Creator:solar22
Credit: Getty Images/iStockphoto

Diagnosing Breast Cancer With Machine Learning

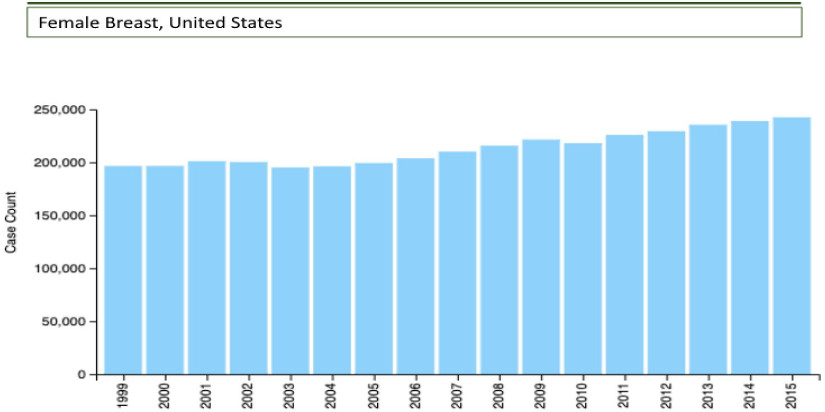
By Lisa Dawes
June 6, 2019

Thinkful Unit 3 Capstone Project

Early Detection Of Breast Cancer Is Key To Saving Lives

- While rates of cancer diagnoses and cancer deaths continue to decline each year, the ***number of new cases and deaths is rising***. This can be attributed to population growth and aging.
- If breast cancer is ***detected early***, there are more treatment options.
- Women whose breast cancer is detected at an early stage have a ***93 percent or higher survival rate*** in the first five years.

Incidence of Breast Cancer



Sources: Centers for Disease Control and Prevention, https://gis.cdc.gov/Cancer/USCS/DataViz.html; Carol Milgard Breast Center, http://www.carolmilgardbreastcenter.org/early-detection

Research Questions & Target Variable

Research Questions

1. How can we use features computed from a digitized image of a breast mass to best predict benign vs malignant masses?
2. What traits are most indicative of whether or not an individual will be diagnosed?

Target Variable

The target variable for my analysis is ‘diagnosis’ which was originally a list of “Ms” (Malignant) and “Bs” (Benign). A malignant diagnosis means that the mass is cancerous while a benign diagnosis means that the mass is most likely harmless.

Data Description

- This data set was pulled from the UCI Machine Learning Repository.
- It contains a total of 30 measurements derived from fine needle aspirate (FNA) images of a breast mass taken from patients at three (3) separate times.
- The 30 features mentioned above were all floats. The ‘ID number’ is an integer and the target variable ‘Diagnosis’ was initially an object.
- All variables (with exception of ‘ID number’ and ‘Diagnosis’) were continuous and rounded to the fourth decimal place.

Link:https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

Data Description (Continued)

Data Shape
RangeIndex: 568 entries, 0 to 567
Data columns: total of 32

M - Mean
SE - Standard Error
Worst - Worst

The mean, standard error and “worst” or largest (mean of the three largest values)of these features were computed for each sample resulting in 30 features excluding the ‘Id Number’ and ‘Diagnosis’ columns.

id_number	diagnosis	M_radius	M_texture	M_perimeter	M_area	M_smoothness	M_compactness	M_concavity	M_concave points	M_symmetry	M_fractal dimension
842517	M	20.570	17.770	132.900	1326.000	0.085	0.079	0.087	0.070	0.181	0.057

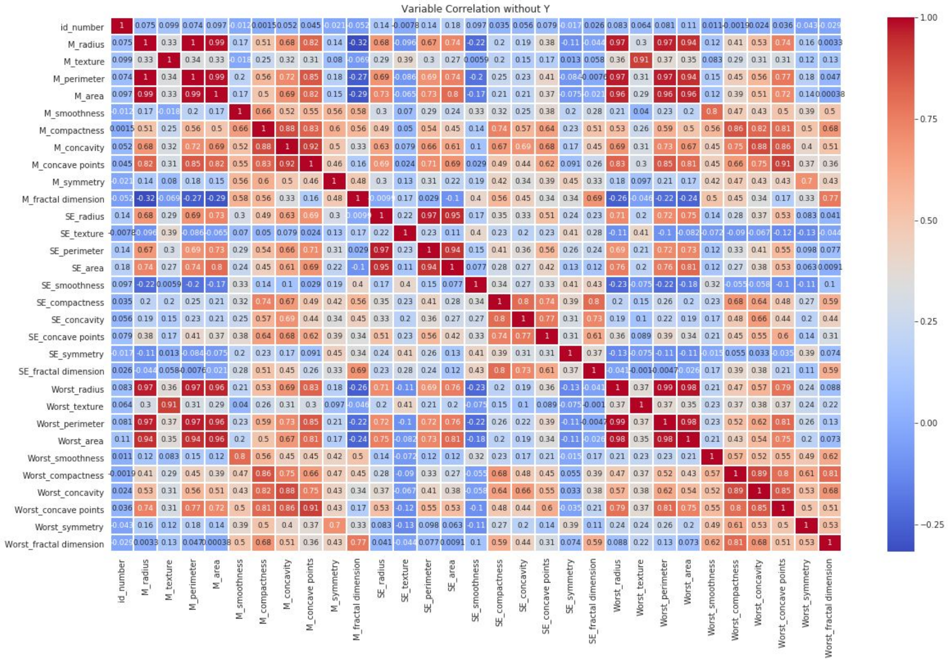
SE_radius	SE_texture	SE_perimeter	SE_area	SE_smoothness	SE_compactness	SE_concavity	SE_concave points	SE_symmetry	SE_fractal dimension
0.543	0.734	3.398	74.080	0.005	0.013	0.019	0.013	0.014	0.004

Worst_radius	Worst_texture	Worst_perimeter	Worst_area	Worst_smoothness	Worst_compactness	Worst_concavity	Worst_concave points	Worst_symmetry	Worst_fractal dimension
24.990	23.410	158.800	1956.000	0.124	0.187	0.242	0.186	0.275	0.089

Initial Data Exploration

Feature Selection Strategies

- Basic geometric principles
- Created a heat map to see how variables would be correlated with each other.
- There were no missing values in this dataset.
- Based on the heat map I dropped 10 features.



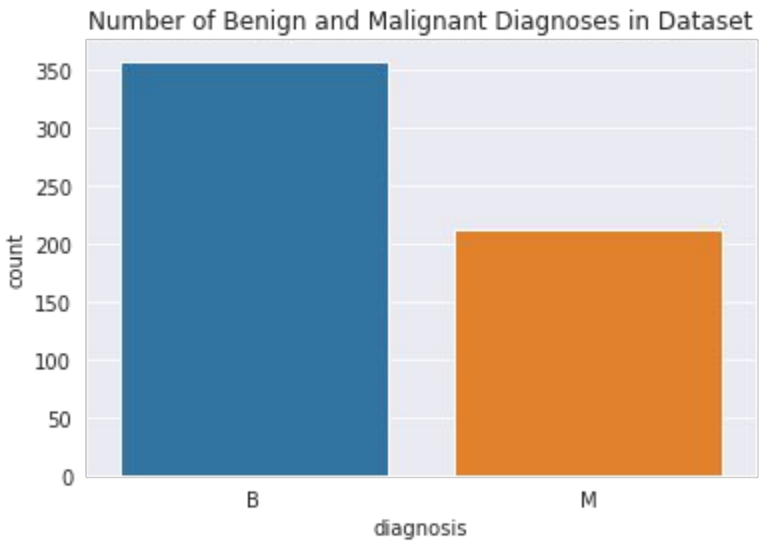
Lost of multicollinearity!

Initial Data Exploration (Continued)

Categorical Data

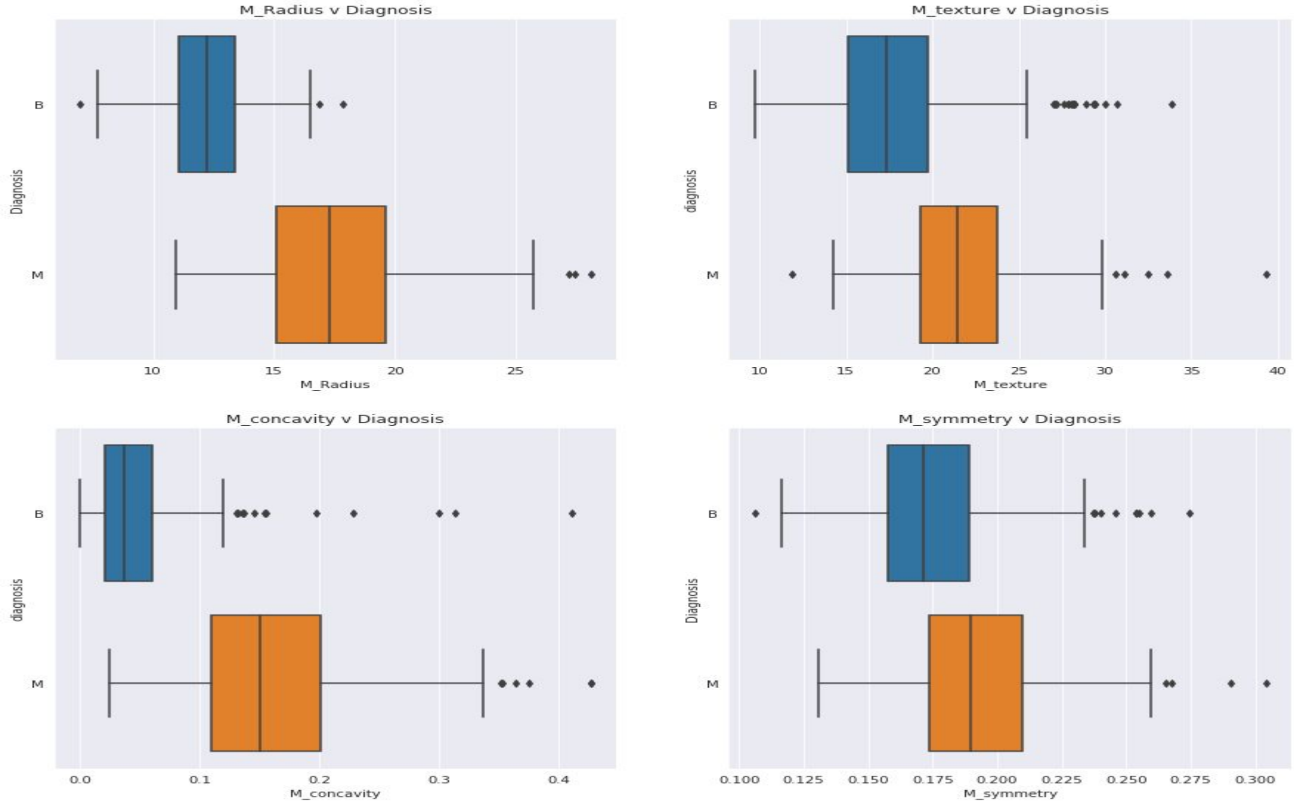
- The target variable – ‘Diagnosis’– was converted into a binary categorical variable. As a result, answering my research question is a ***classification problem.***

Class Imbalance



Initial Data Exploration (Continued)

- The boxplots to the right show the relationship of select features with the target variable.
- From the graphs we can show that the dataset contained some outliers.



Model Overview And Initial Results

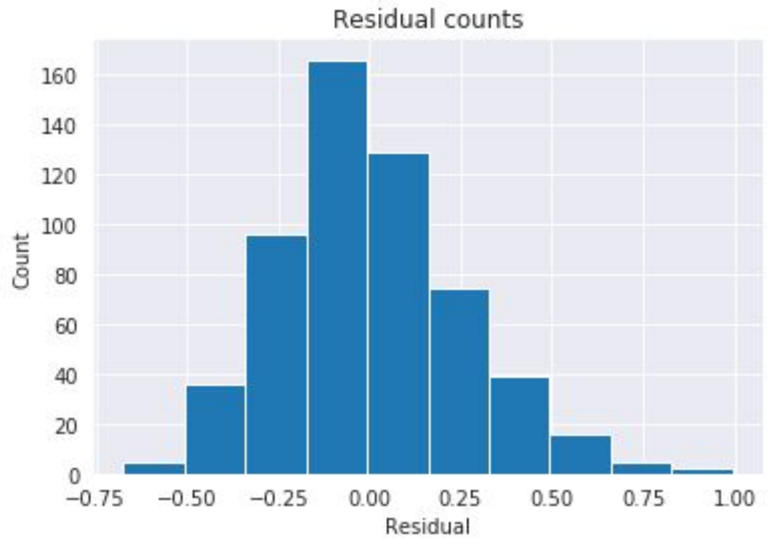
Data Split Method:

- SKLearn - Train, Test, Split

Model Type	Train_Initial Results	Test_Initial Results
Linear Regression	Percentage Accuracy: 0.7267851336392251	Percentage Accuracy: 0.6864551287423677
Logistic Regression	Percentage Accuracy: 0.9700704225352113	Percentage Accuracy 0.9436619718309859
Random Forest	Prediction Accuracy: 0.9806338028169014	Prediction Accuracy: 0.9524647887323944
Random Forest with Gradient Boosting	Percentage Accuracy 0.9876760563380281	Percentage Accuracy 0.9647887323943662

Results here look pretty good!

Model #1 - Linear Regression (Not the Best!)



Coefficients

M_texture	-0.000
M_area	0.000
M_smoothness	1.617
M_compactness	0.272
M_symmetry	-0.051
M_fractal dimension	-13.12
SE_texture	0.013
SE_smoothness	15.358
SE_compactness	-3.923
SE_concavity	-3.341
SE_concave points	4.565
SE_symmetry	1.843
SE_fractal dimension	26.059
Worst_texture	0.010
Worst_smoothness	0.512
Worst_compactness	-0.183
Worst_concavity	0.616
Worst_concave points	2.648
Worst_symmetry	0.650
Worst_fractal dimension	2.897

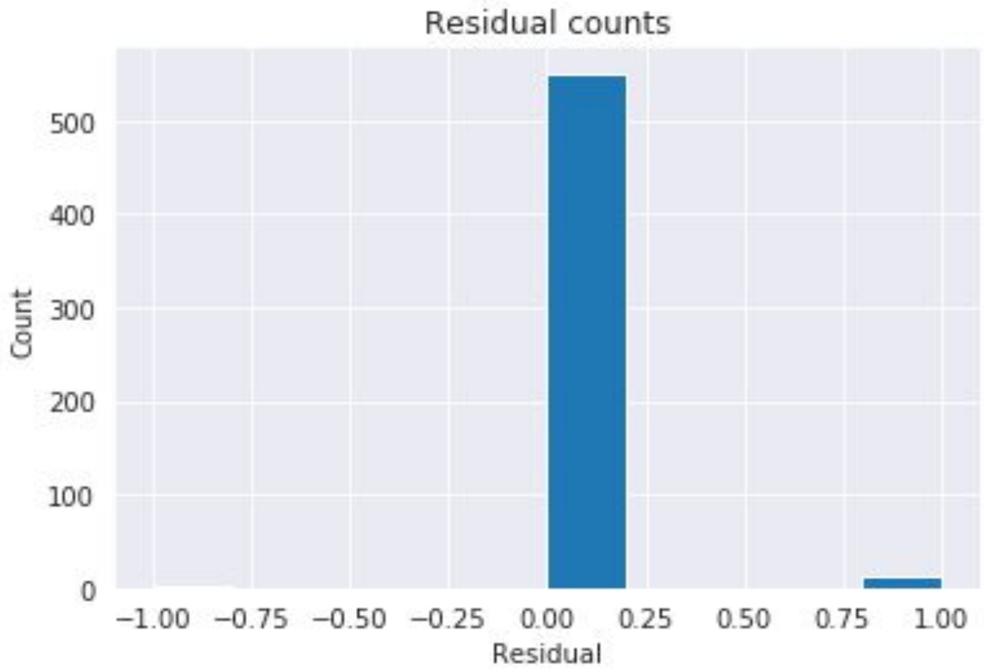
Model #2 - Logistic Regression

Accuracy by Diagnosis - Train

Diagnosis	B	M
B	354	14
M	3	197

Accuracy by Diagnosis - Test

	B	M
B	337	12
M	20	199



Model #3 - Random Forest

Model Type	Train_Initial Results	Test_Initial Results
Random Forest	`Prediction Accuracy: 0.9806338028169014`	`Prediction Accuracy: 0.9524647887323944`

The Black Box!

Model #4 - Random Forest With Gradient Boost

Parameters

- 500 iterations, using 2-deep trees, and loss function.
- I chose to use ‘deviance’ here because it is used with logistic regression.
- **params** = {'n_estimators': 500, 'max_depth': 2, 'loss': 'deviance'}

Training Set Accuracy:

Percent Type I errors: 0.0

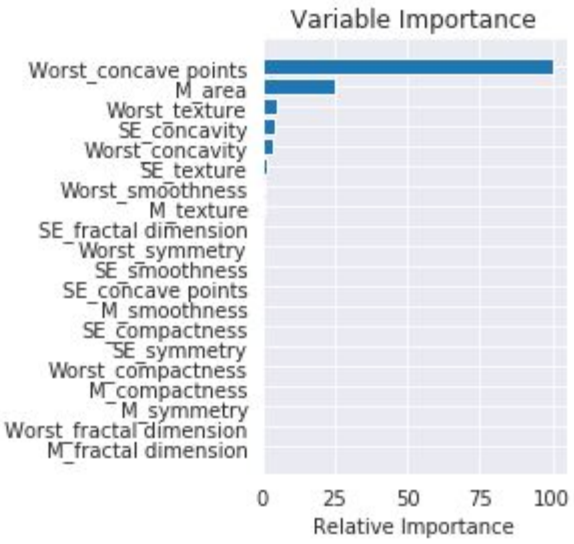
Percent Type II errors: 0.0

Test Set Accuracy:

Percent Type I errors: 0.005847953216374269

Percent Type II errors: 0.03508771929824561

Model #4 - Random Forest With Gradient Boost



- Looks like a lot of these features may be useless.
- Worst concave points (number of concave portions of the contour)
- As a next step, I should run the models again using only these features to see if I get better accuracy.

Conclusion

- Measurements derived from fine needle aspirate (FNA) images of a breast mass can predict with great accuracy whether or not a mass is malignant or benign.
- False Positives/False Negatives
- Next Steps
  - Run the models again with less features
  - Use gridsearch to tune Random Forest Model
  - Try PCA (but data set is small and models ran quite well)
  - Get more data
- Additional Research
  - Look at tumor growth over time

Thank You!