Breast Cancer Prediction

Aash Makwana

27 September, 2024

UNIVERSITY OF
CALGARY

University of Calgary, Mathematics and Statistics Department

Contents

List of Figures

List of Tables

1. Introduction

Breast cancer is one of the most common cancers globally, and it remains a major health concern for women worldwide. It is the leading cause of cancer-related deaths among women, although survival rates have improved significantly over the past few decades due to advances in early detection, treatment, and awareness. Breast cancer occurs when cells in the breast tissue begin to grow uncontrollably. These cells can form a lump or mass, which may be malignant (cancerous) or benign (non-cancerous).

There are two main types of breast cancer: Invasive Ductal Carcinoma (IDC): IDC is the most common form of breast cancer, making up approximately 70-80% of all cases. It begins in the milk ducts and then invades surrounding tissue, Invasive Lobular Carcinoma (ILC): ILC starts in the milk-producing lobules of the breast and can spread to other parts of the body. Other less common types include inflammatory breast cancer, triple-negative breast cancer, and ductal carcinoma in situ (DCIS), which is a non-invasive cancer confined to the ducts.

The risk of developing breast cancer is influenced by various factors, both modifiable and non-modifiable. Non-modifiable factors include age, gender, family history, genetic mutations (such as BRCA1 and BRCA2), and a personal history of breast cancer. Modifiable risk factors include lifestyle factors such as diet, physical activity, alcohol consumption, and hormonal treatments like hormone replacement therapy (HRT), Genetic Mutations: Mutations in genes such as BRCA1, BRCA2, and TP53 significantly increase the risk of breast cancer. Individuals with a family history of these genetic mutations are at a higher risk, Age: The risk of breast cancer increases with age, with the majority of cases diagnosed in women over the age of 50, Hormonal Factors: Prolonged exposure to estrogen (e.g., early menarche, late menopause, and hormone replacement therapy) can increase the risk of breast cancer.

Early detection of breast cancer is critical for improving treatment outcomes. Several diagnostic techniques are employed to detect and diagnose breast cancer: - Mammography: This is the most commonly used screening tool. It involves X-ray imaging of the breast tissue to detect abnormalities such as lumps or calcifications. - Ultrasound: An ultrasound uses sound waves to create images of the internal structures of the breast, helping to differentiate between solid tumors and fluid-filled cysts. - Biopsy: A biopsy involves removing a small sample of tissue from the breast for examination under a microscope. It is the definitive method for diagnosing whether a lump is malignant or benign.

The treatment of breast cancer depends on various factors, including the stage of cancer, the type of breast cancer, the patient's age, and overall health. Common treatment options include: Surgery: The goal of surgery is to remove the tumor and, in some cases, the entire breast (mastectomy), Radiation Therapy: High-energy rays are used to target and destroy cancer cells, Chemotherapy: Chemotherapy drugs are used to kill or slow the growth of cancer cells. It can be administered before (neoadjuvant) or after (adjuvant) surgery.

While machine learning models can improve the accuracy of breast cancer diagnosis, several challenges remain: - Data Quality: High-quality data is crucial for model accuracy. Missing data, noisy variables, or unbalanced datasets can lead to overfitting or biased predictions. - Interpretability: Many machine learning models, particularly deep learning models, are considered "black boxes," meaning their predictions are hard to interpret. This makes it challenging to understand why a model classifies a tumor as malignant or benign. - Clinical Integration: Integrating prediction models into clinical practice requires robust validation and regulatory approval to ensure they are safe and effective for patient care.

Breast cancer remains a major public health issue, but advancements in medical research, diagnostic technologies, and treatment options have significantly improved survival rates. Machine learning methods provide powerful tools for predicting whether a breast tumor is malignant or benign, which can assist clinicians in making more informed decisions. The combination of traditional medical expertise and machine learning holds promise for improving the accuracy and timeliness of breast cancer diagnoses, ultimately leading to better patient outcomes.

2. Description of the dataset

The Breast Cancer Wisconsin (Diagnostic) Dataset is a collection of data points that are used to classify whether a tumor is benign or malignant based on various features derived from digitized images of breast mass samples. The dataset was initially gathered by Dr. William H. Wolberg from the University of Wisconsin Hospitals, Madison, and it has been widely used for machine learning and data analysis tasks, particularly in binary classification problems.

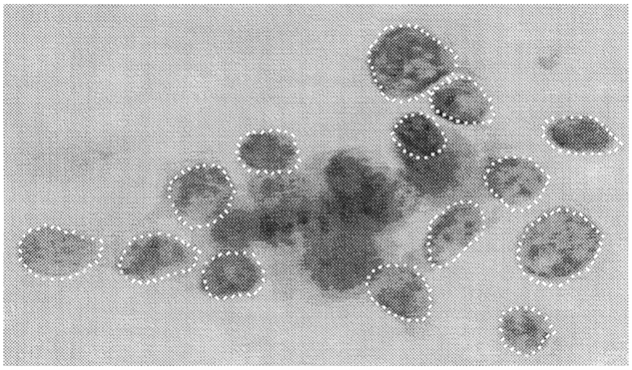


Figure 1: Initial Approximate Boundaries of Cell Nuclei

This dataset is designed to provide a foundation for building predictive models that can assist medical professionals in identifying whether breast cancer cells are malignant or benign based on a set of measurable tumor characteristics.

The dataset consists of 569 instances, each representing a breast cancer sample. Each instance includes 10 real-valued feature columns, which are numerical values derived from the digitized image of a breast mass. These 10 real-valued data is further modified to 30 real0-valued features. For example: The data collected for radius of cell nuclei is modified to radius_mean (mean of radius), radius_se (standard deviation of radius) and radius_worst (mean of 3 largest values of radius.) These features describe various aspects of the tumor, such as its size, texture, smoothness, and shape. The dataset is typically used for a classification task where the goal is to predict whether the tumor is "Malignant" (M) or "Benign" (B). Most of the features in this dataset are numerical and continuous, and they are scaled on a different range, making normalization or standardization important when building machine learning models. Below table 1 shows the 10 main real-valued features, in original dataset this features are modified to 30 features, _mean, _se and _worst as explained above.

Table 1: Table containing variables and their description

No	Variable	Type	Description
1	id	Categorical	ID number
2	diagnosis	Binary	(target) M = malignant, B = benign
3	radius_mean	Continuous	radius: mean of distances from centre to points on the perimeter
4	texture_mean	Continuous	texture: standard deviation of gray-scale values
5	perimeter_mean	Continuous	perimeter
6	area_mean	Continuous	area
7	smoothness_mean	Continuous	smoothness: local variation in radius lengths
8	compactness_mean	Continuous	compactness: perimeter^2 / area - 1.0
9	concativity_mean	Continuous	concavity: severity of concave portions of the contour
10	concave_points_mean	Continuous	concave points: number of concave portions of the contour
11	symmetry_mean	Continuous	symmetry

3. Question of Interest

The central aim of this project is to develop a predictive model that can accurately classify whether a breast cancer tumor is malignant or benign based on a set of tumor characteristics. Early detection and accurate classification of breast cancer are crucial for effective treatment and improved patient outcomes. With advancements in machine learning and data science, predictive models can assist healthcare professionals in making more informed decisions, offering the potential for earlier intervention and personalized treatment plans.

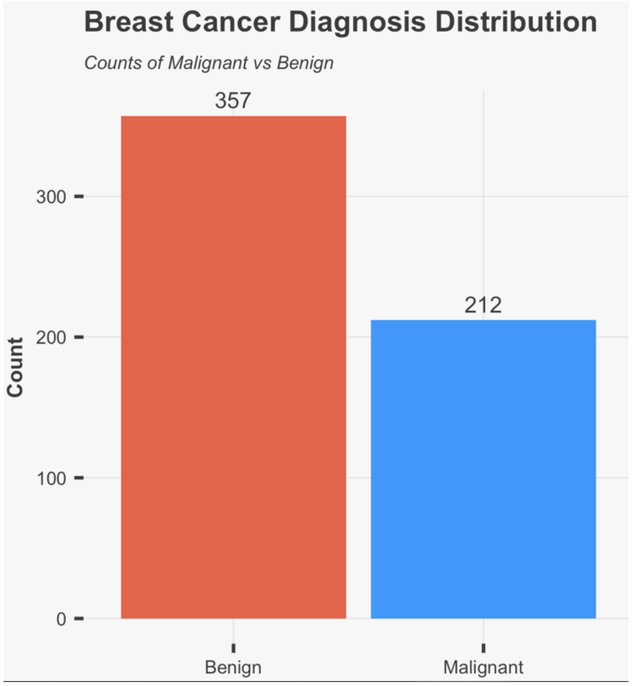


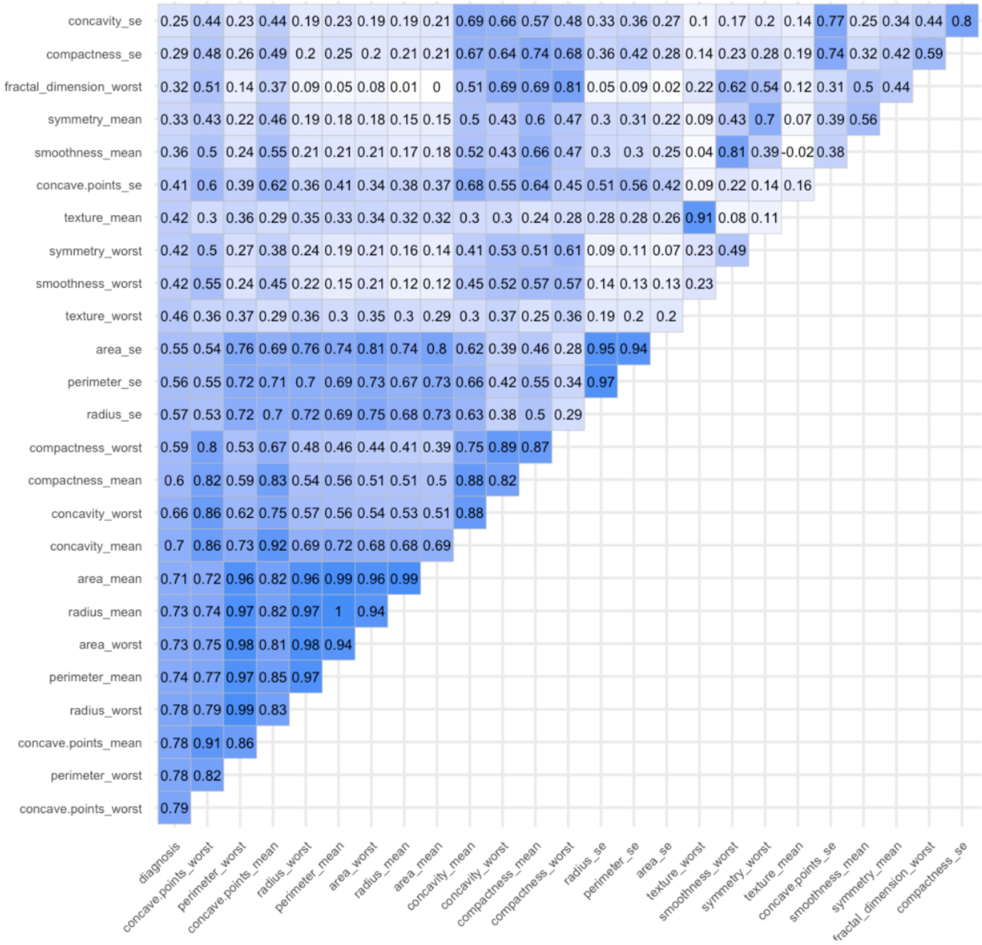
Figure 2: Distributions of the Breast Cancer Diagnosis

Out of 569 observations, there are 357 benign samples and 212 malignant samples in the dataset, meaning that the data is somewhat imbalanced, with more malignant instances than benign.

4. Exploratory Data Analysis

4.1 Correlation Matrix

The correlation matrix helps to understand the relationships between different variables. In a dataset with multiple features, some features may be highly correlated with each other, which can affect model performance (for example, causing multicollinearity in linear models). It's important to identify which features are highly correlated, spot any potential redundancy between features, guide feature selection or dimensionality reduction if needed.



	diagnosis	concave.points_worst	perimeter_worst	concave.points_mean	radius_worst	perimeter_mean	area_worst	radius_mean	area_mean	concavity_mean	concavity_worst	compactness_mean	compactness_worst	radius_se	perimeter_se	area_se	texture_worst	smoothness_worst	symmetry_worst	texture_mean	concave.points_se	smoothness_mean	symmetry_mean	fractal_dimension_worst	compactness_se
concavity_se	0.25	0.44	0.23	0.44	0.19	0.23	0.19	0.19	0.21	0.69	0.66	0.57	0.48	0.33	0.36	0.27	0.1	0.17	0.2	0.14	0.77	0.25	0.34	0.44	0.8
compactness_se	0.29	0.48	0.26	0.49	0.2	0.25	0.2	0.21	0.21	0.67	0.64	0.74	0.68	0.36	0.42	0.28	0.14	0.23	0.28	0.19	0.74	0.32	0.42	0.59	
fractal_dimension_worst	0.32	0.51	0.14	0.37	0.09	0.05	0.08	0.01	0	0.51	0.69	0.69	0.81	0.05	0.09	0.02	0.22	0.62	0.54	0.12	0.31	0.5	0.44		
symmetry_mean	0.33	0.43	0.22	0.46	0.19	0.18	0.18	0.15	0.15	0.5	0.43	0.6	0.47	0.3	0.31	0.22	0.09	0.43	0.7	0.07	0.39	0.56			
smoothness_mean	0.36	0.5	0.24	0.55	0.21	0.21	0.21	0.17	0.18	0.52	0.43	0.66	0.47	0.3	0.3	0.25	0.04	0.81	0.39	-0.02	0.38				
concave.points_se	0.41	0.6	0.39	0.62	0.36	0.41	0.34	0.38	0.37	0.68	0.55	0.64	0.45	0.51	0.56	0.42	0.09	0.22	0.14	0.16					
texture_mean	0.42	0.3	0.36	0.29	0.35	0.33	0.34	0.32	0.32	0.3	0.3	0.24	0.28	0.28	0.28	0.26	0.91	0.08	0.11						
symmetry_worst	0.42	0.5	0.27	0.38	0.24	0.19	0.21	0.16	0.14	0.41	0.53	0.51	0.61	0.09	0.11	0.07	0.23	0.49							
smoothness_worst	0.42	0.55	0.24	0.45	0.22	0.15	0.21	0.12	0.12	0.45	0.52	0.57	0.57	0.14	0.13	0.13	0.23								
texture_worst	0.46	0.36	0.37	0.29	0.36	0.3	0.35	0.3	0.29	0.3	0.37	0.25	0.36	0.19	0.2	0.2									
area_se	0.55	0.54	0.76	0.69	0.76	0.74	0.81	0.74	0.8	0.62	0.39	0.46	0.28	0.95	0.94										
perimeter_se	0.56	0.55	0.72	0.71	0.7	0.69	0.73	0.67	0.73	0.66	0.42	0.55	0.34	0.97											
radius_se	0.57	0.53	0.72	0.7	0.72	0.69	0.75	0.68	0.73	0.63	0.38	0.5	0.29												
compactness_worst	0.59	0.8	0.53	0.67	0.48	0.46	0.44	0.41	0.39	0.75	0.89	0.87													
compactness_mean	0.6	0.82	0.59	0.83	0.54	0.56	0.51	0.51	0.5	0.88	0.82														
concavity_worst	0.66	0.86	0.62	0.75	0.57	0.56	0.54	0.53	0.51	0.88															
concavity_mean	0.7	0.86	0.73	0.92	0.69	0.72	0.68	0.68	0.69																
area_mean	0.71	0.72	0.96	0.82	0.96	0.99	0.96	0.99																	
radius_mean	0.73	0.74	0.97	0.82	0.97	1	0.94																		
area_worst	0.73	0.75	0.98	0.81	0.98	0.94																			
perimeter_mean	0.74	0.77	0.97	0.85	0.97																				
radius_worst	0.78	0.79	0.99	0.83																					
concave.points_mean	0.78	0.91	0.86																						
perimeter_worst	0.78	0.82																							
concave.points_worst	0.79																								

Figure 3: Correlation plot

4.2 Visualizing the data distributions

For below plots, I have only plotted the ones which are not highly correlated as we are going to remove multicollinearity during Data preprocessing.

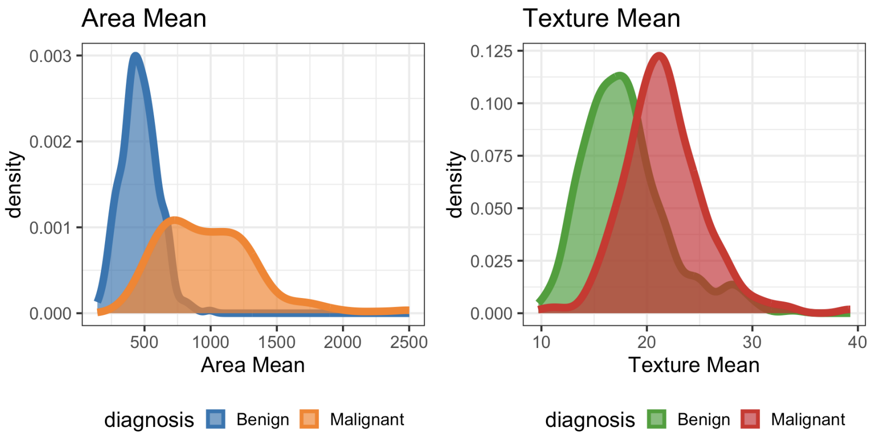


Figure 4: Distributions of the quantitative variables

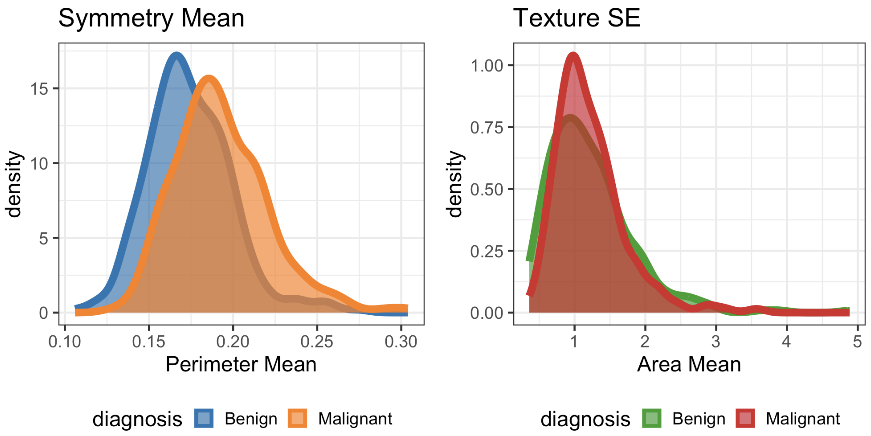


Figure 5: Distributions of the quantitative variables

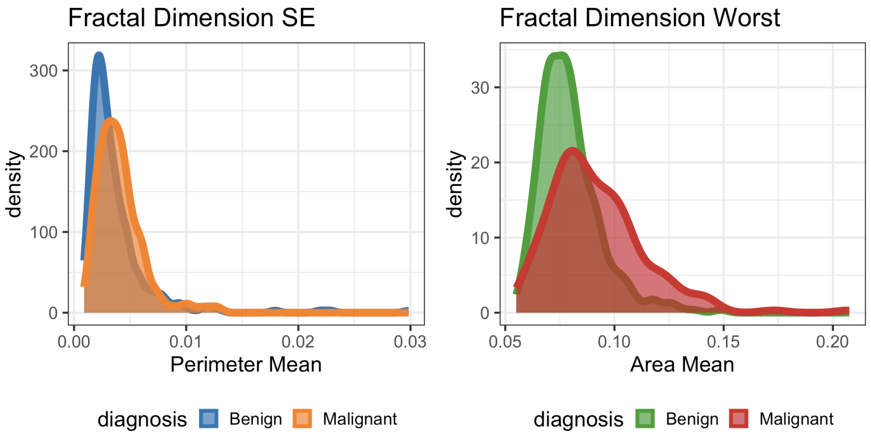


Figure 6: Distributions of the quantitative variables

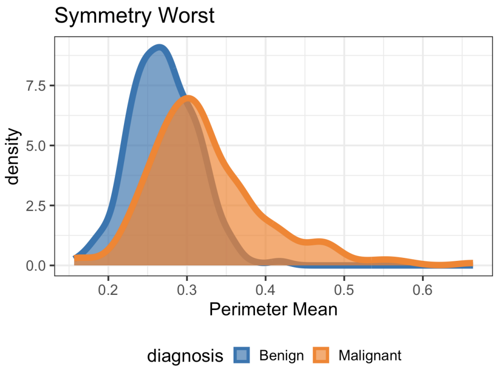


Figure 7: Distributions of the quantitative variables

5 Data Preprocessing

First step was to check if there are any missing values, but the data set doesnot contain any missing values. Now we will move further to handle highly correlated features.

5.1 Multicollinearity

In the process of building a predictive model, especially when using regression-based algorithms like Logistic Regression, one important aspect to consider is multicollinearity. Multicollinearity occurs when two or more independent features in a dataset are highly correlated with each other. This can cause several issues in the model-building process, particularly for Logistic Regression, which relies on the assumption that the predictors are not highly correlated.

To mitigate the impact of multicollinearity, I will remove highly collinear features before fitting the Logistic Regression model. Specifically, I will calculate the correlation matrix of the features and identify pairs of features with a correlation coefficient greater than 0.8. These highly correlated features will be removed, as they contribute redundant information to the model. The threshold of 0.8 is commonly used to identify and remove features that are highly correlated. This value is chosen to capture strong correlations, while leaving moderate correlations (such as 0.6 or 0.7) intact. By removing features with a correlation greater than 0.8, we reduce redundancy and retain only the most relevant predictors, ensuring that the model remains stable and interpretable.

New Data set after removing highly correlated features we are left with this final features:

diagnosis, area_mean, symmetry_mean, fractal_dimension_mean, radius_se, texture_se, smoothness_se, concavity_se, concave.points_se, symmetry_se, fractal_dimension_se, texture_worst, smoothness_worst, symmetry_worst, fractal_dimension_worst.

5.2 Cross-Validation

Cross-validation is a key technique for assessing the performance of a machine learning model while minimizing the risk of overfitting. It helps ensure that the model generalizes well to unseen data and is not merely memorizing the training data. In this project, we perform k-fold cross-validation to evaluate the performance of our predictive model for breast cancer classification. Cross-validation involves partitioning the data into multiple subsets (or "folds") and then training and testing the model multiple times. The model is trained on several different subsets of the data and tested on the remaining fold. This process is repeated for each fold, and the overall performance is averaged to provide a more robust estimate of model performance. In this project, I have chosen to use k = 10 folds for cross-validation. This is a common choice because: It strikes a balance between bias and variance: With 10 folds, we get a sufficient number of training and testing iterations, reducing the bias in model performance estimates while still keeping computation time manageable, It provides a stable estimate of model performance: Using 10 folds is a widely accepted practice and tends to give reliable and consistent results for many types of models.

Below is the list of 10 folds:

Fold01: int [1:57] 9 11 22 39 58 79 85 108 127 132 ...
Fold02: int [1:58] 30 52 63 71 82 95 96 106 117 131 ...
Fold03: int [1:57] 3 13 19 36 51 55 70 74 78 94 ...
Fold04: int [1:57] 2 10 34 40 41 48 49 56 87 89 ...
Fold05: int [1:56] 16 18 28 42 44 53 66 67 84 86 ...
Fold06: int [1:57] 21 38 47 69 76 81 97 104 120 121 ...
Fold07: int [1:56] 6 12 17 23 37 77 90 91 99 116 ...
Fold08: int [1:56] 5 14 20 32 33 35 45 61 72 80 ...
Fold09: int [1:57] 1 7 8 24 31 43 54 59 62 73 ...
Fold10: int [1:58] 4 15 25 26 27 29 46 50 57 60 ...

6. Statistical Analysis

In this phase of the project, I will train and evaluate several machine learning models to predict whether a breast cancer tumor is malignant or benign. Specifically, I will focus on Logistic Regression, a widely-used model for binary classification, and incorporate Elastic Net regularization, which combines both Lasso (L1) and Ridge (L2) regularization techniques. Regularization helps prevent overfitting by penalizing overly complex models, improving generalization to unseen data. I will compare the performance of these models using cross-validation and assess their effectiveness in accurately predicting the malignancy of tumors.

6.1 Logistic Regression

In this stage of the project, we will apply Logistic Regression to predict whether a breast cancer tumor is malignant or benign. Logistic Regression is a widely-used statistical method for binary classification, where the goal is to model the probability of a binary outcome based on a set of predictor variables.

To ensure the model performs well and avoids issues related to multicollinearity, we will use only the features that remain after removing highly collinear variables. Multicollinearity can lead to unstable estimates of the regression coefficients, which can affect the model's performance and interpretability. By removing correlated features (those with a correlation coefficient above 0.8), we aim to improve the stability and reliability of the Logistic Regression model. The model will be trained using cross-validation to assess its generalization ability, ensuring that it is not overfitting to the training data.

For this logistic model, the formula of the model is

$$\log\frac{p}{1-p} = \beta_0 + \sum_{i=1}^{14} \beta_i x_i \tag{1}$$

where p is the probability to get the heart disease, and β_i, $i = 1, 2, \ldots, 14$ is the coefficients of the parameter, β_0 is the intercept. So we fitted the model, and got the estimated parameters shown in the table below.

Table 2: Summary of the full Logistic regression model

Variable	Coefficient	Std. Error	p-value
Intercept	-5.561e+01	5.40e-06	***
area_mean	1.894e-02	2.48e-06	***
symmetry_mean	-2.287e+00	0.928164	Not Significant
fractal_dimension_mean	-6.664e+01	0.644640	Not Signifciant
radius_se	2.115e+01	0.000768	***
texture_se	-2.336e+00	0.078993	Not Significant
smoothness_se	9.413e+01	0.643839	Not Significant
concavity_se	1.668e+01	0.272138	Not Significant
concave.points_se	3.646e+02	0.021276	*
symmtery_se	-3.044e+01	0.743683	Not Significant
fractal_dimension_se	-1.790e+03	0.002511	**
texture_worst	5.389e-01	0.000114	***
smoothness_worst	8.063e+01	0.013170	*
symmetry_worst	1.551e+01	0.263348	Not Significant
fractal_dimension_worst	1.746e+02	0.031778	*

As we can see from the table, there are some variables that are not significant. I will drop some variables to make the model simpler. I choose backward step selection to do so. The new logistic regression model equation is given by,

$$\log\frac{p}{1-p} = \beta_0 + \sum_{i=1}^{7} \beta_i y_i \tag{2}$$

where p is the probability to get the heart disease, and β_i, $i = 1, 2, \ldots, 14$ is the coefficients of the parameters (area_mean, radius_se, concave.points_se, fractal_dimension_se, texture_worst, smoothness_worst, fractal_dimension_worst), β_0 is the intercept. So we fitted the model, and got the estimated parameters shown in the table below.

Table 3: Summary of the full Logistic regression model

Variable	Coefficient	Std. Error	p-value
Intercept	-5.288e+01	2.17e-10	***
area_mean	2.046e-02	8.06E-09	***
radius_se	1.355e+01	0.000364	***
concave.points_se	3.999e+02	0.000849	***
fractal_dimension_se	-1.792e+03	1.49e-05	***
texture_worst	3.738e-01	1.59e-07	***
smoothness_worst	6.730e+01	0.003229	**
fractal_dimension_worst	2.033e+02	6.67e-06	***

Now for statistical analysis we will use different types of methods consisting of Model Accuracy, Receiver operating characteristics (ROC), Psuedo-R square, Deviance Goodness of Fit, The Akaike information criterion (AIC) and Binned Residual Plots.

6.1.1 Model Accuracy

In logistic regression, accuracy is a commonly used metric to evaluate the performance of the model. It measures the proportion of correctly predicted instances (both true positives and true negatives) out of all predictions made by the model. However, accuracy alone may not provide a complete picture, especially when dealing with imbalanced datasets.

A confusion matrix is a more detailed tool for assessing the performance of a logistic regression model. It displays the counts of the following four categories, true positives, true negative, false positives, false negatives. From the R code I calculated model accuracy,

Table 4: Confusion matrix

		Actual Class	
		Positve (Malignant)	Negative (Benign)
Predicted Class	Positive (Malignant)	True Positive (TP) = 349	False Positive (FP) = 8
	Negative (Benign)	False Negative (FN) = 8	True Negative (TN) = 204

Different statistics can be calculated using the confusion matrix as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = 0.9719 \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} = 0.9623 \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} = 0.9623 \tag{5}$$

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 0.9623 \tag{6}$$

- Accuracy: 0.9719, Accuracy is the proportion of correct predictions (both true positives and true negatives) out of all predictions made by the model. An accuracy of 0.9719 means that approximately

97.19% of the predictions made by your logistic regression model were correct. This indicates that the model is performing very well in terms of overall prediction accuracy. The model is correctly predicting almost 97% of all cases, which suggests that the model is performing well overall.

- F1 Score: 0.9623, The F1 score is the harmonic mean of precision and recall, providing a balanced measure of the model's ability to correctly identify positive cases (true positives) and avoid false positives and false negatives. An F1 score of 0.9623 indicates that the model is performing very well with a strong balance between precision and recall. The F1 score is also very high, indicating that the model is effectively balancing precision and recall, meaning that it not only predicts positives accurately (precision) but also identifies most of the positive cases (recall).

6.1.2 Receiver operating characteristics (ROC)

The Receiver Operating Characteristic (ROC) curve is a graphical representation of the performance of a classification model at all classification thresholds. It is widely used for evaluating binary classifiers like logistic regression.

The ROC curve plots the following two metrics:

- True Positive Rate also known as Recall or Sensitivity
- False Positive Rate

The best possible classifier will have a TPR (sensitivity) of 1 and a FPR (fallout) of 0, meaning it correctly classifies all positive cases and never misclassifies a negative case as positive. This would result in a point in the top-left corner of the ROC curve.The closer the ROC curve is to the top-left corner, the better the model. A curve that bows toward the upper left indicates that the model has a high true positive rate and a low false positive rate.

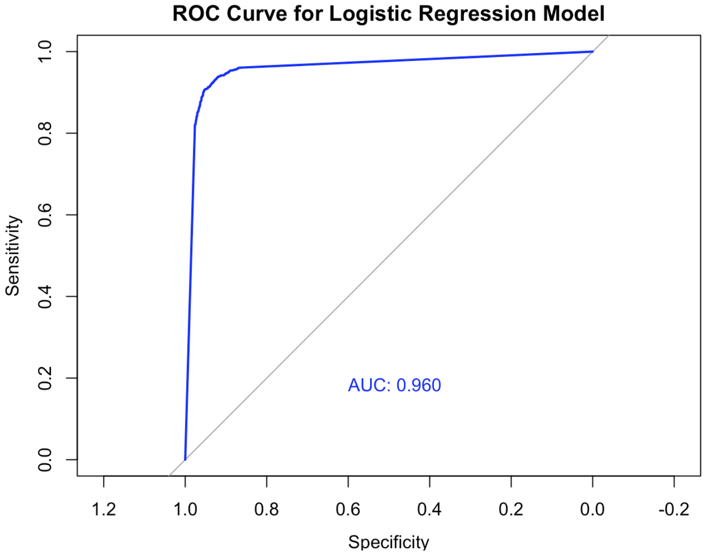


Figure 8: ROC curve of Logistic Regression

Area under the curve: AUC (Area Under the Curve) is a numerical summary of the ROC curve. It quantifies the overall ability of the model to distinguish between positive and negative cases. A higher AUC indicates a better model. In your case, if the ROC curve shows an AUC close to 1, it suggests that your logistic regression model has excellent discriminatory power.

- For our trained model, AUC = 0.960

6.1.3 Deviance Goodness of Fit

In this step, I calculated the residual deviance for the final models. We have the null hypothesis that is - H_0: the model fits well vs H_1: the model fits poorly.

In logistic regression, deviance is a measure of the goodness of fit of a model. It is derived from the likelihood ratio test and compares the likelihood of the model in question to the likelihood of a baseline model. The null deviance represents the deviance of a model that includes no predictors, i.e., a model that only predicts the mean of the outcome variable (the baseline model). This is a measure of how well the outcome variable can be predicted with no information other than the overall distribution of the target variable (e.g., class probabilities for logistic regression). The residual deviance represents the deviance of the fitted logistic regression model, which includes the predictors (features). It measures the goodness of fit of the model with all the predictors included. The deviance goodness of fit is the difference between the null deviance and the residual deviance, measures how much better the fitted model is compared to the null model.

- The null deviance is 751.44, which is the deviance of the model with just the intercept (no features).
- The residual deviance is 77.08, which is much lower than the null deviance. This suggests that our model with predictors is a much better fit than the baseline model (with no predictors).
- The deviance goodness of fit if $(751.44 - 77.08) = 674.36$, indicates that the inclusion of your predictors has resulted in a substantial improvement in model fit, as compared to the null model.

6.1.4 McFadden Psuedo R-Square

McFadden's Pseudo R-squared is a statistic used to assess the fit of a logistic regression model. It is similar to the R-squared used in linear regression, but it is specifically designed for models like logistic regression where the dependent variable is binary. Unlike traditional R-squared, which represents the proportion of variance explained by the model, McFadden's Pseudo R-squared does not have a direct interpretation in terms of "variance explained," but it still provides a measure of how well the model fits the data. McFadden's Pseudo R-squared is calculated as:

$$R^2_{MF} = 1 - \frac{\text{log-likelihood of the fitted model}}{\text{log-likelihood of the null model}} \tag{7}$$

McFadden's Pseudo R-squared values typically range from 0 to 1, but they are generally much lower than R-squared values in linear regression. $R^2 = 0$ indicates that the fitted model does not improve on the null model at all (i.e., the predictors do not explain any of the variability in the outcome). R^2 close to 1 suggests a perfect model, where the fitted model explains almost all of the variability in the dependent variable.

- For our model, McFadden's Pseudo R-squared = 0.8974199 which justifies a good fit.

6.1.5 Binned Residual Plot

A Binned Residuals Plot is a diagnostic plot used to evaluate the fit of a logistic regression model (or other predictive models) by visualizing the residuals in relation to the predicted probabilities. It is a way of assessing whether the model's predictions are well-calibrated across different levels of predicted probabilities. The plot helps identify if the model is systematically underestimating or overestimating the likelihood of certain outcomes. The Binned Residuals Plot shows the mean residuals (the difference between the observed and predicted outcomes) for different bins of predicted probabilities. It plots these residuals against the predicted probability values, with each bin representing a range of predicted probabilities. This is particularly useful in logistic regression, where the outcome is binary and the predicted values are probabilities between 0 and 1. In a well-calibrated model, the residuals should be close to zero across all bins of predicted probabilities. This indicates that the model's predicted probabilities are close to the actual observed outcomes, and there is no systematic over- or under-prediction across any predicted probability ranges. A well-calibrated model should have residuals that are centered around zero across all bins of predicted probability. If the residuals significantly deviate from zero, it suggests that the model may not be properly calibrated, and the predicted probabilities may not match the true probabilities.

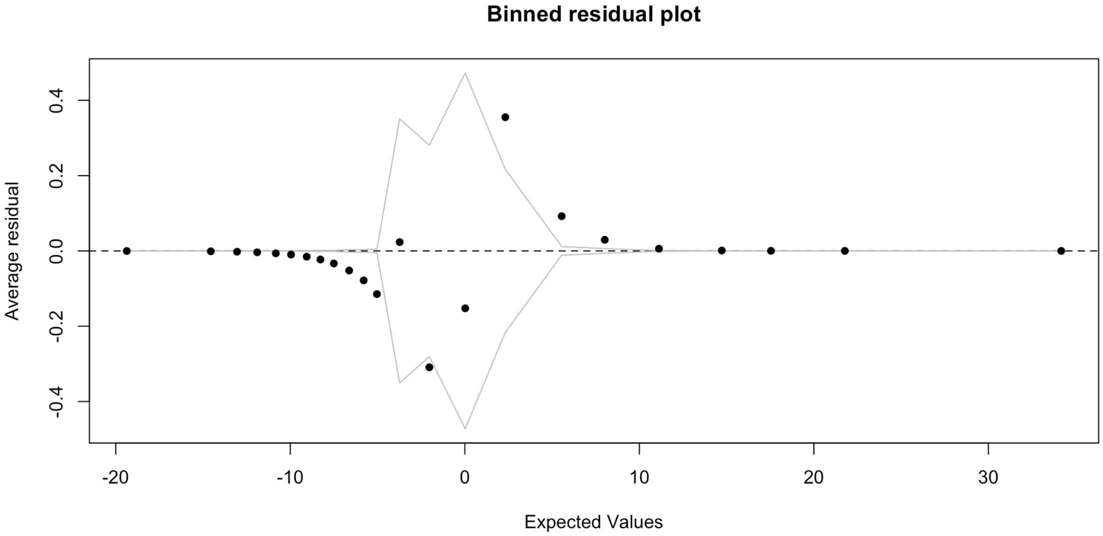


Figure 9: Binned Residual Plot

Since most of the residuals lie within the confidence limits and also most of the residuals are centered around zeo, our model shows a good fit.

6.2 Elastic Net Regularization (Lasso and Ridge Regression)

Elastic Net is a regularization technique used in linear models, particularly when the goal is to improve model generalization, handle multicollinearity, and perform feature selection. It combines the strengths of two other regularization methods: Lasso (Least Absolute Shrinkage and Selection Operator) and Ridge regression.

Elastic Net is particularly useful when the number of predictors (features) is large or when there is multicollinearity (i.e., high correlation) among the predictors. Lasso adds a penalty term proportional to the absolute value of the coefficients. The L1 penalty encourages sparsity in the model, meaning that it drives some coefficients to exactly zero. This can be useful for feature selection because predictors with coefficients of zero are effectively excluded from the model. Ridge adds a penalty term proportional to the square of the coefficients. The L2 penalty shrinks the coefficients toward zero but does not set them exactly to zero, meaning that all predictors remain in the model, though their effects are diminished.

Elastic Net is a regularized regression technique that combines the penalties of Lasso (L1 regularization) and Ridge (L2 regularization), making it particularly useful when dealing with multicollinearity and large feature sets.

Performing this model using “glmnet” method in R we get the following output as shown in below table:

Table 5: Summary of the full Elastic net regularization model

alpha	lambda	ROC	Sens	Spec
0.10	0.0007673665	0.9853296	0.9688880	0.9135023
0.10	0.0076736649	0.9887056	0.9763520	0.9187379
0.10	0.0767366489	0.9894536	0.9866150	0.8920281
0.55	0.0007673665	0.9845430	0.9676428	0.9113971
0.55	0.0076736649	0.9875209	0.9757290	0.9140176
0.55	0.0767366489	0.9874432	0.9887947	0.8632075
1.00	0.0007673665	0.9816840	0.9617209	0.9024662
1.00	0.0076736649	0.9844558	0.9704321	0.9082337
1.00	0.0767366489	0.9835438	0.9869285	0.8364618

This table presents the results of a logistic regression model, using different values for alpha and lambda in an Elastic Net regularization framework. The output provides performance metrics, including:

- ROC (Receiver Operating Characteristic curve): Measures the model’s ability to distinguish between classes.
- Sens (Sensitivity): Also known as Recall or True Positive Rate, indicates the proportion of actual positives (malignant cases) correctly identified by the model.
- Spec (Specificity): Indicates the proportion of actual negatives (benign cases) correctly identified by the model.

The alpha parameter controls the mix between Lasso (L1) and Ridge (L2) regularization in Elastic Net. Alpha = 0.10: This means the model is more influenced by Lasso regularization (closer to Lasso). Alpha = 0.55: This is a middle-ground between Lasso and Ridge, meaning the model uses both L1 and L2 penalties. Alpha = 1.00: The model uses Ridge regularization (L2 penalty) exclusively.

The lambda parameter controls the strength of the regularization. Larger lambda values result in stronger regularization (shrinking coefficients towards zero). As lambda increases, the model becomes simpler, reducing the potential for overfitting.

The Elastic Net regularization with alpha = 0.55 (a balanced L1 and L2 penalty) and moderate values of lambda seems to be the most optimal configuration in terms of balancing ROC, sensitivity, and specificity. Specifically, lambda values around 0.0077 give the best performance overall, especially in distinguishing malignant (positive) and benign (negative) cases giving ROC value of 0.9875209.

6.2.1 Receiver operating characteristics (ROC)

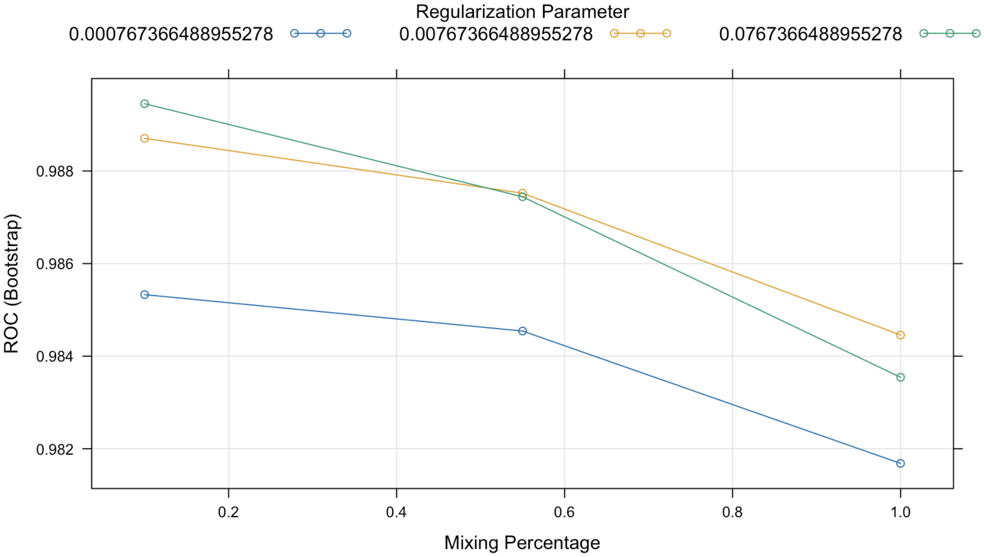


Figure 10: ROC curve of Elastic Net Regularization

From the above ROC plot, it is justified that we get the best model fit for $\alpha = 0.55$ and $\lambda = 0.00767$. For this we have ROC = 0.9875209 with Sensitivity = 0.9757290 and Specificity = 0.9140176. Below plot shows different value of AUC for different values of lambda.

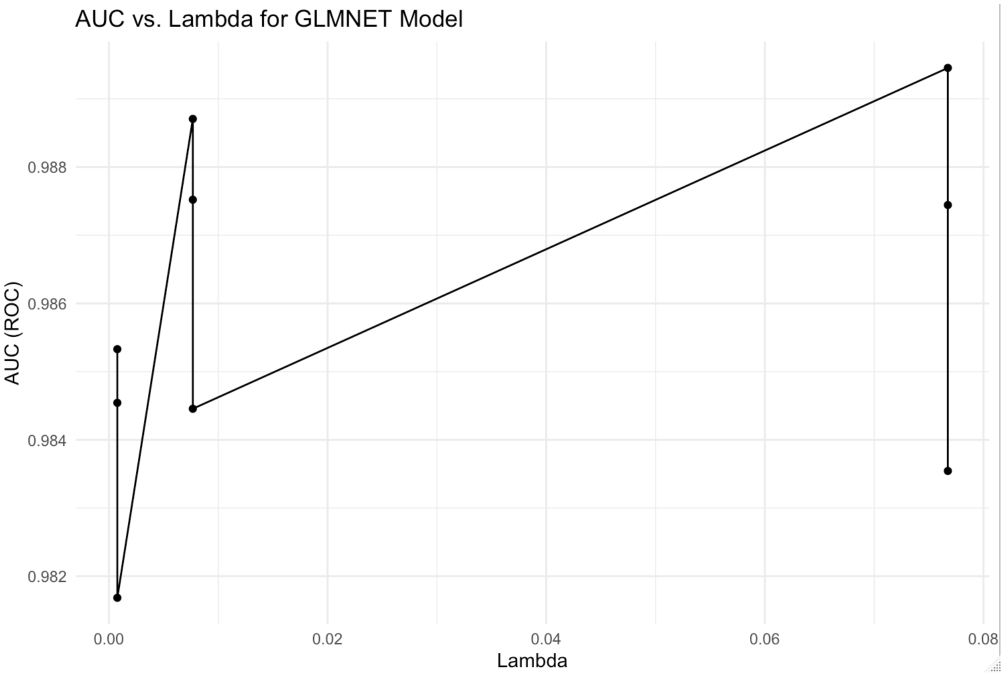


Figure 11: AUC for different lambda

6.2.2 Importance of variables

Below plot shows the ten features which are most important for glmnet model. This shows fractal_dimension_se is the most important feature.

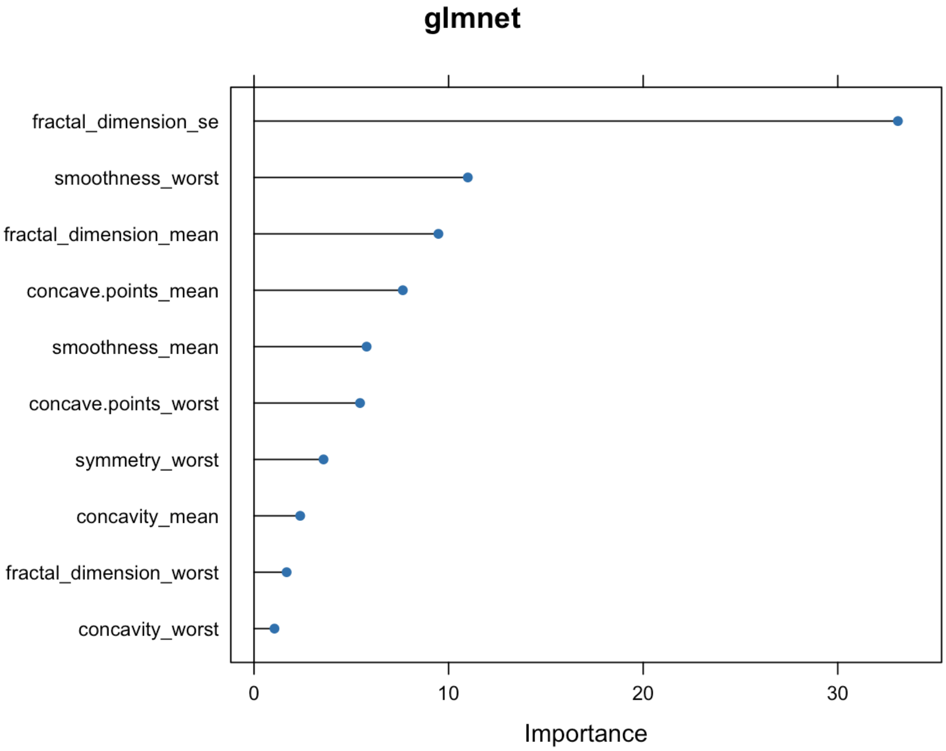


Figure 12: Importance of variables

7 Conclusion

In this study, we developed and evaluated two type of predictive models for classifying breast cancer as either malignant or benign: logistic regression model and an elastic net regularization model.

- Logistic Regression Model: The final logistic regression model achieved an accuracy of 0.9719, indicating that the model correctly classified 97.19% of the samples, which reflects a strong overall performance. For this model we get AUC = 0.960 with precision and recall of 0.9623.
- Elastic net regularization Model: The elastic net model with $\alpha = 0.55$ and $\lambda = 0.00767$ performed exceptionally well, with an ROC score of 0.9875, indicating that the model has an excellent ability to distinguish between the two classes (malignant and benign). The Sensitivity of 0.9757 indicates the high detection rates for malignant tumors.

The developed logistic regression model and elastic net regularization model achieved a satisfactory performance in predicting breast cancer (malignant or benign), as evidenced by the accuracy and sensitivity of the model.

Both models demonstrate high classification accuracy and robustness, with the logistic regression model excelling in terms of precision and recall, while the elastic net regularization model shows superior performance in terms of ROC and sensitivity. These results indicate that both models are reliable tools for predicting the malignancy of breast tumors and can be used to assist in clinical decision-making. However, depending on the specific application or the trade-off between sensitivity and specificity, either model could be preferred in practice.

In conclusion, the combination of logistic regression and elastic net regularization provides complementary strengths in breast cancer classification, offering valuable insights into potential treatment pathways and aiding in early diagnosis.

8 References

[1] Street, W.N., Wolberg, W.H., & Mangasarian, O.L. (1993). Nuclear feature extraction for breast tumor diagnosis. Electronic imaging.

[2] Wolberg W, Mangasarian O, Street N, Street W. Breast Cancer Wisconsin (Diagnostic) [dataset]. 1993. UCI Machine Learning Repository. Available from: https://doi.org/10.24432/ C5DW2B.3