**Predicting Signs of Malignant Breast Cancer:**

**a Logistic Regression Analysis**

Joshua Avery

Connor Davis

Sujata Duwal

David Stempnakowski

December 7, 2021

Prepared for

Dr. Rasim Musal

QMST 3339

Department of Computer Information Systems and Quantitative Methods

McCoy Hall Rm 404

601 University Drive

San Marcos Texas 78666

# **Executive Summary**

We performed a logistic regression analysis of ten variables in the Breast Cancer Wisconsin Data Set (K.P. Bennet, 2016) hosted on Kaggle using the R language (Team, 2021). Based on information gathered from an article by Dr. Susan Klein on breast masses (Susan Klein, 2005) and additional guidance from Dr. Theodore Drell (Regulatory Affairs, Bayer AG), we reduced the scope to three types of variables: compactness, concavity, and texture (Drell, 2021). Utilizing the forward model building methodology, we determined that the count of concave points on a mass and the texture score are the best predictors of a malignant tumor.

# **Problem Description**

This paper explores breast cancer data gathered on Kaggle to predict whether breast masses are benign or malignant. Dr. William H. Wolberg, Nick Street, and Olvi L. Mangasarian of the University of Wisconsin created the dataset using fine needle aspirations (FNAs) (William H. Wolberg, 1995). This FNA procedure is performed by inserting a small needle into the suspected mass and drawing out a small sample and is sent off for evaluation under a microscope (*Cytopathology: Fine Needle Aspiration (FNA)*, 2021). The dataset creators took the microscopic slides, created digitized images, and computed ten different features based on those images. The complete data set consisted of 569 observations of ten real-valued features calculated for each nucleus.

**Data Exploration**

We used RStudio as our data analysis and visualization tool for this project. The libraries used were "tidyverse" (Hadley Wickham, 2019) and "InformationValue" (Prabhakaran, 2016). The raw data had 569 rows and 32 columns. Since the creators had already cleaned the data, the only change we made was assigning '1' to malignant tumors and '0' to benign tumors in the diagnosis column. The data is slightly unbalanced, as seen in Figure 1. There are 212 malignant observations and 357 benign observations, giving us a malignant-benign ratio of about 0.5939.

Chart, bar chart

Description automatically generated

Figure 1: Balance of the Diagnosis

Given that we do not have experience in cytopathology, we consulted Dr. Theodore Drell, who performed his internship and pre-doctoral study in breast cancer research, to aid us in deciding which columns will provide us with the most information. After the consultation, we dropped twenty columns and used the remaining ten columns for computation. Dr. Drell suggested we investigate features related to compactness, concavity, and smoothness. Thus, we decided to explore smoothness\_mean, texture\_mean, and fractal\_dimension\_mean for measures of smoothness, concavity\_mean and concave.points\_mean for concavity, and compactness\_mean for compactness. Additionally, there seemed to be a high correlation between the size descriptors (i.e., radius\_mean, perimeter mean, and area\_mean), so we added these variables to our analysis[[1]](#footnote-2).

**Analysis**

Since our dependent variable, diagnosis, is binary, we want to select a framework that will classify our diagnosis as benign (0) or malignant (1). Thus, the options for the analysis are taking a naive Bayes approach, using a decision tree, or performing logistic regression. We chose logistic regression to facilitate the model-building process.

TRAINING

To prepare for analysis, we created a 30-70 test-train split of the dataset (398 training observations and 171 test observations) and decided on a required significance level of a = 0.01. We began by analyzing each of the nine variables individually. We created models with single variables and found the best predictor under each category by comparing Akaike's Information Criteria (AIC) for the features if we found them to be significant (p < 0.01). The best single predictors for each category were concave.points\_mean, texture\_mean, compactness\_mean, and perimeter\_mean for the concavity, smoothness, compactness, and size categories, respectively. We then used these features for the rest of the analysis.

When creating models with two or more variables, one must consider the possibility of multicollinearity. We assumed that a correlation greater than 0.8 was grounds for multicollinearity for our purposes. Thus, we had to rule out two of the six possible two-variable models: concavity with compactness and concavity with perimeter[[2]](#footnote-3). Additionally, this rules out using a three or four-variable model. Table 1 showcases the AIC scores for each of the remaining four possible models. We then selected the model with concavity and texture as our predictors since it scored the lowest AIC, even among the single-variable models.

|  |  |
| --- | --- |
| Model | AIC Score |
| concave.points\_mean + texture\_mean | 156.35 |
| compactness\_mean + textre\_mean | 184.07 |
| Perimeter\_mean + texture\_mean | 207.77 |
| Compactness\_mean + texture\_mean | 299.25 |

Table 1: Two-Variable Models and AIC Scores

TESTING

Now that we have obtained our best model, it is time to test it. First, we double-checked to make sure that each predictor in our model was still significant with the test data, and they were. Next, we examined the receiver operating characteristic (ROC) curve, pictured in Figure 1. The Area Under the ROC curve (AUROC) justifies the model's goodness-of-fit, which in this case was 0.9869. In other words, 98.69% of the time, the concavity-texture model will correctly assign a malignancy label to a tumor to a randomly selected tumor with higher levels of concave points and a higher texture score.

Graphical user interface, application

Description automatically generated

Figure 2: The Receiver Operating Characteristic Curve for the Selected Model

Next, we computed some statistics about our model's prediction ability. Since this dataset is unbalanced, we found the optimal point to start assigning malignancy to observations using the InformationValue library. The cutoff is 0.560. Once we obtained this value, we created a confusion matrix (pictured in Table 2) to see how our model performed. Lastly, we calculated the sensitivity, specificity, accuracy, and precisions for malignant and benign tumors (Table 3).

|  |  |  |
| --- | --- | --- |
|  | Actual Malignant Tumor | Actual Benign Tumor |
| Predicted Benign Tumor | 105 | 3 |
| Predicted Malignant Tumor | 3 | 60 |

Table 2: Confusion Matrix for the Selected Model

|  |  |
| --- | --- |
| **Sensitivity** | 95.24% |
| **Specificity** | 97.22% |
| **Accuracy** | 96.59% |
| **Precision for Malignant Tumors** | 95.24% |
| **Precision for Benign Tumors** | 97.22% |

Table 3: Key Model Statistics

# **Summary**

In this report, we have curated a model that takes data about a sample's concavity and texture and predicts whether a tumor is benign or malignant with 96% accuracy. Although not perfect, with an AUROC score of 98.7%, this model is a significantly better predictor of malignancy than the average rate of malignant tumors. However, due to the sensitive nature of cancer diagnosis, using this model as the *only* tool is not appropriate. Instead, cytopathologists should use this tool in symphony with the other standard tools used to diagnose a patient. Regardless, it will be a powerful addition to a cytopathologist's toolbox.

## **Appendix I: Code**

citation()

##########################################################

# #

# Group 14 Prediction of Breast Tumor Malignancy #

# #

##########################################################

library(tidyverse)

library(InformationValue)

# Read data file

# cancer\_data is the original data frame

df <- read.table(file = "data.csv", header = TRUE, sep = ",")

####################### DATA FILE CLEANING #######################

Diagnosis <- as.factor(df$diagnosis)

# Bar Graph of the data

df %>%

ggplot(mapping = aes(x = Diagnosis,

fill = Diagnosis)) +

geom\_bar() +

ggtitle(label = "Balance of Data") +

xlab("Diagnosis") +

ylab("Amount")

# Convert the diagnosis from M and B to 1 and 0

df$diagnosis <- ifelse(df$diagnosis == "M", 1, 0)

# Creates our data frame with only the columns of interest

drop\_col <- c(

"id", "radius\_worst", "texture\_worst", "perimeter\_worst", "area\_worst",

"smoothness\_worst", "compactness\_worst", "concavity\_worst",

"concave.points\_worst", "symmetry\_worst", "fractal\_dimension\_worst",

"radius\_se","texture\_se", "perimeter\_se", "area\_se", "smoothness\_se", "compactness\_se",

"concavity\_se", "concave.points\_se", "symmetry\_se","fractal\_dimension\_se"

)

# Full Data Set

cancer\_data <- df[, !(names(df) %in% drop\_col)]

# Ratio of malignant to benign tumors in this set

length(cancer\_data$diagnosis[cancer\_data$diagnosis == 1])/

length(cancer\_data$diagnosis[cancer\_data$diagnosis == 0])

### Dividing the data into train/test

set.seed(666)

n <- as.vector(1:(nrow(cancer\_data)))

n\_test <- nrow(cancer\_data)\*.3

n\_train <- nrow(cancer\_data) - n\_test

# Generate n\_train random numbers without replacement

train\_index <- sort(sample(n, n\_train))

test\_index <- sort(setdiff(n, train\_index))

# Create training and test dataframes

train <- cancer\_data[train\_index, ]

test <- cancer\_data[test\_index, ]

####################### MODEL BUILDING #######################

attach(train)

# According to research, compactness, concavity, and smoothness are indicators

# Forwards Model Building

cor(train)

### Single variables

Con1 <- glm(diagnosis ~ concavity\_mean, family = binomial(link = "logit"))

summary(Con1)

# Better concavity predictor, keep it

Con <- glm(diagnosis ~ concave.points\_mean, family = binomial(link = "logit"))

summary(Con)

# Better "smoothness" (texture) predictor, keep it

texture <- glm(diagnosis ~ texture\_mean, family = binomial(link = "logit"))

summary(texture)

smooth <- glm(diagnosis ~ smoothness\_mean, family = binomial(link = "logit"))

summary(smooth)

Frac <- glm(diagnosis ~ fractal\_dimension\_mean, family = binomial(link = "logit"))

summary(Frac)

# Also wanted to see if the tumor size is a good predictor since the

# correlations were high for those. I'll try perimeter first since the cor is highest.

perm <- glm(diagnosis ~ perimeter\_mean, family = binomial(link = "logit"))

summary(perm)

# Higher AIC

rad <- glm(diagnosis ~ radius\_mean, family = binomial(link = "logit"))

summary(rad)

# Also higher AIC

area <- glm(diagnosis ~ area\_mean, family = binomial(link = "logit"))

summary(area)

# Concavity and Compactness are highly correlated, so since

# AIC(Concavity) < AIC(compactness), leave out compactness to avoid

# multicollinearity, which can wreck the model

Comp <- glm(diagnosis ~ compactness\_mean, family = binomial(link = "logit"))

summary(Comp)

### Two variables

ConText <- glm(diagnosis ~ concave.points\_mean + texture\_mean, family = binomial(link = "logit"))

summary(ConText)

anova(Con, ConText, test = "Chisq") # Significant, nice

# AIC is way higher. Scrap.

CompText <- glm(diagnosis ~ compactness\_mean + texture\_mean, family = binomial(link = "logit"))

summary(CompText)

anova(texture, CompText, test = "Chisq") # Significant, but still scrap

# Try perimeter and texture

PermText <- glm(diagnosis ~ perimeter\_mean + texture\_mean, family = binomial(link = "logit"))

summary(PermText)

anova(perm, PermText, test = "Chisq") # Significant, but still scrap

# Compactness and Perimeter

CompPerm <- glm(diagnosis ~ perimeter\_mean + compactness\_mean, family = binomial(link = "logit"))

summary(CompPerm)

anova(perm, CompPerm, test = "Chisq") # Significant, but still scrap.

# AIC is much higher than concavity + texture

####################### TESTING #######################

detach(train)

attach(test)

# The best model

ConText <- glm(diagnosis ~ concave.points\_mean + texture\_mean, family = binomial(link = "logit"))

summary(ConText)

predicted <- predict(ConText,type="response")

predicted

# ROC curve

plotROC(actuals = diagnosis, predictedScores = predicted)

# Find the optimal cutoff for this set

optimal <- optimalCutoff(actuals = diagnosis, predictedScores = predicted)

# Actual: top-bottom, Predicted: left-right

confusionMatrix(actuals = diagnosis, predictedScores = predicted, threshold = optimal)

predicted <- (predict(ConText,type="response")>optimal)\*1

# Sensitivity P(Y\_hat=1|Y=1)

sum((predicted==1 & diagnosis==1))/sum(diagnosis==1)

# Specificity

sum((predicted==0 & diagnosis==0))/sum(diagnosis==0)

# Accuracy

(sum((predicted==1 & diagnosis==1))+sum((predicted==0 & diagnosis==0)))/length(diagnosis)

# Precision P(Y=1|Y\_hat=1)

sum((predicted==1 & diagnosis==1))/sum(predicted==1)

# Precision, but for 0s P(Y=0|Y\_hat=0)

sum((predicted==0 & diagnosis==0))/sum(predicted==0)

**Appendix II: Correlation Table on Training Data**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | diagnosis | radius | texture | perimeter | area | smoothness |
| diagnosis | 1 | 0.741 | 0.474 | 0.752 | 0.719 | 0.348 |
| radius[[3]](#footnote-4) | 0.741 | 1 | 0.376 | 0.998 | 0.987 | 0.159 |
| texture | 0.474 | 0.376 | 1 | 0.383 | 0.366 | 0.009 |
| perimeter | 0.752 | 0.998 | 0.384 | 1 | 0.986 | 0.197 |
| area | 0.719 | 0.987 | 0.366 | 0.986 | 1 | 0.168 |
| smoothness | 0.348 | 0.159 | 0.009 | 0.197 | 0.168 | 1 |
| compactness | 0.565 | 0.474 | 0.293 | 0.527 | 0.466 | 0.668 |
| concavity | 0.676 | 0.650 | 0.348 | 0.691 | 0.661 | 0.508 |
| concave.points | 0.779 | 0.819 | 0.362 | 0.849 | 0.819 | 0.545 |
| symmetry | 0.309 | 0.121 | 0.068 | 0.159 | 0.128 | 0.575 |
| fractal\_dimension | -0.047 | -0.331 | -0.066 | -0.279 | -0.301 | 0.596 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | compactness | concavity | concave.points | symmetry | fractal\_  dimension |
| diagnosis | 0.565 | 0.676 | 0.779 | 0.309 | -0.047 |
| radius | 0.474 | 0.650 | 0.819 | 0.121 | -0.331 |
| texture | 0.293 | 0.348 | 0.362 | 0.068 | -0.067 |
| perimeter | 0.527 | 0.691 | 0.849 | 0.159 | -0.279 |
| area | 0.466 | 0.661 | 0.819 | 0.128 | -0.301 |
| smoothness | 0.668 | 0.508 | 0.545 | 0.575 | 0.596 |
| compactness | 1 | 0.868 | 0.817 | 0.617 | 0.584 |
| concavity | 0.868 | 1 | 0.913 | 0.507 | 0.343 |
| concave.points | 0.817 | 0.913 | 1 | 0.467 | 0.160 |
| symmetry | 0.617 | 0.507 | 0.467 | 1 | 0.506 |
| fractal\_dimension | 0.584 | 0.343 | 0.160 | 0.506 | 1 |

**Appendix III: Data Dictionary**

Bennet and Mangasarian give ten real-valued features computed for each cell nucleus image:

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

**References**

*Cytopathology: Fine Needle Aspiration (FNA)*. (2021). AF Pathology Laboratories. Retrieved December 5 from <https://pathlabs.ufl.edu/tests/test-directory-c/cytopathology-fine-needle-aspiration-fna/>

Drell, D. T. L. (2021). Variables which may correlate or inversely correlate to malignancy. In D. A. Stempnakowski (Ed.).

Hadley Wickham, M. A., Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, Hiroaki Yutani. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. <https://doi.org/10.21105/joss.01686>

K.P. Bennet, O. L. M. (2016). *Breast Cancer Wisconsin (Diagnostic) Data Set* UCI Machine Learning. <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data?select=data.csv>

Prabhakaran, S. (2016). InformationValue: Performance Analysis and Companion Functions for Binary Classification Models. <https://CRAN.R-project.org/package=InformationValue>

Susan Klein, M. D. (2005). Evaluation of Palpable Breast Masses. *American Family Physician*, *71*, 1731-1738. <https://www.aafp.org/afp/2005/0501/p1731.html#:~:text=Malignant%20masses%20generally%20are%20hard,nonfixed%20masses%20can%20be%20cancerous>.

William H. Wolberg, W. N. S., Olvi L. Mangasarian. (1995). *Breast Cancer Wisconsin (Diagnostic) Data Set*. <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>

1. See Appendix III: Data Dictionary for a description of each variable. [↑](#footnote-ref-2)
2. See Appendix II: Correlation Table on Training Data for correlations between these variables. [↑](#footnote-ref-3)
3. “\_mean” was removed from the name of all features except diagnosis to fit the chart to the page. [↑](#footnote-ref-4)