Breast Cancer Wisconsin (Diagnostic) Data

TEAM MISNEACH

Nadadur, Karuna

Ogilvie, Jessa

Sharma, Gaurav

Yang, Siwen

Executive Summary

Breast cancer is the second most common cause of cancer death in women and is responsible for approximately 15% of cancer deaths in the United States (Goddard et al, 2013). Once a cancerous mass has been identified many oncologists will order 2-5 time consuming and expensive tests to confirm a diagnosis. Studies have shown the total cost to diagnose a single patient can be as high as $28,000 (Honein-AbouHaidar et al, 2017). This is due in part to the cost of testing as well as the consultation time that a doctor must use in order to make the determination of cancer. The follow study seeks to determine whether or not data analytics can significantly and accurately assist in the diagnosis of breast cancer using only information gained from the relatively cheap procedure of fine needle aspiration[1].

In doing so it is thought that the number of tests and the time spent interpreting their results will be decreased, resulting in a significant savings for both hospitals and patients. Several tests were run to determine if the measurements from the data set were significantly different between groups – malignant and benign – and confidence intervals were tested to determine if simple thresholds could be used to recommend a diagnosis. Unfortunately, we were unable to establish such parameters and subsequently built a logistic regression model which was able to predict the diagnosis with an accuracy rate of 98.24% . As such it is recommended that hospitals and medical practitioners consider including analytics models, such as the logistic regression presented in this case, in conjunction with diagnostic tests in order to decrease the time spent analyzing results and the number of test needed to diagnose breast cancer.

[1] In this method, a fine needle of specific measurements is thrust into the lump and the cells (specifically the nuclei of the cells) of the growth thus obtained are studied under a microscope.

Methodology

The data used in this analysis was created using cancer diagnostic data for the Wisconsin area and contained 569 rows and 32 attributes(SOURCE). Data points were obtained from fine needle aspiration tests. The distinct attributes of the nuclei studied here are:

1. **Radius**: mean of distances from center to points on the perimeter
2. **Texture:** standard deviation of gray-scale values
3. **Perimeter:** circumference
4. **Area:** mean Pi*(radius^2)
5. **Smoothness:** local variation in radius lengths
6. **Compactness:** perimeter^2 / area - 1.0
7. **Concavity:**severity of concave portions of the contour
8. **Concave points:** number of concave portions of the contour
9. **Symmetry:** Regularity or consistency of structure
10. **Fractal dimension:** "coastline approximation" - 1

Further, the data set chosen also gives the means, the standard error readings and the worst (i.e., mean of 3 largest) values for all the above attributes for meticulous analyzation.

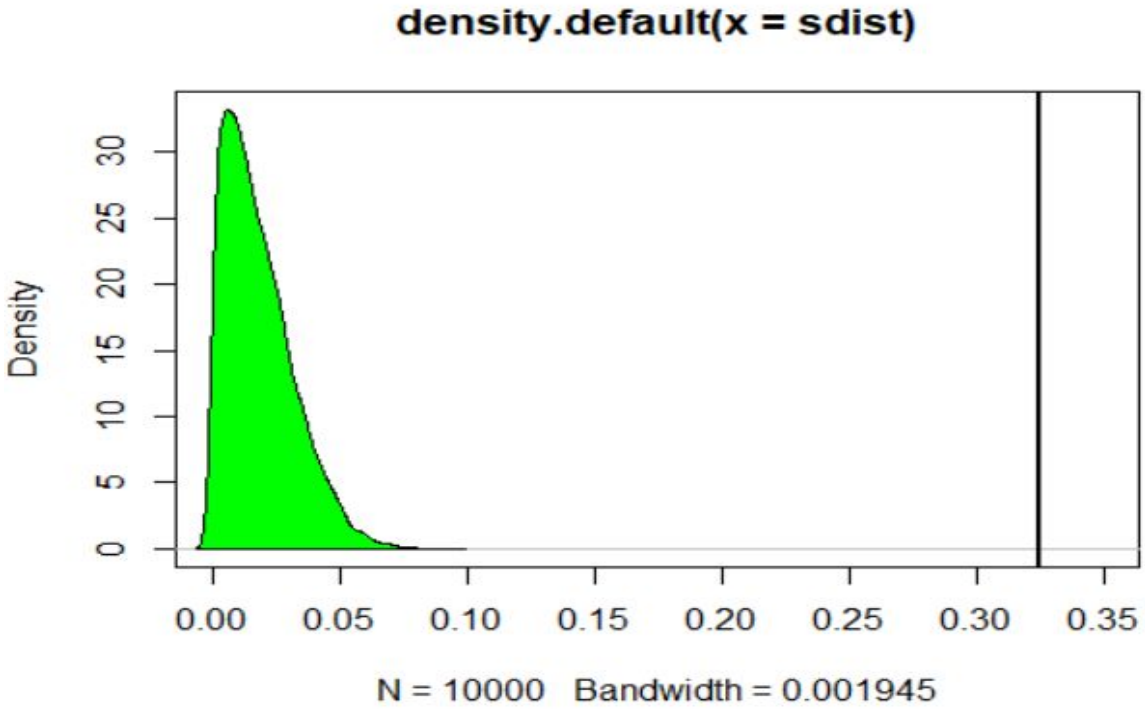
Modifications & Pre-processing

The data obtained contained no missing values and no outliers. Two columns were added, the first was a binomial column corresponding to the diagnosis where “M” = 1 and “B”=0. Second, the “cut off” column “X” was created for the two-sample test where any value greater than 14 became 1 for malignant and any value less than 14 became 0 for benign.

Results: Statistical Testing

Two Sample Test (Non-Parametric):

In this test, we evaluated the null hypothesis that the radius for both the groups Malignant and Benign is same. We attempted to compare the difference of the average radius for the categorical target variable Diagnosis and the difference between the average value of random sample taken the dataset. The result of the two sample test states that the average radius for both the group is not same i.e we reject the null hypothesis. We also evaluated that if the average for both the diagnosis group is not same than which group has larger average value, the result states that Malignant group average radius is more than Benign.



Test to evaluate which higher average radius for Malignant and Benign group :

```
        Welch Two Sample t-test

data:  g1 and g2
t = 13.301, df = 237.95, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2768640 0.3731364
sample estimates:
mean of x mean of y
0.6090825 0.2840824
```

Statistical Testing: 75th quantile for the data

The attribute chosen for analysis for this test is radius_mean and smoothness factor. The computation of the 75th quantile value of the radius_mean will help technicians in understanding the average values for nucleus radius of malignant cancers and of benign cancers. Using these values, the doctors can determine if the candidate needs further examination or can be cleared for cancer.

Computing the 75th quantile value for the data:

1. For mean texture of malignant cancer types:

```
s1=data[data$diagnosis=="M",4]
View(s1)
t1=s1
View(t1)
quantile(t1,probs = 0.75)
```

Output:

quantile(t1,probs = 0.75)

75%

23.765

From the output, we incur that 75% of the time the mean texture value is around 23.765 when the cancer is malignant.

2. For mean texture of benign cancer types:

```
s2=data[data$diagnosis=="B",4]
View(s2)
t2=s2
View(t2)
quantile(t2,probs = 0.75)
```

Output:

quantile(t2,probs = 0.75)

75%

19.76

From the output, we incur that 75% of the time the mean texture value is around 19.76 when the cancer is benign.

By comparing the mean texture values for malignant and benign cancer nuclei, we can conclude that malignant cancer nuclei have higher texture value on an average.

The doctors can predict potential cancers by comparing the texture of cancerous cells with the texture of normal cells extracted from the same candidate.

3. For mean smoothness factor of malignant cancer types:

```
s3=data[data$diagnosis=="M",7]
View(s3)
t3=s3
View(t3)
quantile(t3,probs = 0.75)
```

Output:

quantile(t3,probs = 0.75)

75%

0.110925

From the output, we can incur that 75% of the time the mean smoothness is around 0.110925 when the cancer is malignant.

4. For mean smoothness factor of benign cancer types:

```
s4=data[data$diagnosis=="B",7]
View(s4)
t4=s4
View(t4)
quantile(t4,probs = 0.75)
```

Output:

quantile(t4,probs = 0.75)

75%

0.1007

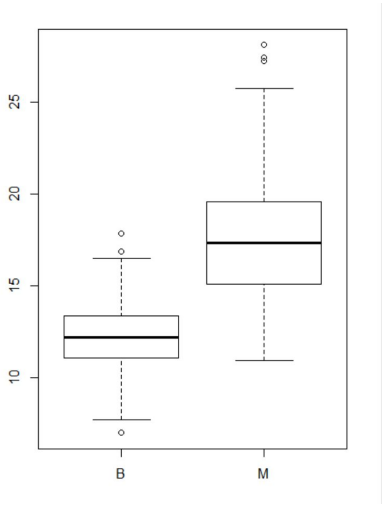
From the output, we can incur that 75% of the time the mean smoothness is around 0.1007 when the cancer is benign.

By comparing the mean smoothness values for malignant and benign cancer nuclei, we can conclude that malignant cancer nuclei have higher smoothness value on an average.

Now, since cancer is known to have irregular shapes, it is usually expected that benign cells would be smoother compared to malignant cells. But this test suggested that the opposite is true. This is an interesting find for our dataset.

The doctors can predict potential cancers by comparing the smoothness factor of cancerous cells with that of normal cells extracted from the same candidate.

Chi-Square: Creating Thresholds: After visualizing the data and creating confidence intervals we attempted to create a cut point for the radius mean which would identify cell means above the threshold as malignant and below as benign. The chosen cut point was 14, as show in the Figure #, at this size only the outliers of each group overlap, making it a strong candidate as the cut point.



In order to perform the chi-square test the data was re-coded into a new column “X”, where any value greater than 14 became “1” for malignant and “0” for benign. Unfortunately, the test produced a p-value of .00000004, the null hypothesis was rejected; indicating that there is a significant difference between the diagnosis group and the re-coded group using our proposed cut point. As a result we determined that a more complex model would be needed to determine if a sample was cancerous.

```
p_diag = prop.table(table(data$diagnosis))
p_rad = prop.table(table(data$X))
p = p_diag %*% t(p_rad)
n = nrow(data)
E = p*n
O = table(data$diagnosis,data$X)
tstat = sum((O-E)^2/E)

f1 = function()
{
  s1 = sample(x = c('M',"B"),size = n,replace = T,prob = p_diag)
  s2 = sample(x = c("0","1"),size = n,replace = T,prob = p_rad)
  O = table(s1,s2)
  return(sum((O-E)^2/E))
}
f1()
sdist = replicate(10000,f1())
plot(density(sdist))
polygon(density(sdist),col="green")
abline(v=tstat,lwd=2)
rset = sdist[sdist>=tstat]
p_value = length(rset)/length(sdist)
p_value
#.0000000004
```

Statistical Estimation: Maximum Likelihood Method : We want to know probabilities for the variable “diagnosis”: whether the tumor is malignant or benign. Due to our target is a categorical variable with two classes, we chose to build a binomial logistic regression. In our data, “1” represents malignant and “0” represents benign.

To achieve this, we conducted a logistic regression using mle function that explains the impact of each variable selected i.e., smoothness_mean, compactness_mean, symmetry_mean, fractal_dimension_mean, texture_se, area_se, smoothness_se, compactness_se, concavity_se, concave.points_se, symmetry_se, fractal_dimension_se, texture_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave.points_worst, symmetry_worst and fractal_dimension_worst. Although we identified some strong predictors, we still develop the function that contain 20 variables to find their likelihood .

The combination of predictors impacting outcome can be shown as log(probability(malignant) /probability(non-malignant)) = b0 + b1*smoothness_mean + b2*compactness_mean +

b3*symmetry_mean + b4*fractal_dimension_mean + b5*texture_se + b6*area_se + b7*smoothness_se + b8*compactness_se + b9*concavity_se + b10*concave.points_se + b11*symmetry_se + b12*fractal_dimension_se + b13*texture_worst+b14*area_worst + b15*smoothness_worst + b16*compactness_worst + b17*concavity_worst + b18*concave.points_worst + b19*symmetry_worst + b20*fractal_dimension_worst. Then we use the function mle2() to find the maximum likelihood parameter value. The summarized coefficient is shown as follow:

```
Maximum likelihood estimation

Call:
mle2(minuslogl = f1, start = list(b0 = 0, b1 = 0, b2 = 0, b3 = 0,
    b4 = 0, b5 = 0, b6 = 0, b7 = 0, b8 = 0, b9 = 0, b10 = 0,
    b11 = 0, b12 = 0, b13 = 0, b14 = 0, b15 = 0, b16 = 0, b17 = 0,
    b18 = 0, b19 = 0, b20 = 0))

Coefficients:
       Estimate  Std. Error   z value     Pr(z)
b0  -43.1510085   0.4752446  -90.7975 < 2.2e-16 ***
b1   25.2596843   0.0768817  328.5527 < 2.2e-16 ***
b2    1.2410400   0.1856747    6.6839 2.326e-11 ***
b3  -31.9622027   0.1616269 -197.7530 < 2.2e-16 ***
b4    0.2866859   0.0524206    5.4690 4.527e-08 ***
b5   -2.0495329   1.3162075   -1.5572  0.119435
b6    0.2646907   0.0502558    5.2669 1.388e-07 ***
b7   21.8460996   0.0049775 4388.9658 < 2.2e-16 ***
b8  -35.7410708   0.0850851 -420.0624 < 2.2e-16 ***
b9  -42.1749832   0.1949490 -216.3385 < 2.2e-16 ***
b10  19.9056274   0.0213618  931.8310 < 2.2e-16 ***
b11  -9.1456033   0.0304006 -300.8360 < 2.2e-16 ***
b12  -5.7328690   0.0078498 -730.3174 < 2.2e-16 ***
b13   0.5126416   0.0874046    5.8652 4.487e-09 ***
b14   0.0077143   0.0026932    2.8644  0.004178 **
b15  46.6359071   0.1174377  397.1118 < 2.2e-16 ***
b16 -18.9861871   0.6361032  -29.8477 < 2.2e-16 ***
b17  17.8885364   1.0894074   16.4204 < 2.2e-16 ***
b18  66.2885526   0.1789693  370.3906 < 2.2e-16 ***
b19  25.9222871   0.3454160   75.0466 < 2.2e-16 ***
b20   8.2235386   0.0993554   82.7689 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

-2 log L: 53.0829
> AIC(res)
[1] 95.0829
```

The estimate of Coefficients gives us the function: - 43.15 + 25.25*smoothness_mean + 1.24*compactness_mean - 31.96*symmetry_mean + 0.28*fractal_dimension_mean - 2.04*texture_se + 0.26*area_se + 21.84*smoothness_se - 35.74*compactness_se - 42.17*concavity_se + 19.91*concave.points_se - 9.14*symmetry_se - 5.73*fractal_dimension_se + 0.51*texture_worst + 0.007*area_worst + 46.63*smoothness_worst - 18.98*compactness_worst + 17.88*concavity_worst + 66.28*concave.points_worst + 25.92*symmetry_worst + 8.22*fractal_dimension_worst. Excepting the texture_se variable, the function shows that all of variables are higher than 0.05 which means that they have strong impact on the outcome of diagnosis_b.

Regression model using glm() function: Firstly, we use cor() function to find highly correlated variables and then reduce some of variables as follows: compactness_mean, area_worst, compactness_se, concavity_se, concave points_se, compactness_worst, concave points_worst, fractal_dimension_se. Then we make another model based on logistic regression function glm() after finding logistic results using mle2().

Secondly, we build a logistic regression of outcome by glm() function, and the function is shown as follows. reg1 = glm(diagnosis_b~.,family = "binomial",data=cancer). followed by coefficient summary.

```
Call:
glm(formula = diagnosis_b ~ ., family = "binomial", data = cancer)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6329  -0.0288  -0.0021   0.0001   3.6207

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -24.52103    8.61812  -2.845 0.004437 **
smoothness_mean           41.23562   82.41502   0.500 0.616835
symmetry_mean            -26.07035   30.46828  -0.856 0.392188
fractal_dimension_mean  -102.64178  173.53890  -0.591 0.554210
texture_se                -2.05948    1.36988  -1.503 0.132737
area_se                    0.28035    0.06598   4.249 2.14e-05 ***
smoothness_se            421.34932  378.22117   1.114 0.265267
concave.points_se       -292.21124  235.59165  -1.240 0.214854
symmetry_se             -134.88417  130.00682  -1.038 0.299495
texture_worst              0.45480    0.12726   3.574 0.000352 ***
smoothness_worst           0.66342   57.95729   0.011 0.990867
concavity_worst            6.73267    4.22245   1.594 0.110825
concave.points_worst      80.90727   26.29988   3.076 0.002096 **
symmetry_worst            32.73479   18.16521   1.802 0.071536 .
fractal_dimension_worst  -62.81652   60.99280  -1.030 0.303057
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.440  on 568  degrees of freedom
Residual deviance:  62.902  on 554  degrees of freedom
AIC: 92.902
```

The result shows AIC as 92.902. The regression structure is Outcome = 24.52 + 41.23* smoothness_mean - 26.07* symmetry_mean - 102.64* fractal_dimension_mean - 2.06 *texture_se + 0.28* area_se + 421.35* smoothness_se - 292.21* concave.points_se - 134.88* symmetry_se + 0.45* texture_worst + 0.66* smoothness_worst + 6.73* concavity_worst + 80.91* concave.points_worst + 32.73* symmetry_worst - 62.82*fractal_dimension_worst. The top three variables that impact the outcome are area_se, texture_worst and concave.points_worst this time.

In addition, we use step() function to build a better model reg2 = glm (diagnosis_b ~. + smoothness_mean:texture_se - smoothness_worst,data = cancer,family = “binomial”). The AIC is 88.891 which is lower than reg1’s AIC.

To further optimize the model, we use library(MASS) and step() function to build reg3 model: reg3 = glm(diagnosis_b ~ smoothness_mean + texture_se + area_se + smoothness_se + concave.points_se + texture_worst + concavity_worst + concave.points_worst + symmetry_worst + fractal_dimension_worst + smoothness_mean : texture_se , data = cancer , family = “binomial”). Reg3’s AIC is 85.7. After that, we use AIC(reg1,reg2,reg3) to compare AIC. We found a significant improvement in AIC from 88.891(reg1) to 85.7(reg3).

```
Call:
glm(formula = diagnosis_b ~ smoothness_mean + texture_se + area_se +
    smoothness_se + concave.points_se + texture_worst + concavity_worst +
    concave.points_worst + symmetry_worst + fractal_dimension_worst +
    smoothness_mean:texture_se, family = "binomial", data = cancer)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4045  -0.0229  -0.0018   0.0000   3.6976

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -13.3691     8.0498  -1.661  0.09675 .
smoothness_mean           -146.0558    85.3478  -1.711  0.08703 .
texture_se                 -16.0692     6.8597  -2.343  0.01915 *
area_se                      0.3054     0.0658   4.641 3.46e-06 ***
smoothness_se              374.5051   196.1210   1.910  0.05619 .
concave.points_se         -513.0842   177.3862  -2.892  0.00382 **
texture_worst                0.5365     0.1311   4.092 4.28e-05 ***
concavity_worst              8.7370     4.4401   1.968  0.04910 *
concave.points_worst        98.4159    24.8265   3.964 7.37e-05 ***
symmetry_worst              12.2663     8.1515   1.505  0.13238
fractal_dimension_worst    -92.7614    38.4166  -2.415  0.01575 *
smoothness_mean:texture_se 138.3552    66.8228   2.070  0.03841 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.44  on 568  degrees of freedom
Residual deviance:  61.70  on 557  degrees of freedom
AIC: 85.7

Number of Fisher Scoring iterations: 10
```

We build a confusion matrix to find out the actual and predicted value to see the total accuracy of the model. The confusion matrix is illustrated below:

The total accuracy (reg3) = (353+206)/569 = 0.9824. The accuracy of 1s = 98.24%

```
> AIC(reg1,reg2,reg3)
     df       AIC
reg1 15 92.90224
reg2 15 88.89105
reg3 12 85.70015
```

```
> confusion_matrix(reg3)
          Predicted 0 Predicted 1 Total
Actual 0          353           4   357
Actual 1            6         206   212
Total             359         210   569
```

In a nutshell, reg3 which has the lowest AIC is the best model in this logistic regression. The accuracy 98.24% is relatively high. Area_se, texture_worst and concave.points_worst are the top three variables that impact the outcome.

Recommendations

Due to the high degree of accuracy produced by our model it is recommended that hospitals considering incorporating analytics and data modeling into their diagnostic practices. The regression highlighted in this paper produced an accuracy of 98.24% and could prove to be extremely useful and accurate in real world diagnostic scenarios; resulting in less time and money spent diagnosing individual patients. Additionally, as time goes on and as more data becomes available information from fine needle aspirations could be built into additional models which determine nucleus size threshold for various types of cancer such as HER2 and BRCA1&2 mutations, each of which requires different treatment plans and whose successful treatment is highly dependent upon early diagnosis.. As such, treatment plans could be formed using the same information already being gathered and facilitate faster treatments & subsequently higher remission and survival rates.

References

Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)]. Irvine, CA: University of California, School of Information and Computer Science.

Goddard, K. A. B., Weinmann, S., Richert-Boe, K., Chen, C., Bulkley, J., & Wax, C. (2011). HER2 Evaluation and Its Impact on Breast Cancer Treatment Decisions. Public Health Genomics, 15(1), 1–10. http://doi.org/10.1159/000325746

Honein-AbouHaidar, G. N., Hoch, J. S., Dobrow, M. J., Stuart-McEwan, T., McCready, D. R., & Gagliardi, A. R. (2017). Cost analysis of breast cancer diagnostic assessment programs. *Current Oncology*, *24*(5), e354–e360. http://doi.org/10.3747/co.24.3608