Different Method for Analyzing Risk Factor of Breast Cancer

Introduction

In this project, we are trying to answer what characteristics of cell nuclei from tumors might be the risk factor of breast cancer by a breast cancer data set on Kaggle. This breast cancer data set contains 30 predictors and 569 observations. The 30 predictors are three measurements of 10 tumors' features measuring by image: the mean, standard error, and worst(largest) of radius(mean of distances from the center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness(local variation in radius lengths), compactness (perimeter^2 / area - 1), concavity(severity of concave portions of the contour, concave points(number of concave portions of the contour), and symmetry, fractal dimension("coastline approximation" -1). Our primary outcome is whether the tumors are malignant or benign. We used 70% of it to train the models and 30% to test model performance.

Data Processing

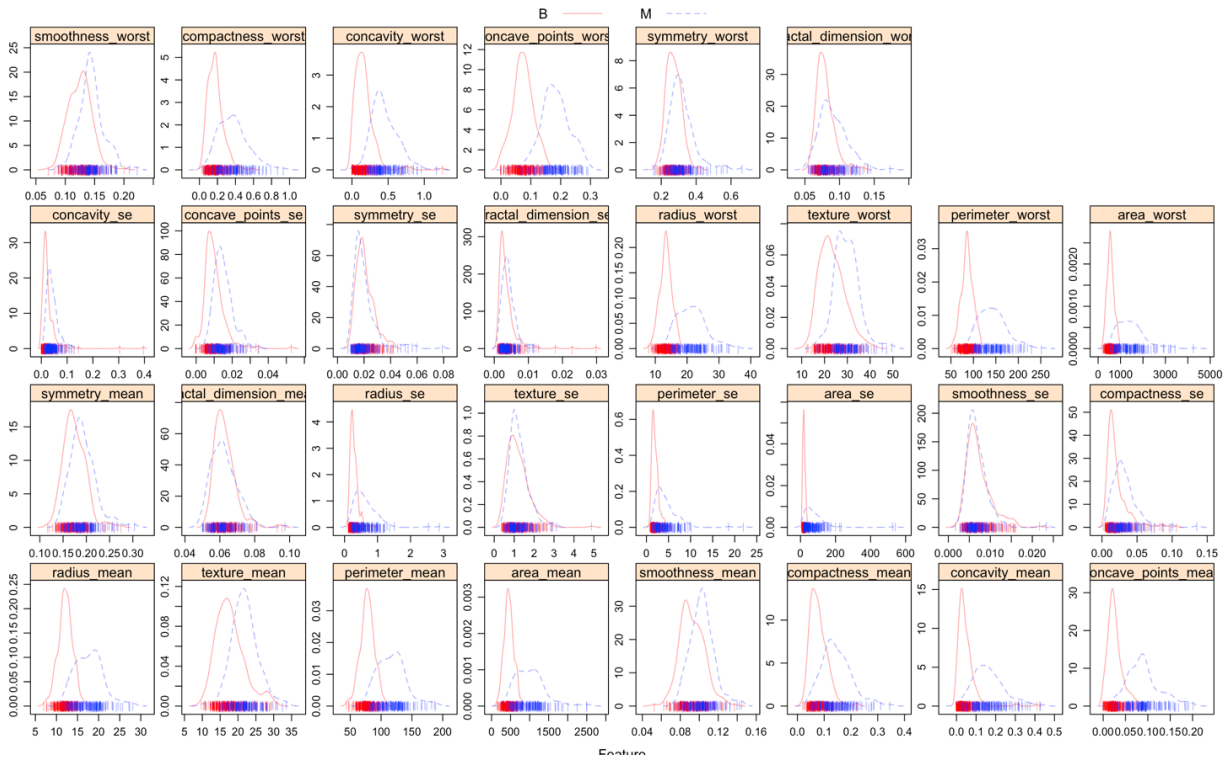
The data set contains no missing value. The only change we did was change the predictors into numeric class and outcome into factor class. Table 1 contains all the variables in this data.

Predictors	**Definition**
radius_mean	Radius of Lobes
texture_mean	Mean of Surface Texture
perimeter_mean	Outer Perimeter mean of Lobes
area_mean	Mean Area of Lobes
smoothness_mean	Mean of Smoothness Levels
compactness_mean	Mean of Compactness
concavity_mean	Mean of Concavity
concave points_mean	Mean of Cocave Points
symmetry_mean	Mean of Symmetry
fractal_dimension_mean	Mean of Fractal Dimension
radius_se	Standard error of Radius
texture_se	Standard error of Texture
perimeter_se	Standard error of Perimeter
area_se	Standard error of Area
smoothness_se	Standard error of Smoothness
compactness_se	Standard error of compactness
concavity_se	Standard error of concavity
concave points_se	Standard error of concave points
symmetry_se	Standard error of symmetry
fractal_dimension_se	Standard error of Fractal Dimension
radius_worst	Worst (largest) Radius
texture_worst	Worst (largest) Texture
perimeter_worst	Worst (largest) Permimeter
area_worst	Worst (largest) Area
smoothness_worst	Worst (largest) Smoothness
compactness_worst	Worst (largest) Compactness
concavity_worst	Worst (largest) Concavity
concave points_worst	Worst (largest) Concave Points
symmetry_worst	Worst (largest) Symmetry
fractal_dimension_worst	Worst (largest) Fractal Dimension

Table1, 30 predictors in this data

Methods

Exploratory analysis-Distribution of Variables grouped by outcome



Plot 1, Distribution of Variables grouped by outcome

Visualization was applied to understand the variables' distribution of malignant and benign tumors. By plot 1, it is obliviously that some variables are particularly significant in classifying malignant and benign tumors. Such as standard error of concave points (concave points_se), worst concave points (concave points_worst), worse compactness(compactness_worst), worst concavity (concavity_worst), mean of concavity(concavity_mean), mean of concave Points (concave points_mean). By this observation, it seems that concave points, compactness, concavity might be the important classifier for classifying malignant and benign tumor

Statistical Methods In this project, six classification methods were explored, including Logistics Regression, Penalized Logistic Regression, Multivariate Adaptive Regression Splines, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Naive Bayes. All of the models were trained by 10-fold cross-validation.

Model 1- Logistics Regression Considering that this data have 30 highly correlated predictors, logistics regression is not the best idea for this data. However, for the reference of our future models, a logistics regression is still built as a reference. Our final logistics regression model run by R has 0.9412 accuracy on test data which seems like a great prediction. However, none of the predictors in this model is significant (P value<0.05). A good predicted accuracy but

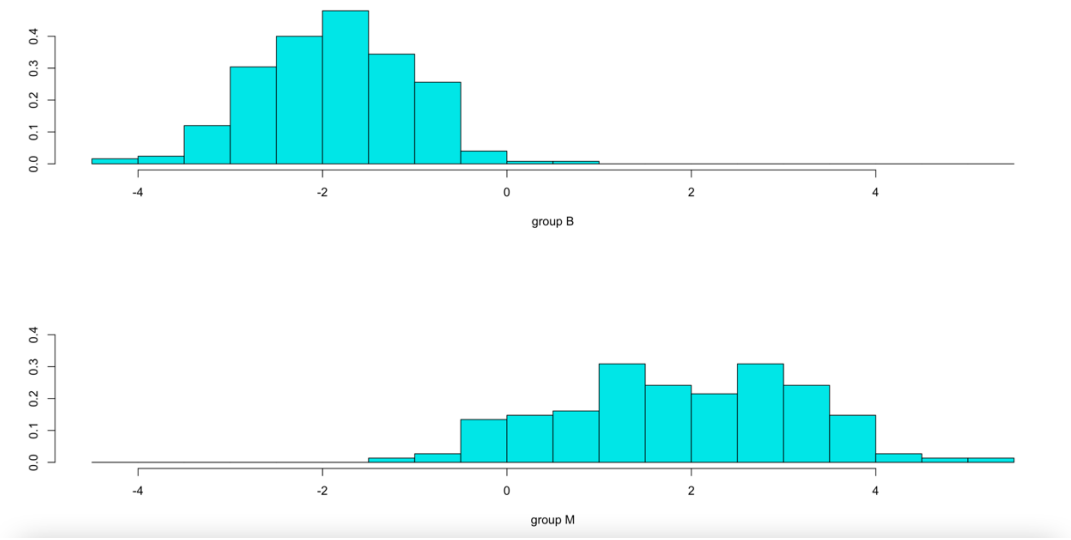
0 significant predictors might cause by too many predictors in this model. Therefore, it might be more appropriate for this data to do logistics model with penalization.

Model 2- Penalized Logistic Regression The Penalized logistics regression model built by 10-fold cross-validation has a 0.9941 accuracy rate and Kappa value 0.9873 on test data. This model contains predictors include radius_mean, texture_mean, perimeter_mean , area_mean, concavity_mean, concave_points_mean, radius_se, perimeter_se, area_se, compactness_se, symmetry_se, fractal_dimension_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave_points_worst , symmetry_worst, fractal_dimension_worst. The best tunning parameters for this model is alpha equals to 0.3, lambda equals to 0.01907868.

Model 3- Multivariate Adaptive Regression Splines The chosen degree of interactions and the number of retained terms are 1 and 11.The final model contain predictors including area_worst(with a knot at 1226), radius_mean(with a knot at 16.84), concave_points_worst(with a knot at 0.09173), texture_worst(with a knot at 35.34), symmetry_worst (with a knot at 0.2213), perimeter_worst (with a knot at 123.8),symmetry_mean(with a knot at 0.1547), perimeter_worst (97.67), concavity_mean(with a knot at 0.1975), concavity_se(with a knot at 0.002074). The accuracy and Kappa of this multivariate adaptive regression spline model on test data are 0.9765 and 0.9502, which means this model can make a great prediction on whether the tumor is malignant.

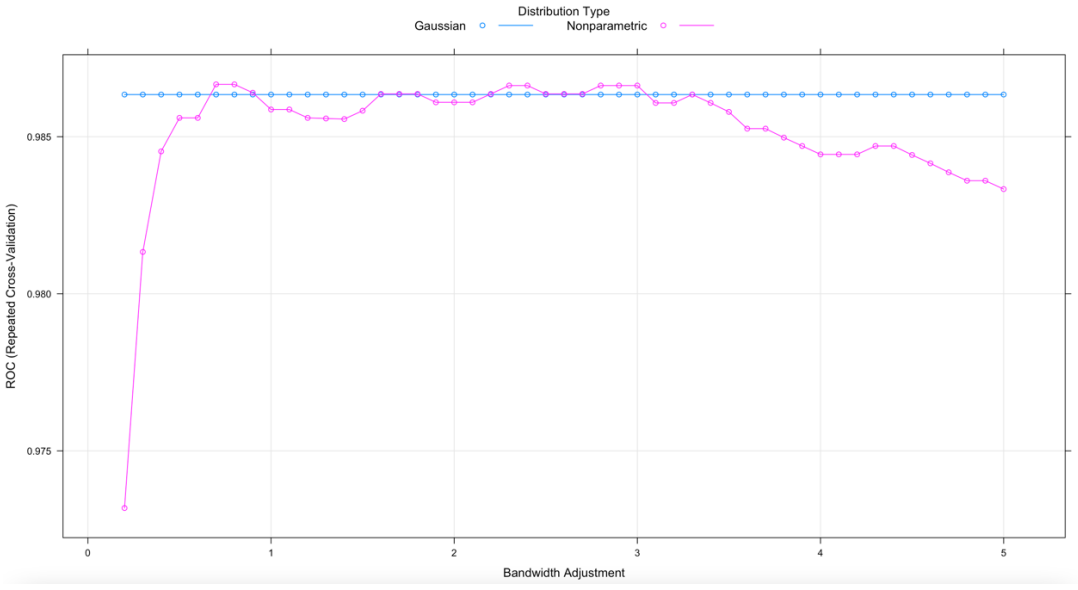
Model 4- Quadratic Discriminant Analysis Since we already used Multivariate Adaptive Regression Splines and Logistics Regression to analyze this model. Next, the discriminant analysis will be explored. However, K-Nearest Neighbor classifiers were not considered due to their lack of insight into predictors and outcomes. The Quadratic Discriminant Analysis assumes different covariance matrices for all the classes, so it is more flexible than Linear Discriminant Analysis which assumes the equality of group covariance matrices. Using the training data, the QDA built by R has a 0.9412 accuracy rate on test data and Kappa 0.8755.

Model 5- Linear Discriminant Analysis Using the training data, the LDA built by 10-fold cross-validation has 0.9647 accuracy on test data and Kappa 0.9228. Plot 4 shows how the LDA classifier $\tilde{X}$ (Linear discriminant variable) has classified the response class. When the $\tilde{X}$ is larger, the tumor is more likely to classify as malignant. If the $\tilde{X}$ is smaller, the tumor is more likely to be classified as benign.



Plot 2, how linear discriminant variable classified the tumors into two groups.

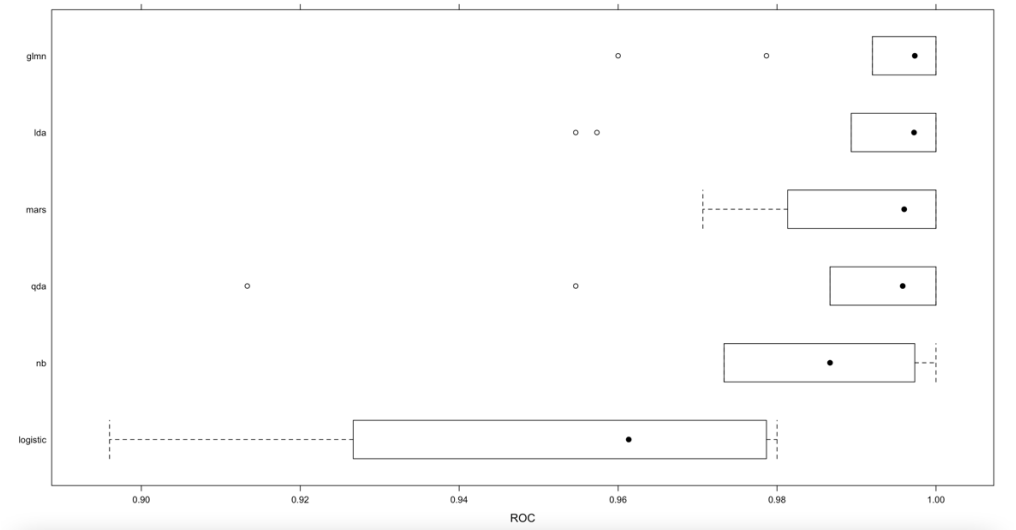
Model 6- Naive Bayes Naive Bayes has the assumption that features are independent in each class, and it is useful when the number of predictors is large. Considering that this data has many predictors, we also consider Naïve Bayes for prediction. The final Naive Bayes has a 0.9471 accuracy and Kappa 0.8883 on test data using the training data. The best Kernel density estimator we got is 0.7.



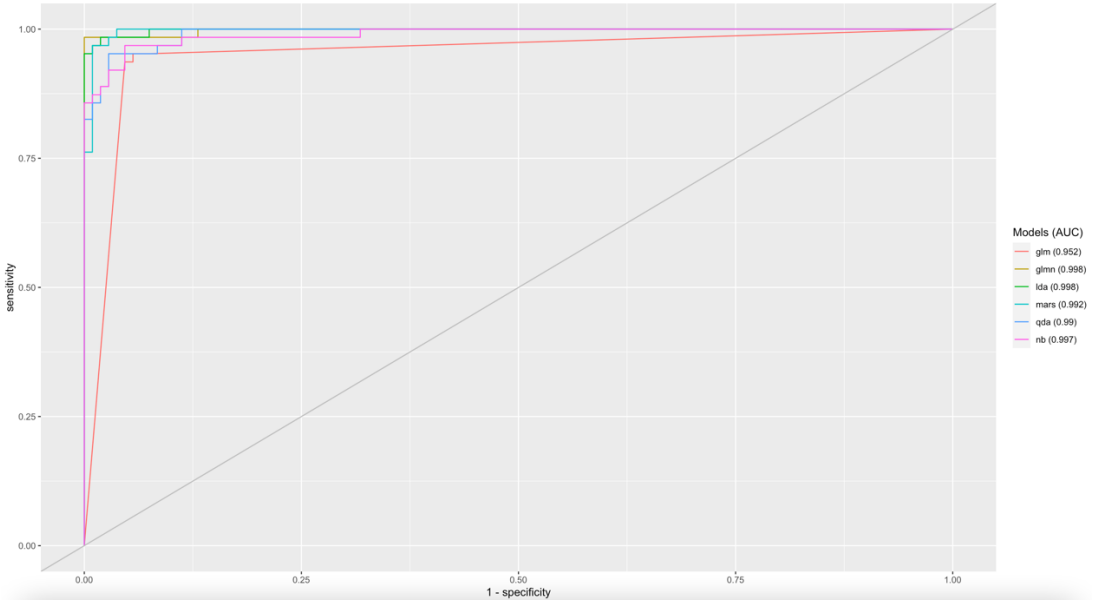
Plot 3, looking for best Kernel density estimator that can give a maximum AUC.

Results

Final model selection Among the six-candidate models, we selected model 4 (Penalized Logistic Regression) as the final model by comparing six models on training data by resampling method. As expected by plot 5 and 6 we can see that logistic regression has the smallest training AUC (0.9511) and testing AUC (0.952). The best performing model is Penalized Logistic Regression with the largest training AUC (0.992) and testing AUC (0.998). LDA also has the same testing AUC (0.998) as Penalized Logistic Regression on testing data; this might be due to the randomness.



Plot 4, Compare six models' performance on training data



Plot 5, Compare six models' performance on testing data

Model Interpretation The final Penalized logistics regression model has a 0.9941 accuracy rate on testing data, which means that this model can make a pretty good prediction on breast cancer. The models contain the mean of radius, texture, perimeter, area, concavity, concave points. Standard error of radius, perimeter, area, compactness, symmetry, fractal dimension. Largest smoothness, compactness, concavity, concave points, symmetry, fractal dimension. The Penalized logistics regression can be interpreted as when all other predictors are fixed. An increase of 1 unit of tumor's radius means multiplies the odds of the tumor being classified as malignant by 1.094379. When all other predictors are fixed, an increase of 1 unit of tumor's texture means multiplies the odds of the tumor being classified as malignant by 1.064703. When all other predictors are fixed, an increase of 1 unit of the tumor's perimeter means multiplies the odds of the tumor be classified as malignant by 1. 012702.When all other predictors are fixed, an increase of 1 unit of the tumor's area mean multiplies the odds of the tumor being classified as malignant by 1.000781. When all other predictors are fixed, an increase of 1 unit of tumor's concavity means multiplies the odds of the tumor being classified as malignant by 47.49644. When all other predictors are fixed, an increase of 1 unit of the tumor's concave points mean multiplies the odds of the tumor being classified as malignant by 77990.46. When all other predictors are fixed, an increase of 1 unit of tumor's radius standard error multiplies the odds of the tumor being classified as malignant by 6.193406. When all other predictors are fixed, an increase of 1 unit of tumor's perimeter standard error multiplies the odds of the tumor being classified as malignant by 1.156259. When all other predictors are fixed, an increase of 1 unit of tumor's area standard error multiplies the odds of the tumor being classified as malignant by 1.006397. When all other predictors are fixed, an increase of 1 unit of tumor's compactness standard error multiplies the odds of the tumor being classified as malignant by 0.0009541761. When all other predictors are fixed, an increase of 1 unit of tumor's symmetry standard error multiplies the odds of the tumor being classified as malignant by 0.0001585510. When all other predictors are fixed, an increase of 1 unit of the tumor's fractal dimension standard error multiplies the odds of the tumor being classified as malignant by 0.01698308. When all other predictors are fixed, an increase of 1 unit of the largest tumor's radius multiplies the odds of the tumor being classified as malignant by 1.122886. When all other predictors are fixed, an increase of 1 unit of the largest tumor's texture multiplies the odds of the tumor being classified as malignant by 1.108832. When all other predictors are fixed, an increase of 1 unit of the largest tumor's perimeter multiplies the odds of the tumor being classified as malignant by 1.015273. When all other predictors are fixed, an increase of 1 unit of the largest tumor area multiplies the odds of the tumor becoming malignant by 1.000765. When all other predictors are fixed, an increase of 1 unit of the largest tumor's smoothness multiplies the odds of the tumor being classified as malignant by 46122600. When all other predictors are fixed, an increase of 1 unit of the largest tumor's compactness multiplies the odds of the tumor being classified as malignant by 1.050738. When all other predictors are fixed, an increase of 1 unit of the largest tumor's concavity multiplies the odds of the tumor being classified as malignant by 4.571850. When all other predictors are fixed, an increase of 1 unit of the largest tumor's concave points multiplies the odds of the tumor being classified as malignant by 5601. 439. When all other predictors are fixed, an increase of 1 unit of the largest tumor's symmetry multiplies the odds of the tumor be classified as malignant by 241.5564. When all other predictors are fixed, an increase of 1 unit of the largest tumor's fractal dimension multiplies the odds of the tumor being classified as malignant by 1.871254.

Conclusion Since this final Penalized logistics regression has a prediction accuracy of 99%, and it means that by the different measuring features of tumors' images, we can detect breast cancer 99% correctly. By the variable importance scores, some important variables in this model include standard error of fractal dimension and symmetry, size of largest concave points and fractal dimension, mean of concave points.

Limitation Although we used ten-fold cross-validation on our final model, it still selected 22 predictors in the model. The Penalized logistics regression model performs well on both testing and training data. However, too many predictors might cause overfitting. It would be better to obtain more testing data to check the model's performance. Because we only have 170 observations (30% of original data) for testing, which might be too small to detect overfitting.