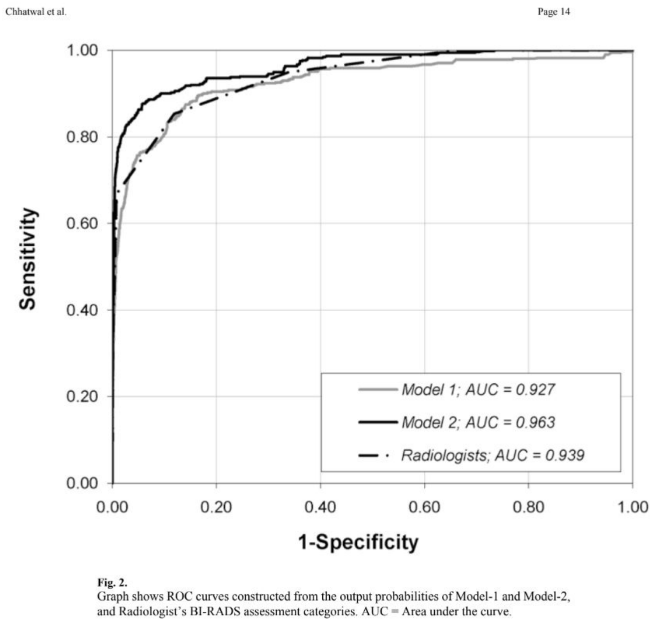
Material and Methods—Institutional Review Board waived this HIPAA-compliant retrospective study from requiring informed consent. We created two logistic regression models based on the mammography features and demographic data for 62,219 consecutive cases of mammography records from 48,744 studies in 18,270 patients reported using the Breast Imaging-Reporting and Data System (BI-RADS) lexicon and NMD format between 4/5/1999 and 2/9/2004. State cancer registry outcomes matched with our data served as the reference standard. The probability of cancer was the outcome in both models. Model-2 was built using all variables in Model-1 plus radiologists' BI-RADS assessment codes. We used 10-fold cross-validation to train and test the model and calculate the area under the receiver operating characteristic (ROC) curves (A_z) to measure the performance. Both models were compared to the radiologists' BI-RADS assessments.

Results—Radiologists achieved an A_z value of 0.939 ± 0.011. The A_z was 0.927 ± 0.015 for Model-1 and 0.963 ± 0.009 for Model-2. At 90% specificity, the sensitivity of Model-2 (90%) was significantly better ($P<0.001$) than that of radiologists (82%) and Model-1 (83%). At 85% sensitivity, the specificity of Model 2 (96%) was significantly better ($P<0.001$) than that of radiologists (88%) and Model-1 (87%).

Conclusions—Our logistic regression model can effectively discriminate between benign and malignant breast disease and identify the most important features associated with breast cancer.

A Logistic Regression Model Based on the National Mammography Database Format to Aid Breast Cancer Diagnosis

Jagpreet Chhatwal, MS, Oguzhan Alagoz, PhD, Mary J. Lindstrom, PhD, Charles E. Kahn Jr., MD, MS, Katherine A. Shaffer, MD, and Elizabeth S. Burnside, MD, MPH, MS



Chhatwal et al.

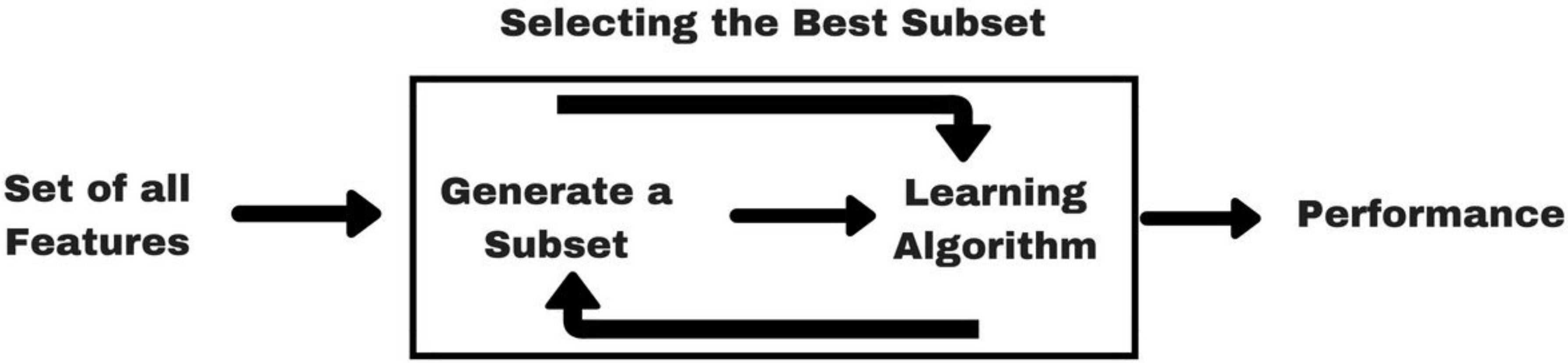
Page 14

Fig. 2.
Graph shows ROC curves constructed from the output probabilities of Model-1 and Model-2, and Radiologist's BI-RADS assessment categories. AUC = Area under the curve.

VARIABLE SELECTION

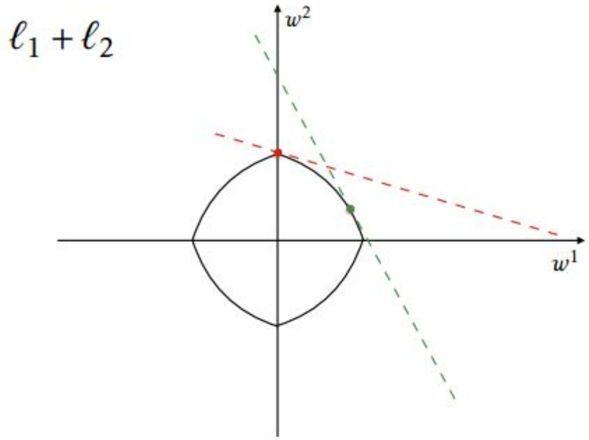
Step Forward Selection

Forward Selection: Step forward feature selection starts with the evaluation of each individual feature, and selects that which results in the best performing selected algorithm model. What's the "best?" That depends entirely on the defined evaluation criteria (AUC, prediction accuracy, RMSE, etc.). Next, all possible combinations of the that selected feature and a subsequent feature are evaluated, and a second feature is selected, and so on, until the required predefined number of features is selected.



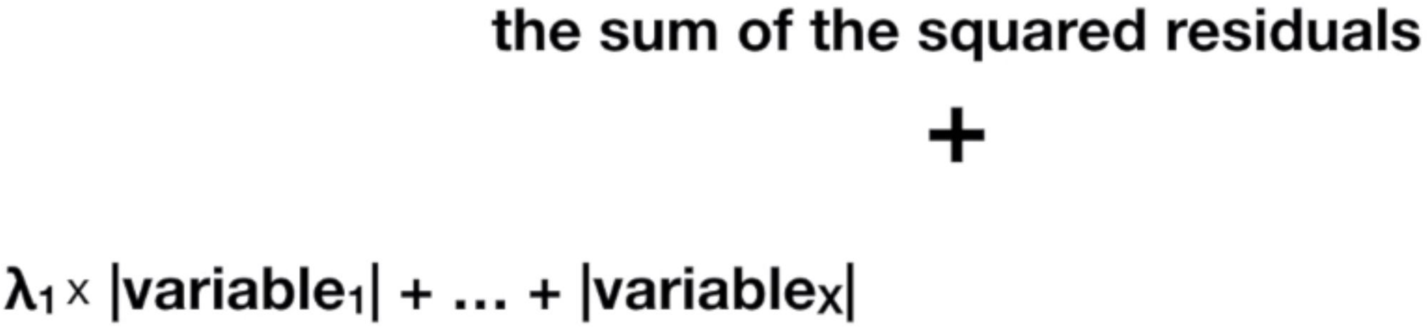
Regularization

A regularization technique is in simple terms a penalty mechanism which applies shrinkage (driving them closer to zero or to zero) of coefficient to build a more robust and parsimonious model.



Regularization: Lasso

L1 Regularization (aka Lasso Regularization) adds regularization terms in the model which are function of absolute value of the coefficients of parameters. The coefficient of the paratmeters can be driven to zero as well during the regularization process. Hence this technique can be used for feature selection and generating more parsimonious model.



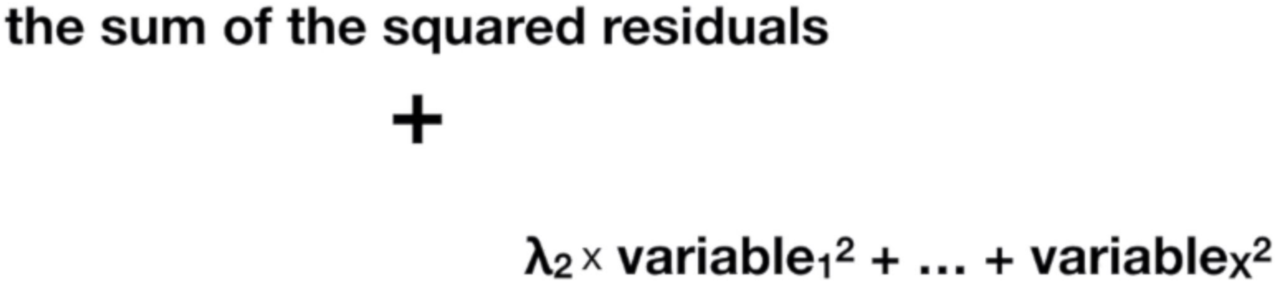
the sum of the squared residuals

+

$$\lambda_1 \times |\text{variable}_1| + \ldots + |\text{variable}_X|$$

Regularization: Ridge

L2 Regularization (aka Ridge Regularization) adds regularization terms in the model which are function of square of coefficients of parameters. Coefficient of parameters can approach to zero but never become zero and hence



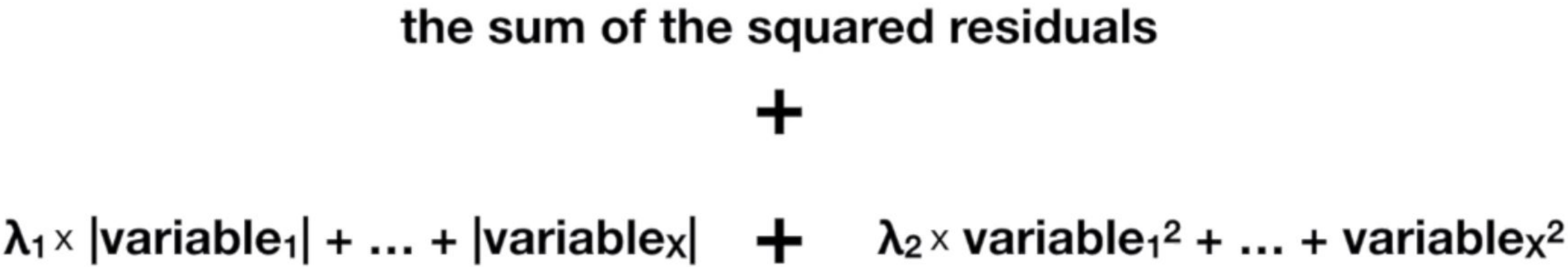
the sum of the squared residuals

+

$\lambda_2 \times \text{variable}_1^2 + \ldots + \text{variable}_X^2$

Regularization: ElasticNet

Elastic net regression combines the power of ridge and lasso regression into one algorithm. What this means is that with elastic net the algorithm can remove weak variables altogether as with lasso or to reduce them to close to zero as with ridge.



the sum of the squared residuals

$$+$$

$$\lambda_1 \times |\text{variable}_1| + \ldots + |\text{variable}_X| \quad + \quad \lambda_2 \times \text{variable}_1^2 + \ldots + \text{variable}_X^2$$