CANCER PREDICTION ANALYSIS

1. Abstract

Breast cancer is one of the most common types of cancer diagnosed in humans. In the current time the causes are generally unknown due to which it is difficult to cure it. Early diagnosis is the best method to prevent and cure this disease. Accurate prediction of the type of tumor can assist in early diagnosis of breast cancer. This can help in providing better treatment to patients and reduce the mortality rate. In this project we develop a model with which we can predict if the breast cancer is malignant or benign. Various methods used in regression analysis are used for achieving this objective. The analysis results in a model that has a predictive ability of the diagnosis of a tumor being benign or malignant with an accuracy of approximately 79.86%.

2. Introduction

As reported by World Health Organization (WHO), in 2017, 7.6 million people died from cancer accounting for 13% of all deaths worldwide. The number of global cancer deaths is projected to increase by 45% between 2017 and 2030. Breast cancer is the one of the most prevalent type of cancer that affects women. The causes of cancer are generally unknown due to which it is difficult to implement prevention measures. It is generally accepted that early diagnosis of cancer can lead to effective treatment and can help the patient getting cured. Thus, the research question that we aim to answer is ~ Can we accurately predict the malignancy or benignancy of tumor? This can assist the doctors in reducing the deaths due to cancer, as early detection of malignant tumors can lead to effective treatment and even a possibility of full recovery of the patient. A tumor can be malignant or benign. We aim to develop a model that can predict the type of tumor with a decent accuracy rate.

It is important to derive key insights from the characteristics of cell nuclei. Features required for this study can be generated from a digitized image of a fine needle aspirate (FNA) of a breast mass. The purpose of this project is to develop a model which can predict whether the cancer is malignant or benign. Using various concepts like correlation, VIF, log-transformation, variable selection, outlier analysis and ROC curve, we developed a model for the diagnostics and prediction of cancer.

3. Data Set

The data was obtained from the National Cancer Institute GDC Data portal: https://portal.gdc.cancer.gov/. The data set consists of features which were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. The data has been generated using a computer vision diagnostic system that extracts ten different features from the snake-generated cell nuclei boundaries. It is a comprehensive dataset of 556 patients. There are multiple attributes available in the dataset. However, for the purposes of keeping the model as simple as possible while accounting for all the relevant attributes, we reviewed some literature[4][5] on cancer prediction and also consulted a specialist to determine the most appropriate attributes for our purpose. There are 33 variables in the entire data set which consists of the mean value, standard error and the worst possibility of each entry. We use the mean value for each data to model the data. The table below provides a summary of the dataset.

Attribute	Detail
Diagnosis	cancer ~ 1=benign or 0=malignant
Radius	mean of distances from center to points on the perimeter
Texture	standard deviation of gray-scale values
Perimeter	size of core of tumor
Area	surface area of tumour
Smoothness	local variation in radius lengths
Compactness	perimeter^2 / area - 1.0
Concavity	severity of concave portions of the contour
Concave points	number of concave portions of the contour
Symmetry	symmetricity of tumour around the core
Fractal dimension	"coastline approximation" - 1

Table 1: Dataset description

To understand some of the key features, we go through some diagrams for better understanding as follows:

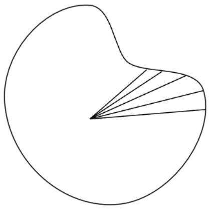


Figure 1: Radial lines used for smoothness

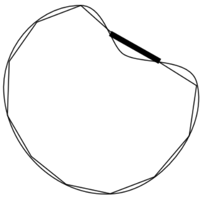


Figure 2: Chords used to compute concavity

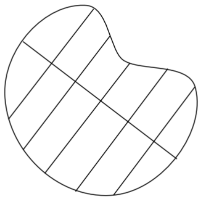


Figure 3: Symmetry segments used in symmetry computation

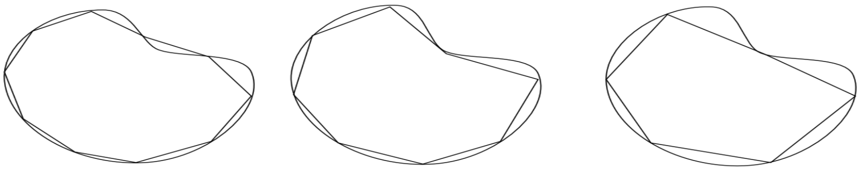


Figure 4: Sequence of measurements for computing fractal

4. Exploratory Data Analysis

An initial analysis of the dataset is performed as per the initial diagnostics that are performed in regression analysis. To begin with, we check for any missing values in the dataset as this may manipulate the analysis. We do not observe any missing values in this dataset which is a satisfactory result and would require no further action. Further, we decide to begin with a model containing all the features and perform a correlation analysis of the predictors to better understand the dataset and the behavior of the initial model.

The first model used for analysis is given below:

diagnosis ~ (radius + texture + perimeter + area + smoothness + compactness + concavity + concave.points + symmetry + fractal_dimension, family = binomial)

The figure below denotes the scatterplot that we obtain from our analysis.

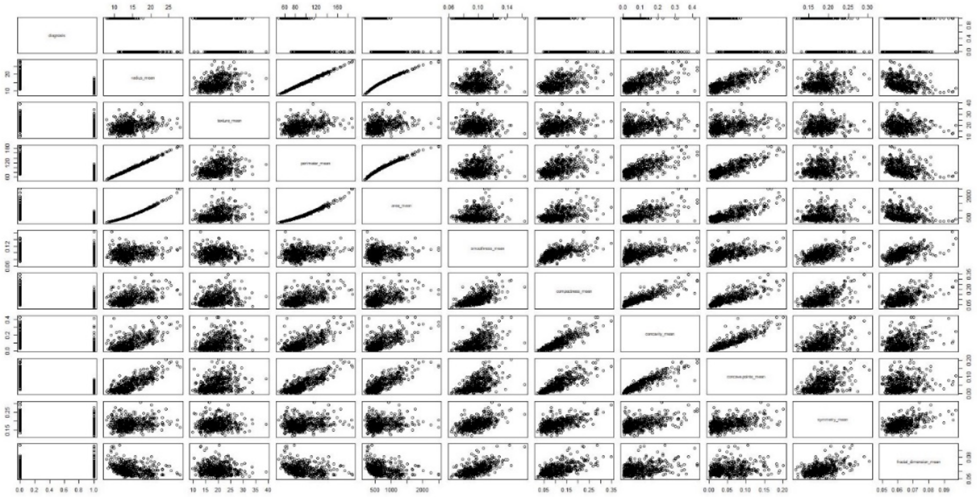


Figure 5: Scatterplot of variables

From the scatterplot analysis, we see high correlations amongst various predictors and hence decide to perform a further detailed study of the dataset and the model.

Additionally, we check the marginal model plots of the model shown in the figure below for better clarity of the data.

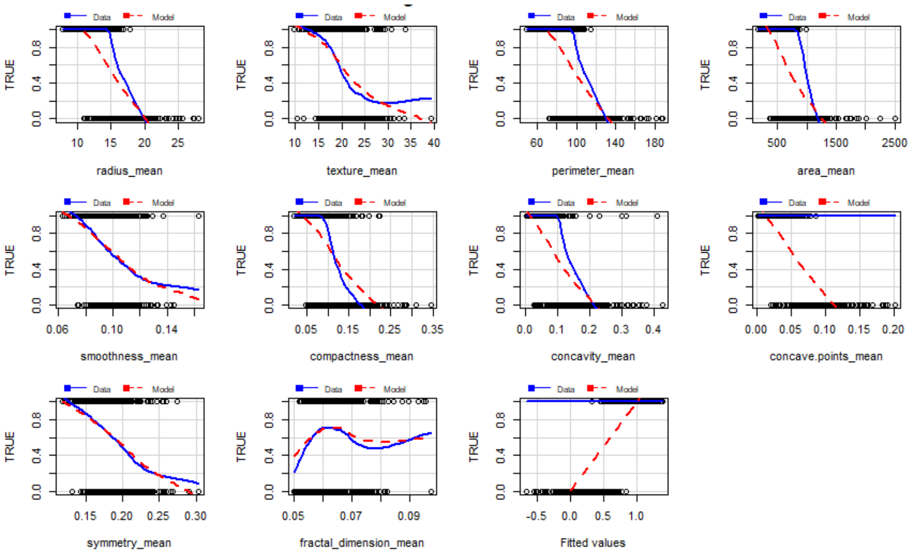


Figure 6: Marginal Model Plots for Model 1

As we observe from the figure above, the line obtained from the model doesn’t align with the non-parametric function line. This shows a poor fit and thus we decide to further analyze the data. In addition, we also calculate VIF values for all features. The table below shows the result.

Attribute	Radius	Texture	Perimeter	Area	Smoothness
VIF	1546.28	1.20	1890.47	59.11	2.99
Attribute	Compactness	Concavity	Concave points	Symmetry	Fractal dimension
VIF	22.70	11.52	21.59	1.79	6.58

Table 2: VIF analysis output

We follow the general heuristic that high VIF (>10) indicates higher correlation between predictors. Using this, we see that radius, perimeter, area, compactness, concavity and concave points have high value of VIF (>10) and using the heuristic we infer that the multi-collinearity issue exists in the dataset.

5. Modeling

5.1 Removing highly correlated variables:

The scatterplot analysis initially in the exploratory data analysis, showed that certain features such as radius, area and perimeter are highly correlated. According to the dataset we observed that radius is the lowest level of granularity in the data among these variables, while perimeter and area are derived from the radius feature. Hence, we drop the area and perimeter features and then check the behavior of the resulting model. The second model:

Diagnosis ~ (radius + texture + smoothness + compactness + concavity + concave.points + symmetry + fractal_dimension , family = binomial)

Plot for this model:

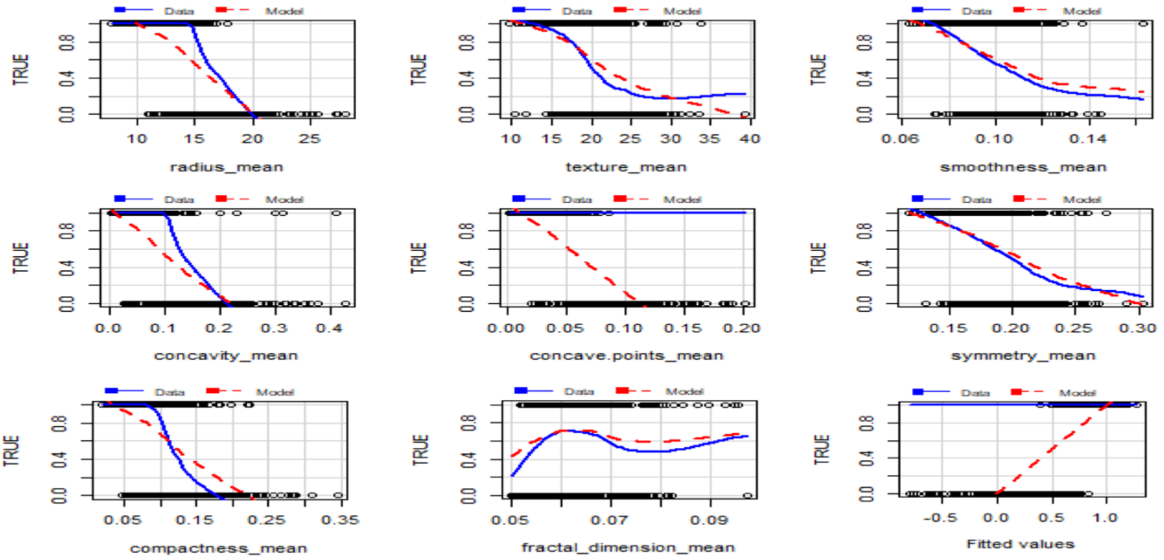


Figure 7: Marginal Model Plot for Model 2

5.2 Log Transformation:

The regression analysis performed up to this point have resulted in models that have features that are highly skewed. Hence, we decide to perform log transformations for the predictors and check the nature of the resulting model. Following is the third model:

Diagnosis ~ (logRadius + logTexture + logSmoothness + logCompact + logConcavity + logConcave + logSymmetry + logFractal)

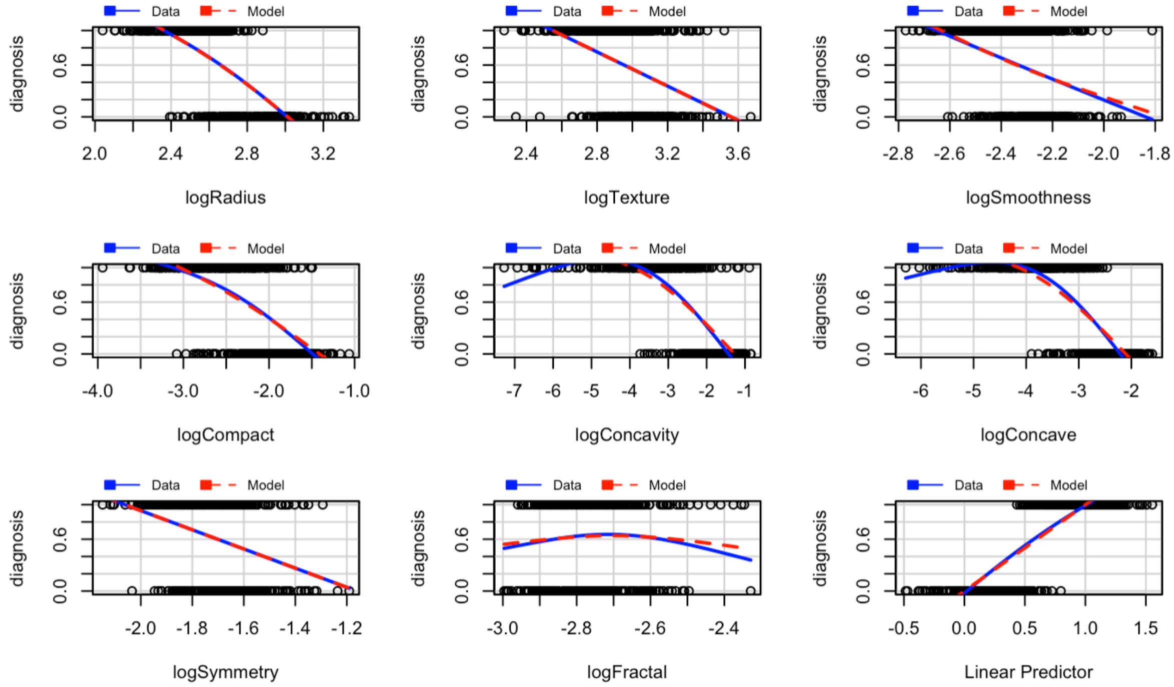


Figure 8: Marginal Model Plots for Model 2

From the marginal model plots above, we see that there has been substantial improvement in the model. Thus, we obtain a good model that fits the data better than previous models which were tested. However, we observe that the marginal model plot of the 'logSmoothness' feature still does not have a satisfactory fit.

5.3 Variable Selection:

Variable selection is the process of iteratively adding and removing predictors, in the predictive model, in order to find the subset of variables in the data set resulting in the best performing model which decreases prediction error.

There are three strategies of stepwise regression:

1. **Forward selection**: This starts with no predictors in the model, iteratively adds the most contributive predictors, and stops when the improvement is no longer statistically significant.
2. **Backward Elimination:** This starts with all predictors in the model (full model), iteratively removes the least contributive predictors, and stops when you have a model where all predictors are statistically significant.
3. **Stepwise Selection** (Both): This is a combination of forward and backward selections. We start with no predictors, then sequentially add the most contributive predictors (like forward selection). After adding each new variable, remove any variables that no longer provide an improvement in the model fit (like backward selection).

We decide to use the concepts of forward selection Akaike Information Criterion (AIC) and Stepwise selection AIC.

a. Using forward selection AIC, we obtain the model:
   Diagnosis ~ logConcave + logRadius + logTexture + logSymmetry + logSmoothness

b. Using stepwise selection AIC, we obtained the model:
   Diagnosis ~ logRadius + logCompact + logTexture + logSmoothness + logSymmetry

To finally determine the model as per the outputs of these methods, we consider the fact that our previous model has indicated that “compact” is a significant effect, and hence we decided to go for the model suggested by stepwise selection.

Thus, the fourth model is:

Diagnosis ~ logRadius + logCompact + logTexture + logSmoothness + logSymmetry

From the diagnostic analysis we see that our marginal model plot has improved to a great extent, except for the logSmoothness, which can be seen from the below figure:

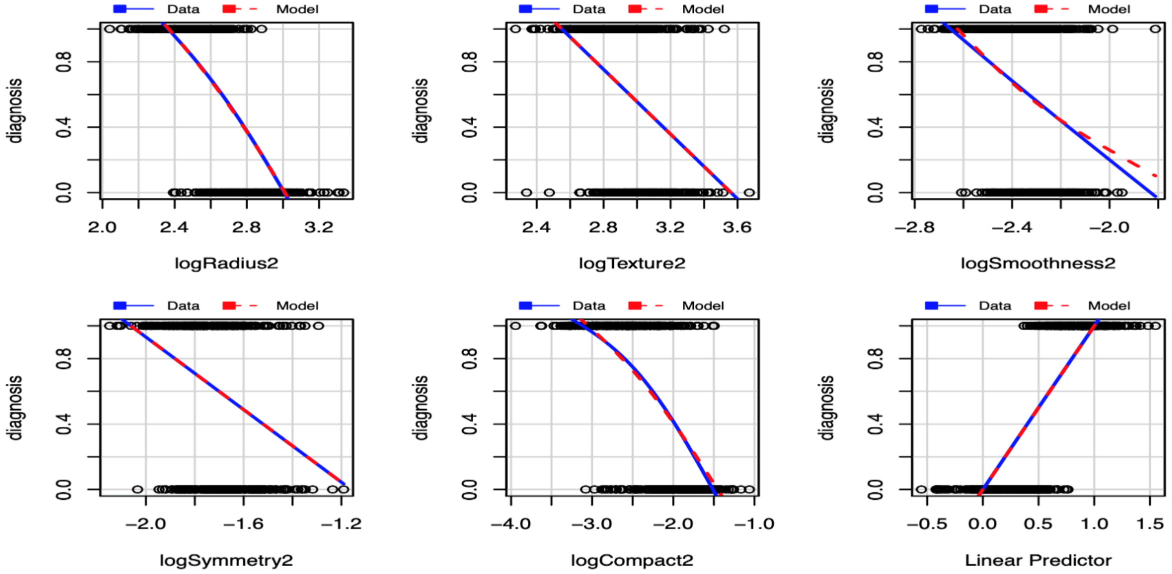
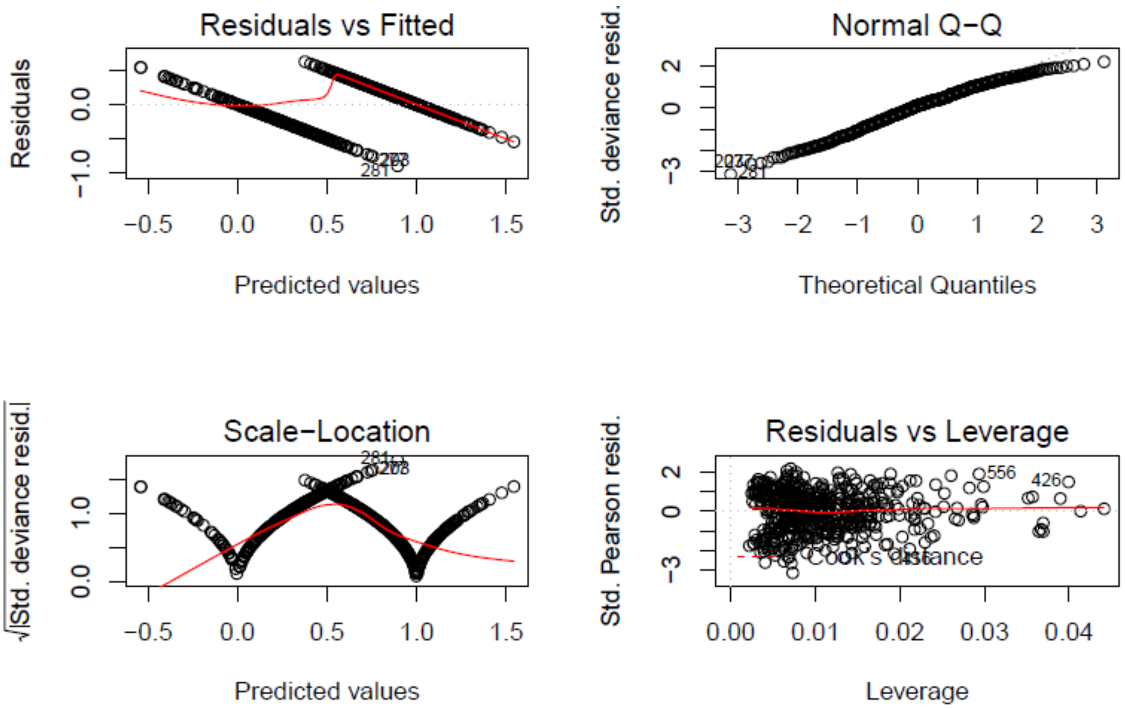


Figure 9: Marginal Model Plot for Model 4

5.4 Outlier Analysis:

We now look for potential outliers. To determine the reason and validity of these points we decided to analyze outliers as many of these points cause the models generated to deviate from the non-parametric fit as observed in the marginal model plots.

The R summary for the above mentioned model shows that all predictors are significant. However, we observe some of the outliers from the diagnostic plots.



Case 426, Case 556, Case 281, Case 207, Case 130. We consulted a specialist for a recommendation over the validity of these cases. Three of the data points had potential erroneous values for smoothness. Generally, ‘smoothness’ has a range [0.05, 0.15] while three of these had values > 0.5 and thus, Case 426, Case 556 and Case 281 can be ignored:

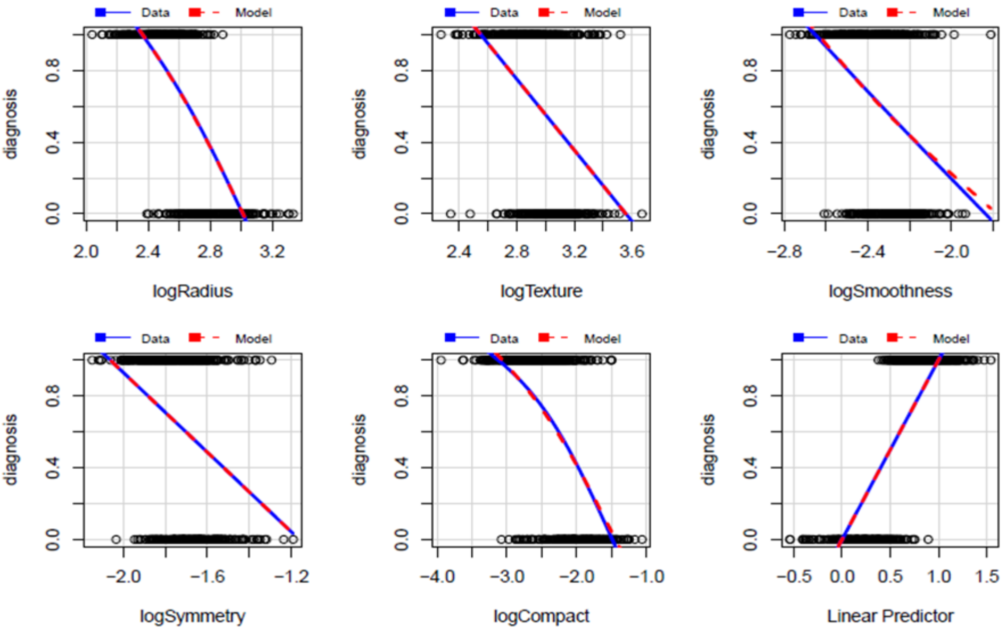


Figure 10: Marginal Model Plots after Outlier points removal

The above marginal model plots are obtained after the removal of the specified data points. It is observed that the plots show better fit. Further, the marginal model plot of ‘logSmoothness’ has improved as compared to previous results and as per our R-Summary, all our coefficients are significant, as seen in the table below. Hence, we have developed a good model.

Coefficients	Estimates	P - Value
Intercept	3.14156	2e-16 ***
logRadius2	-1.19781	2e-16 ***
logTexture2	-0.45025	1.87e-13 ***
logSmoothness2	-0.46991	0.000249 ***
logCompact2	-0.30813	0.002898 **
logSymmetry2	-0.13272	0.002362 **

Table 3: Regression analysis output for Model 6

5.5 Testing the Significance of Interactions:

For further analysis, we decide to test the interactions in the model. The sixth model considered is as follows:

Diagnosis ~ (logRadius2 + logTexture2 + logSmoothness2 + logSymmetry2 + logCompact2 + logSymmetry2: logCompact2)

We do not observe any specific improvement in the marginal model plots and hence use two approaches to find a better model:

1. **Analysis of Deviance**
   Deviance measures the discrepancy between the current model and the full model. The full model is the model that has *n* parameters, one parameter per observation. The full model maximizes the log-likelihood function. The full model provides a point of comparison for models with fewer than *n* parameters. Comparisons to the full model use the scaled deviance.

Residual.df	Residual.Dev	df	Deviance
547	43.324		
546	42.940	1	0.3834

Table 4: Analysis of Deviance

From the above output, we do not notice significant improvement by using the model 6.

2. **K-Fold Cross Validation**
   This analysis is done to compare the predictive performance of the 2 models and to assess whether adding the interaction term results in a significant improvement in prediction ability.

K = 10	Model 5	Model 6
PMSE Final	0.080	0.079

Table 5: K-Fold Cross Validation Results

The above results of K-fold cross validation show insignificant difference in predictive performance as seen by the value of the Prediction Mean Square Error (PMSE).

From both approaches, we could see that model 6 does not provide substantial improvement. Hence, we avoid using model 6 since it consists of additional interaction terms which leads to additional complexity in the model.

Hence, we decide to finalize model 5:

Diagnosis ~ logRadius + logCompact + logTexture + logSmoothness + logSymmetry

5.6 Receiver Operating Characteristic & Area Under Curve:

Area Under Curve (AUC) – Receiver Operating Characteristic (ROC) curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represent degree or measure of separability. It signifies the ability of the model to distinguish between classes. The higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between patients with disease and no disease.

We observe that the Area Under Curve (AUC) = 0.7986 for our model. This indicates that Model 5 predicts the values with 79.86% accuracy, which is a decent value.

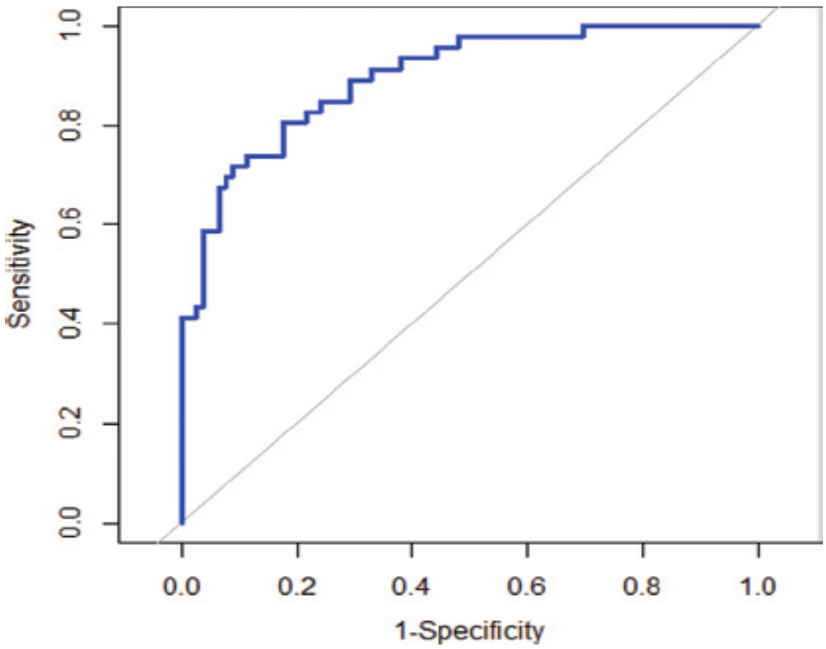


Figure 11: AUC output

6. Results

Regression analysis has been performed in an effort to develop a model that can predict the type of tumor as per the research question which we introduce here. We have built a predictive tool for cancer by creating a logistic regression model. Our final model is:

Diagnosis ~ logRadius + logCompact + logTexture + logSmoothness + logSymmetry

Using the Area Under Curve (AUC) as the performance metric for assessing the predictive ability of the model, we obtain an accuracy of 79.86% for diagnosing the type of tumor.

7. Discussions

The regression analysis performed for addressing the research question of accurately predicting the type of tumor has led to a model that can predict the diagnosis with 79.86% accuracy. However, the robustness of the model is limited by nature due to the limited data used for analysis and generation of the model. In the recent times there have been significant changes in the treatment of breast cancer. A recommendation from the specialist was to use some additional attributes or predictors such as cell wall thickness, clump thickness, epithelial cell size, bare nuclei etc. Modern datasets could be created for this purpose and the model can be extended to consider these factors. This could potentially improve the predictive ability of the model and be more useful to the medical community. Furthermore, for the purposes of this project we have applied regression methods for model development, however for the purposes of substantial and robust predictive ability, more sophisticated methods such as machine learning algorithms etc. can be used to achieve desired results.

8. Citations:

1. W.N. Street, W.H. Wolberg and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993.
2. M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. International Journal of Computer Vision, 1(4):321{331, 1988.
3. R. W. M. Giard and J. Hermans. The value of aspiration cytologic examination of the breast. A statistical review of the medical literature. Cancer, 69:2104{2110, 1992}
4. http://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH/PH709_Cancer/PH709_Cancer7.html
5. https://cancerres.aacrjournals.org/content/60/14/3683