Classification analysis of Breast cancer diagnostic

Xuanyi Wang (Vicky), November 2018

Executive Summary

The dataset is the breast cancer diagnostics in Wisconsin, which was downloaded from the Kaggle website. It contains all the measurements from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

The analysis is based on 569 observations for each of the 32 variables. The first variable is ID number, the second column is diagnosis which is the response variables with two classes: One is M which stand for malignant and the other one is B which stand for benign. The response variable has 357 observations for benign and 212 for malignant.

The rest of the variables is explanatory variables. They are all measurements for each cell nucleus. For example, radius is the distances from center to points on the perimeter and texture of the breast mass. All the measurements have mean, standard error and worst as variables.

After exploring the data by calculating summary and descriptive statistics, and by creating visualizations of the correlation between each numerical variable, several highly correlated variables are found. After exploring the data, four algorithms have been tested for this train dataset and the best model has been choose based on the model accuracy.

Initial Data Exploration

The initial exploration of the data began with some summary and descriptive statistics.

Individual Feature Statistics Summary statistics for minimum, maximum, mean, median, standard deviation, and distinct count were calculated for numeric columns, and the results taken from 216 observations are shown here:

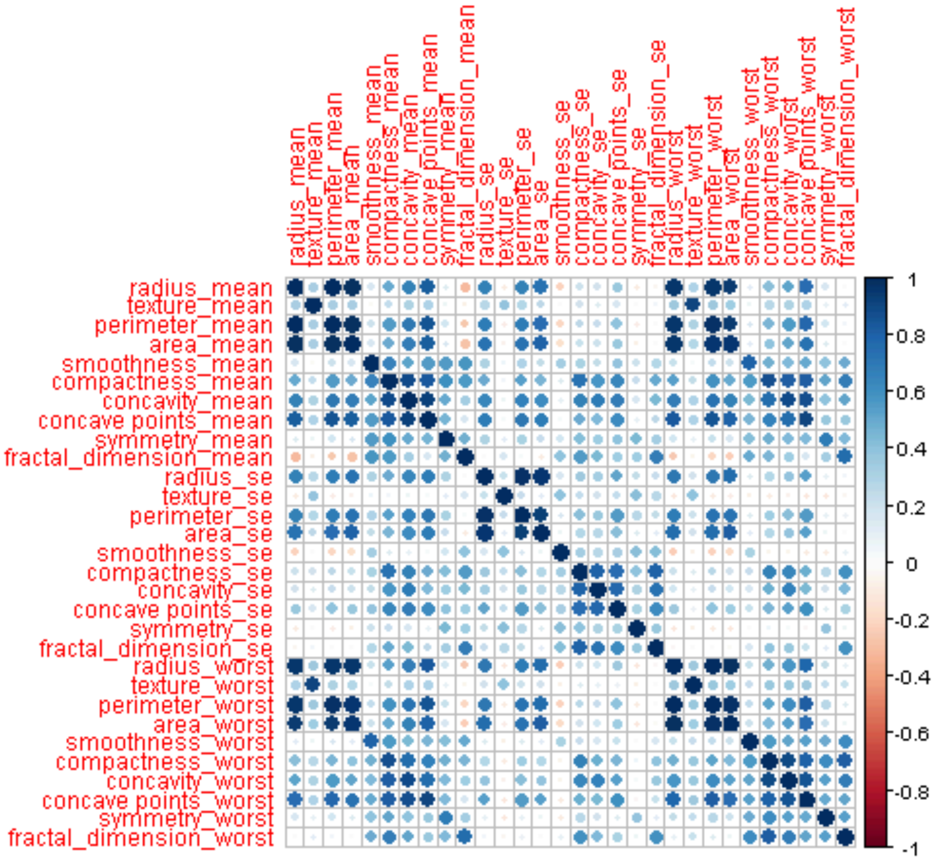
Feature	Min	Max	Mean	Median	Std
radius_mean	6.981	28.11	14.12729174	13.37	3.524048826
texture_mean	9.71	39.28	19.28964851	18.84	4.301035768
perimeter_mean	43.79	188.5	91.96903339	86.24	24.29898104
area_mean	143.5	2501	654.8891037	551.1	351.9141292
smoothness_mean	0.05263	0.1634	0.096360281	0.09587	0.014064128
compactness_mean	0.01938	0.3454	0.104340984	0.09263	0.052812758
concavity_mean	0	0.4268	0.088799316	0.06154	0.079719809
concave points_mean	0	0.2012	0.048919146	0.0335	0.038802845
symmetry_mean	0.106	0.304	0.181161863	0.1792	0.027414281
fractal_dimension_mean	0.04996	0.09744	0.06279761	0.06154	0.007060363
radius_se	0.1115	2.873	0.405172056	0.3242	0.277312733
texture_se	0.3602	4.885	1.216853427	1.108	0.551648393
perimeter_se	0.757	21.98	2.866059227	2.287	2.021854554
area_se	6.802	542.2	40.33707909	24.53	45.49100552
smoothness_se	0.001713	0.03113	0.007040979	0.00638	0.003002518

compactness_se	0.002252	0.1354	0.025478139	0.02045	0.017908179
concavity_se	0	0.396	0.031893716	0.02589	0.03018606
concave points_se	0	0.05279	0.011796137	0.01093	0.006170285
symmetry_se	0.007882	0.07895	0.020542299	0.01873	0.008266372
fractal_dimension_se	0.0008948	0.02984	0.003794904	0.003187	0.002646071
radius_worst	7.93	36.04	16.26918981	14.97	4.83324158
texture_worst	12.02	49.54	25.6772232	25.41	6.146257623
perimeter_worst	50.41	251.2	107.2612127	97.66	33.60254227
area_worst	185.2	4254	880.5831283	686.5	569.3569927
smoothness_worst	0.07117	0.2226	0.132368594	0.1313	0.022832429
compactness_worst	0.02729	1.058	0.254265044	0.2119	0.157336489
concavity_worst	0	1.252	0.272188483	0.2267	0.208624281
concave points_worst	0	0.291	0.114606223	0.09993	0.065732341
symmetry_worst	0.1565	0.6638	0.290075571	0.2822	0.061867468
fractal_dimension_worst	0.05504	0.2075	0.083945817	0.08004	0.018061267

Correlation and Relationships

Numeric Relationships

The correlation between the numeric columns were calculated and observed in the below correlation plot. (The right color bar indicated the correlation values. For example, dark blue means correlation value is 1 and dark red means correlation value is negative 1.)

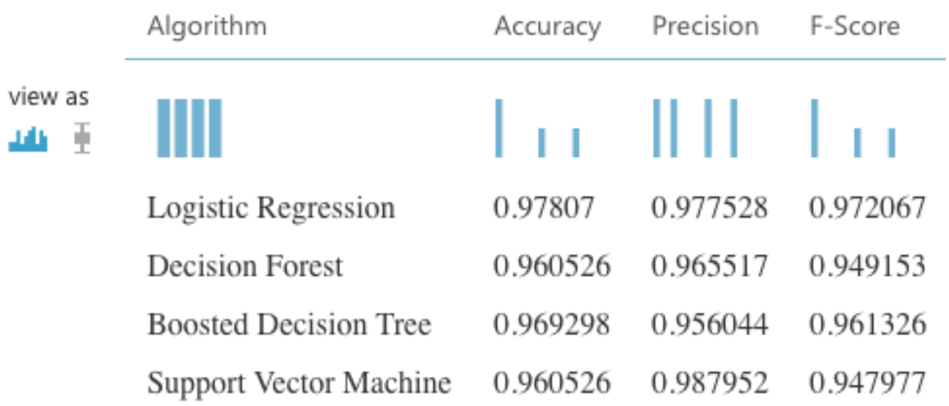


The graph shows that radius, perimeter and area of the breast mass has strong positive correlation with each other. This strong correlation can be seen on mean, se and worst value.

Analysis

In this analysis, four algorithms have been tested, which are Two-Class Logistic Regression, Two-Class Decision Forest, Two-Class Boosted Decision Tree and Two-Class Support Vector Machine.

These algorithms were trained with 60% of the data. Testing the model with the remaining 40% of the data yielded the following results:



Algorithm	Accuracy	Precision	F-Score
Logistic Regression	0.97807	0.977528	0.972067
Decision Forest	0.960526	0.965517	0.949153
Boosted Decision Tree	0.969298	0.956044	0.961326
Support Vector Machine	0.960526	0.987952	0.947977

As shown from the above table Two-Class Logistic regression has the highest Accuracy, Precision and F-score of classify the diagnostics.

Decision for model chosen, process and results

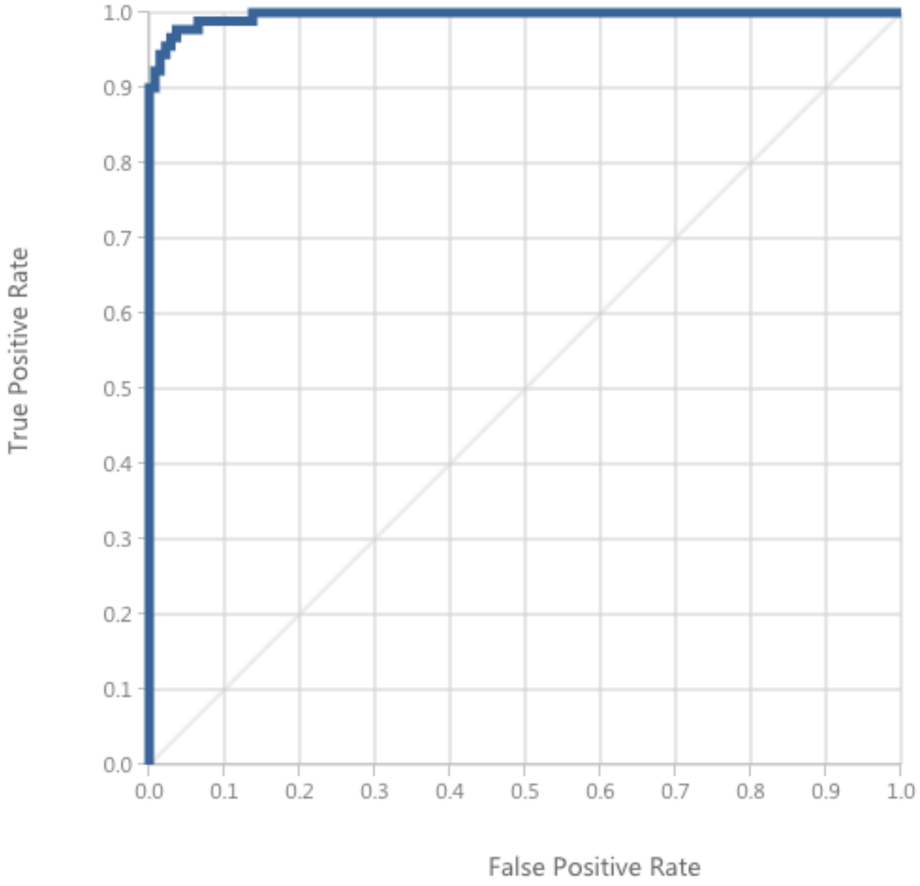
Based on the output of the model evaluation, Two-Class Logistic Regression is chosen as the prediction model.

More details of the performance of this model has been shown as followings;

(The positive label is M(Malignant), the negative label is B(Benign))

- True Positives: 85
- True Negatives: 136
- False Positives: 2
- False Negatives: 5

The Received Operator Characteristic (ROC) curve for the model is shown here, with the blue line indicating the model’s performance at varying classification threshold values, and the diagonal line showing the expected results of a random guess:



This translates in to the following standard performance metrics for classification:

- Accuracy: 97.0%
- Precision: 97.7%
- Recall: 94.4%
- F1 Score: 96.0%

Conclusion

This analysis has shown that the breast cancer diagnostic can be confidently predicted from its measurements of a breast mass. In particular, the two-class logistic regression has the best performance among the four algorithms. The accuracy rate is 97%.