Prediction of cancer using Machine Learning algorithm

1 Introduction

This coursework is to train the Logistic Regression model with the given cancer data set and predict the probability of having cancer. The model learns from the data set using positive samples and negative samples and then predicts the outcome of the new exposed data. Logistic Regression is one of the technique pulled from Statistics under Machine Learning to build a binary classification model.

2 Background

The below concepts are the logistic regression technique's roots recursively addressed before discussing the model in detail.

2.1 Regression Analysis

In statistic modelling, regression analysis is a set of statistical process of estimating the relationship between a dependent variable and one or more independent variables. Mostly, the dependent variable are called 'outcome' or 'response' whereas the independent variable are called the 'predictors' , 'features' or 'explanatory variables'. Regression Analysis is primarily used used for prediction and forecasting. Secondly, it is used to infer causal relationships between the independent and dependent variables. More importantly, regression analysis themselves only reveal relationship between a dependent variable and a collection of independent variables in a fixed dataset. These relationship are represented in the form of equation.

2.2 Binary Regression

A binary regression estimates a relationship between one or more predictor variables and a single binary response variable, a variable with two possible outcomes. It models the probability for the two possible outcomes of the binary response variable with mutually exclusive values. Binary regression is a case of binomial regression with a singe outcome, n=1, where one of the two outcomes is regarded as "success" and labeled as "one". The value is the count of successes in a trial which is either 0 or 1.

2.3 Binary Regression Models

Binary regression models can be classified as

1. Latent variable interpretation together with a measurement model

   In statistics, latent variables are variables that are not directly observed($\hat{y}$) but are rather inferred from other variables that are observed(x). Mathematical models that aim to explain observed variables in terms of latent variables are called latent variable models.

   The logistic regression is finding the β parameters that best fit

   $$y = \begin{cases} 1 & \beta_0 + \beta_1 + e > 0 \\ 0 & else \end{cases}$$

   The latent variable model in binary regression can be written as
   $\hat{y} = \beta_0 + \beta_1 + e$
   if $\hat{y} >= 0, y = 1$
   if $\hat{y} < 0, y = 0$

2. Probabilistic model (Directly modeling the probability) The simplest direct probabilistic model is the logit model, which models the log-odds as a linear function of the explanatory variable or variable. The logit model is "simplest" in the sense of generalized linear models (GLIM): the log-odds are the natural parameter for the exponential family of the Bernoulli distribution, and thus it is the simplest to use for computations.

2.4 Applications

Binary regression is principally applied either for prediction (binary classification), or for estimating the association between the explanatory variables and the output.[1] We will be focusing on building binary classifier using the probabilistic model type.

2.5 Logistic regression model representation

Technically, regression is used to describe the given data in terms of variables that form the relationships whereas the regression model describes the relationship using a mathematical model. This model is then used to predict "new values" of the predictor not observed in the original or given dataset. Most regression models propose , y) as a function of x and β with with e,

$$\hat{y} = f(x, \beta) + e, \tag{1}$$

where β are unknown parameters, x are the independent variables, y are the dependent variables and e is the indirectly observed error in data.

When the response variable is a binary, i.e, it can take only two forms like yes/no or present/absent, etc. ,logistic regression is effectuated. The resulted predicted response values, y*I, is a probability of both of the alternatives or likelihood of getting a positive result for a particular value of a predictor variable in the case of logistic regression. Logistic regression is estimating the parameters of a logistic model.

Logistic model is used for modelling categorical dependent variables like yes/no or true/false. Categorical dependent variables could be of a certain class or category or event. The regression equation holding the relationship between qualitative response Y* and predictor variable X1, x2, x3.. with coefficients and an additional coefficient that provides the intercept or bias can be represented as

$$\hat{y} = \beta_0 + \beta_1 x_1 + ...\beta_n x_p \tag{2}$$

This equation help to estimate the new target variable Y*I, fitting a line through the data or a hyperplane, as per the dimensions of the given data.[2] In linear algebra, with a vector for the coefficients (Beta) and a matrix for the input data (X) and a vector for the output (y).

$$Y = X * \beta$$

In logistic regression, the predicted $\hat{y}$ values(weighted sum) using the equation are squashed and transformed to values between 0 and 1 using logistic function. Thus, the equation is a linear combination of coefficients or weights for every predictor variables(observed data) for any regression model. Linear regression fits the line to the data, which can be used to predict a new quantity, whereas logistic regression fits a line to best separate the two classes.

2.6 Parameters of the model

The coefficients $\beta_0, \beta_1 ... \beta_n$ are unknown and are estimated using the maximum likelihood or Least square Estimation to predict the Y from the given sample of observations. Unlike normal regression, the β parameters cannot be directly expressed by equation represented using x and y values in the observed data. They are found by an iterative search process by software programs that finds the maximum of "likelihood expression" that is function of all observed y and x values.

Maximum Likelihood Estimation is a frequentist probabilistic framework that finds a set of parameters for the model that maximizes a likelihood function.

The estimation of the coefficients must be approximately accurate for the predictions to be closer to the original observed values to fit the model.

Unlike linear regression, there is no formula for the estimates of for logistic regression. Finding the best estimates requires repeatedly improving approximate estimates until stability is reached. This is done easily on a computer, and there are many statistical software packages that perform logistic regression, but it makes logistic regression less understandable and more of a “black box” approach for many researchers.

3 Mathematical model

3.1 Odds ratio

Odds ratio is a measure of the likelihood of a particular outcome. It is the ratio of the number of the events occurrence to the number of the events that do not occur. Lets say p is the probability of the occurrence of event =1 then

$$odds\,ratio = \frac{p}{1-p}$$

3.2 Logit

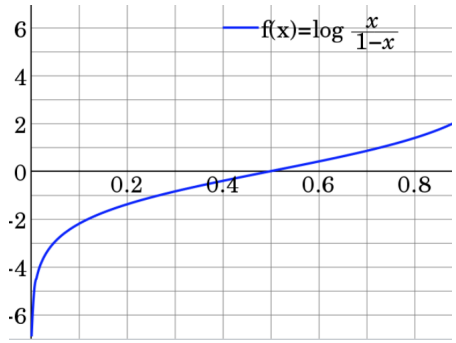

FIGURE 1: Plot of logit(p) in the domain of (0,1), where the base of the logarithm is e

In statistics, the logit is the quantile function associated with the standard logistic distribution. Mathematically, the logit is the inverse of the standard logistic function

$$\sigma(x) = \frac{1}{(1+e^{-x})}$$

Therefore, logit can be defined as

$$\text{logit}(p) = \sigma^{-1}(p) = \ln\left(\frac{p}{1-p}\right) \quad \text{for} \quad p \in (0,1)$$

Hence, logit is also called as log-odds for its equivalence to the logarithm of the odds ratio, $\frac{p}{1-p}$ where p is a probability. It is a fucntion that maps probability values from (0, 1) to real numbers in (-∞, ∞)

3.3 Logistic function

We can understand the core of logistic regressionby knowing the standard logistic function. The logistic function is a sigmoid function, which takes any real input and outputs a value

between 0 and 1. Logit takes input log-odds and gives output probability. A logistic function can be defined as,

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}},$$

where L, curve's maximum value k, logistic growth rate x_0, sigmoid's midpoint

The standard logistic function where $L = 1, k = 1, x_0 = 0$,

$$f(x) = \frac{1}{1 + e^{-x}}, \tag{3}$$

where X is the input to the function

A graph of the logistic function on the t-interval (6,6) is shown in Figure 2 .

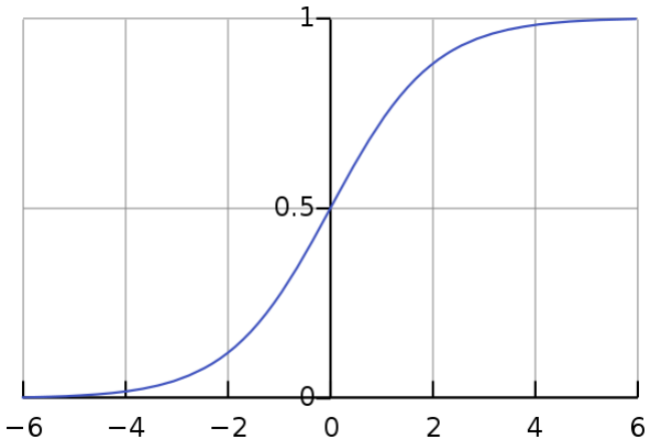

FIGURE 2: Plot of logit(p) in the domain of (0,1), where the base of the logarithm is e

3.4 Derivation of the log-odds of event occurrence

Let us derive that the linear weighted sum of the input predictors calculates the log-odds of event occurrence.

From Standard logistic function ,

$$f(x) = \frac{1}{1 + e^{-x}}$$

Lets consider, f(x) = y (linear combination of coefficients and term, x)

$$\therefore y = \frac{1}{1 + e^{-x}}$$

$$\therefore \frac{1}{y} = 1 + e^{-x}$$

$$\therefore e^{-x} = \frac{1}{y} - 1$$

$$\therefore e^{-x} = \frac{1-y}{1}$$

$$\therefore \frac{1}{e^{x}} = \frac{1-y}{1}$$

Taking inverse,

$$\therefore e^{x} = \frac{y}{1-y}$$

$$\therefore x = \ln\left(\frac{y}{1-y}\right)$$

From equation 2. the linear equation representing the target y is

$$y = \beta_0 + \beta_1 x_1 + ...\beta_n x_p$$

,

$$\therefore \beta_0 + \beta_1 x_1 + ...\beta_n x_p = \ln\left(\frac{y}{1-y}\right)$$

$$odds\, ratio = \frac{y}{1-y}$$

Thus, we have proved that the logistic regression model considers the natural logarithm of the odds as a regression function of the predictors.
For a predictor x, it is in the form

$$\ln(odds(Y=1)) = \beta_0 + \beta_1 x_1 \tag{4}$$

where ln is the natural logarithm,
Y is the outcome and Y=1 is when the event occur and Y=0 means when it did not occur,
β_0 is the intercept term or slope and
β_1 represents the regression coefficient.

β_0 and β_1 represents the regression coefficients to induce change in the logarithm of the odds of the event with a 1-unit change in the predictor, x. The logarithm of the odd ratio which represents the ratio of probability of events occurrence to not occurrence is equal to the difference in the logarithm probability of events occurrence to not occurrence

4 Implementation

4.1 Building the model

The base of the logistic regression model is the odds of a two level outcome of interest. One of the outcome level is assumed as the event of interest and is called an event. The odds of the event can calculated as the ratio of the probability of the event will occur to the probability of event will not occur. The two level outcome of interest or classes in the problem will be labeled as 0 and 1

1. The linear equation is used to predict the target, $\hat{y}$
2. The parameters of the model, β parameters are estimated from the sample of observations using the technique, Maximum Likelihood Estimation
3. Then, the logistic function is used to transform the target $\hat{y}$*(calculated weighted sum) values between 0 and 1 using logistic function.

   Replacing the X by the weighted sum in logistic equation (1)

   $$\hat{y} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_p)}}$$

   The output interpreted is the probability from a binomial probability distribution function for the class labeled 1. Since, the interpreted output is either 0 or can be consigned as probability of belonging to the class labeled 1.
4. The calculated probabilities ranging between 0 and 1 can be classified into 0 and 1 by classifying the probabilities. Eg: if p(class=1) < 0.05 then classify as 0 and if p(class=1) > 0.05 then classify as 1, where p(class=1) is the binomial probability distribution function for the class labeled 1.

The Logistic Regression module of the Scikit-learn python library is built using the above mentioned approach and the designated binary classifier to predict cancer for the given unobserved data is built using this library. The pipeline of building the model is depicted in Figure 3.

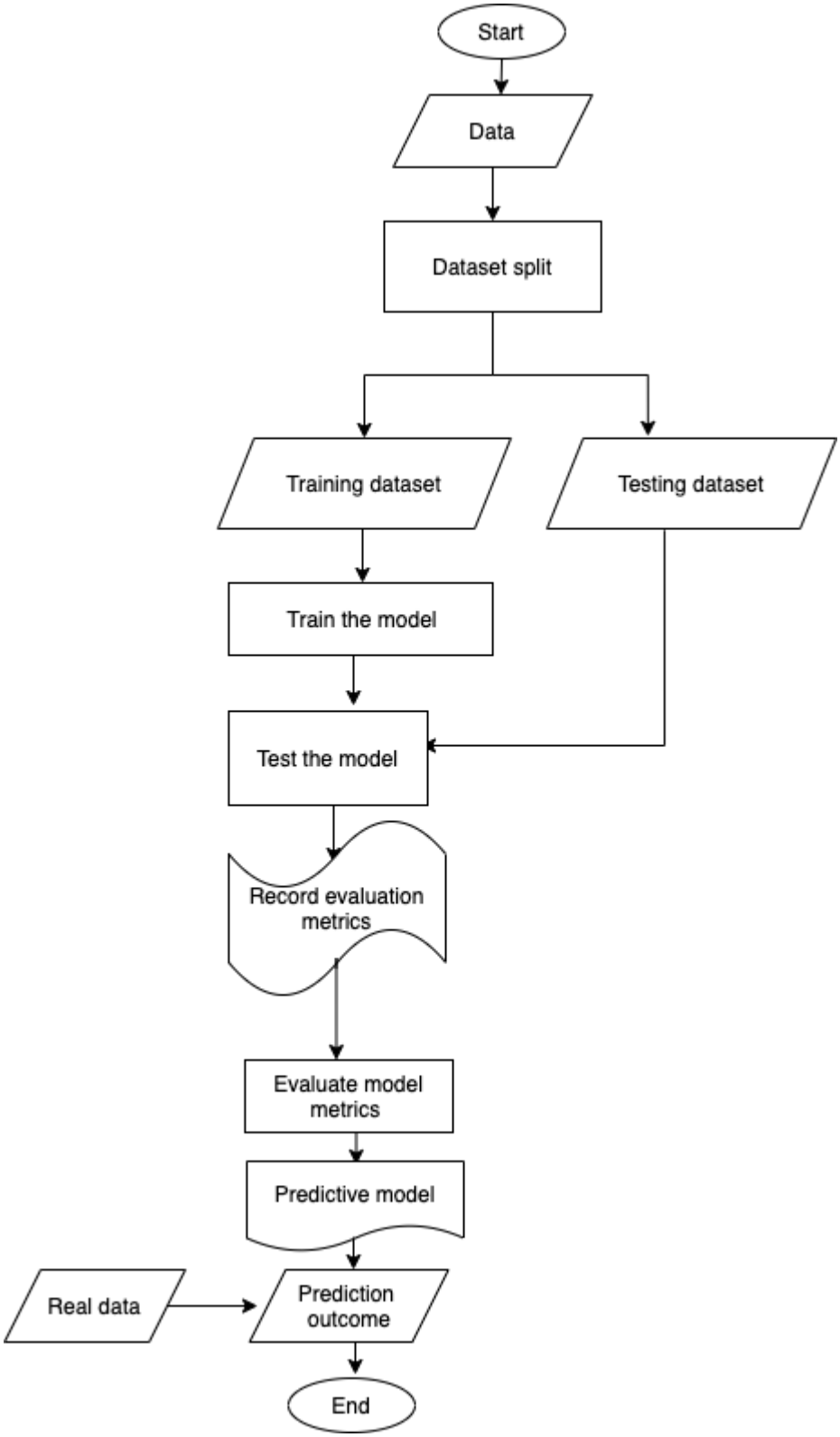

FIGURE 3: Pipeline of building the binary classifier model

4.2 Dataset

The given cancer data set is an imbalanced one. We can in the below Figure 4, that the data for the cancer type, malignant is almost half of the benign data set. Therefore, malignant dataset is consider to be a minority-class over the benign dataset. For a binary classifier, the minority samples are considered as positive whereas the majority samples as negative. The reason behind this assumption is the evaluation model performance metrics uses positive instances. Hence, this assumption will help evaluate the model against the minority class. Nevertheless, the performance of majority are better than minority.

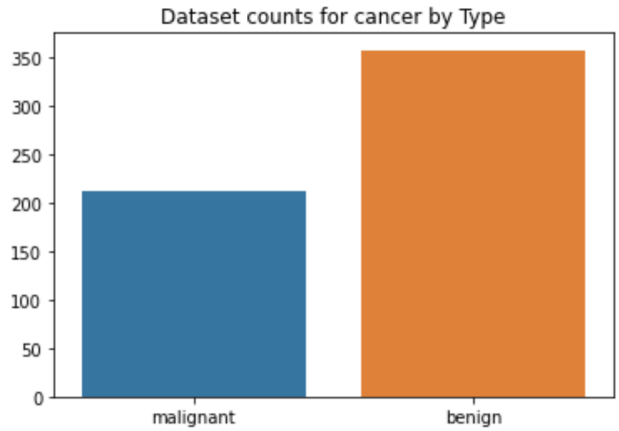

FIGURE 4: Feature extraction using the CNN model to train and test

4.3 Model Evaluation Indicators

The designated binary classifier will either classify the given observation as positive class or negative class. The given data set is an imbalanced data set with benign in majority over malignant class data set.

For binary classification problems, the confusion matrix(Table 1) with the discriminating evaluation metrics helps to evaluate the performance of the classifier[3]. The rows of the confusion matrix represent the actual class and the columns represent the predicted class. The conventional accuracy and error rate metrics are insufficient to evaluate correct and incorrect predictions over total instances due to imbalanced data set.

The model evaluation metrics are built on positive class as mentioned in Table 2. Hence, to evaluate the model the minority-class is assumed as positive. Therefore, the minority-class, malignant, will be assumed as positive.

The binary classifier predict the outcomes as either positive or negative leading to possibilities of four type of outcomes: True positives (TP), True Negatives (TN), False Positive (FP) and False Negative (FN). Focusing on the positive samples, the appropriate performance measures are Precision and Recall. Precision is to measure the accuracy of predicting the positive class whereas Recall is the measure of number of total positives identified accurately[4].

TABLE 1: *Confusion Matrix for binary classification [3]*

	Predicted Negative class	**Predicted Positive class**
Actual Negative Class	True Negative (TN)	False Positive (FP)
Actual Positive Class	False Negative (FN)	True Positive (TP)

TABLE 2: *Model evaluation terminologies and indicators[4]*

Terminology	**Description**	**Formula**
True positives (TP)	Number of positive instances that are classified as positive.	NA
True Negatives (TN)	Number of negative instances that are classified as negative.	NA
False Positive (FP)	Number of negative instances that are classified as positive.	NA
False Negative (FN)	Number of positive instances that are classified as negative.	NA
Accuracy	The ratio of correct predictions over the total number of instances evaluated.	(TP + TN) / (TP + FP + TN + FN)
Error rate	The ratio of incorrect predictions over the total number of instances evaluated.	(FP + FN) / (TP + FP + TN + FN)
Recall	NumberThe proportion of the positive samples correctly identified to the number of the positive samples.	TP / (TP + FN)
Precision	The proportion of the positive samples correctly identified to the number of all classified positive samples.	TP / (TP + FP)
F1 score	Harmonic average value of precision and recall.	(2 * Precision * Recall) / (Precision + Recall)
Sensitivity (True positive rate: TPR)	The proportion of positive instances that are correctly classified as positive.	100.0 * (TP / (TP + FN)
Specificity (True Negative rate: TNR)	The proportion of negative instances that are correctly classified as negative.	100.0 * (TN / (FP + TN))

5 Analysis of the results

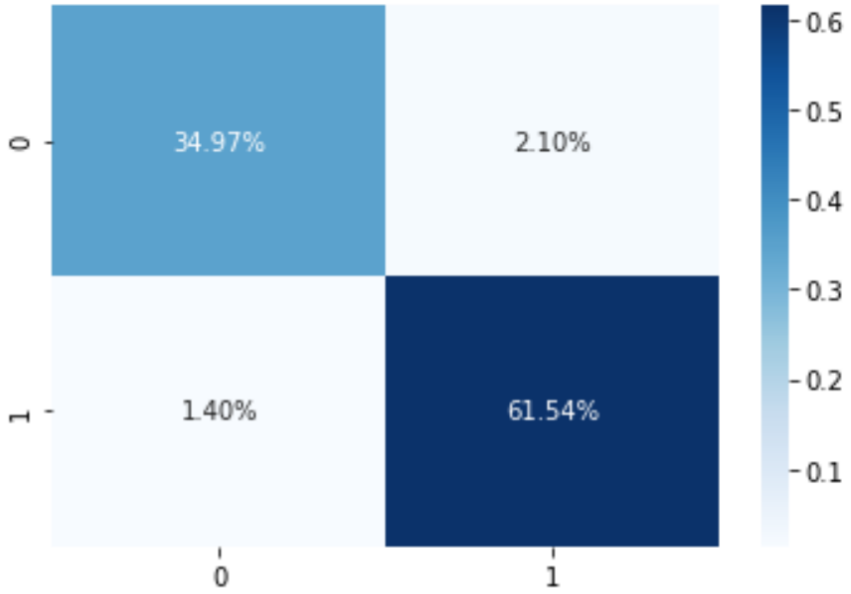

FIGURE 5: 0-positive(minority) class, 1-negative(majority) class; left-bottom represents false negative whereas right-bottom represents true negatives; left-top represent true positives whereas right-top represents false positives

Below are the observations from the confusion matrix(Figure 3):

1. In the above Figure 5, false negatives and false positives are low indicating that the performance for the model is with better accuracy.
2. False positives is more than false negative i.e False predicting as benign is more that false predicting malignant. Falsely predicting as benign may mislead the patient from early treatment which is bad for the patient whereas falsely predicting as malignant may put patient in dangerous situation of going through harmful treatments unnecessarily.

TABLE 3: *Evaluation metrics for the model*

Terminology	**Description**	**%**
Accuracy	(TP + TN) / (TP + FP + TN + FN)	96.50
Error rate	(FP + FN) / (TP + FP + TN + FN)	0.03
Recall	TP / (TP + FN)	97.78
Precision	TP / (TP + FP)	96.70

The metrics mentioned in the Table 3 says that mode has better accuracy accompanied with low error rate and good precision without compromising the sensitivity, recall.

5.1 Summary

To summarize, a binary classifier model was built using logistic regression mathematical model. This model was trained using the given cancer data set. The trained model was then used to predict whether the new unobserved observation is benign or malignant. The model's performance was evaluated using discriminating confusion matrix metrics.

Bibliography

[1]J. I.E.Hoffman, "Basic Biostatistics for Medical and Biomedical Practitioners", Doi.org, 2019. [Online]. Available: https://doi.org/10.1016/B978-0-12-817084-7.00033-4. [Accessed: 02- Oct- 2021].

[2]A. Worster, J. Fan and A. Ismaila, "Understanding linear and logistic regression analyses", CJEM, vol. 9, no. 02, pp. 111-113, 2007. Available: 10.1017/s1481803500014883.

[3]S. Ruuska, W. Hämäläinen, S. Kajava, M. Mughal, P. Matilainen and J. Mononen, "Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle", 2021. [Online].
Available: https://doi.org/10.1016/j.beproc.2018.01.004. [Accessed: 02- Oct- 2021].

[4]"A critical investigation of recall and precision as measures of retrieval system performance | ACM Transactions on Information Systems", Doi.org, 2021. [Online]. Available: https://doi.org/10.1145/65943.65945. [Accessed: 02- Oct- 2021].