zerocode learning

Learning Data Analytics Made Easy

USER GUIIDE

BINARY LOGIT ANALYSIS

TABLE OF CONTENTS

INDEX

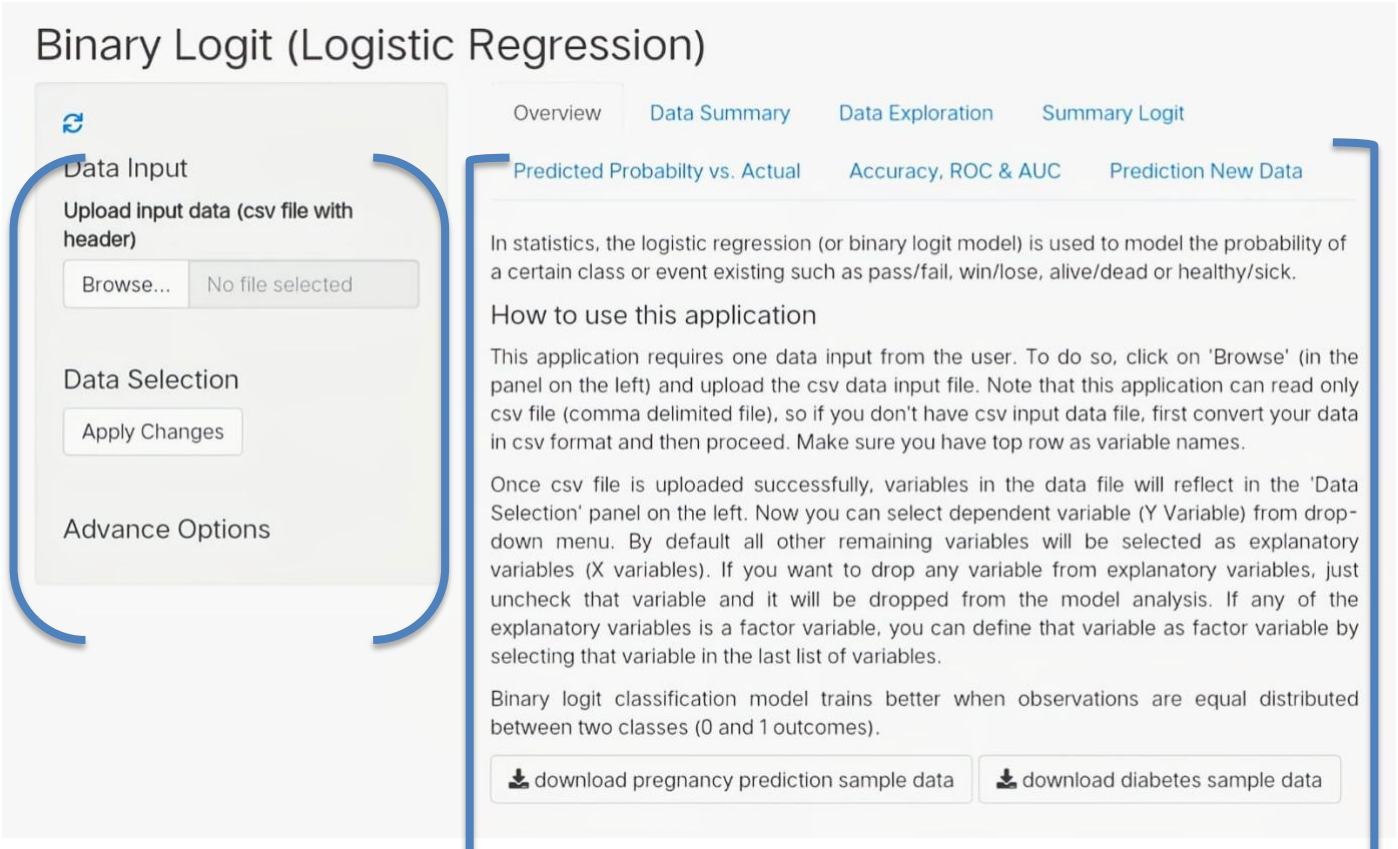
BINARY LOGIT REGRESSION ANALYSIS

01

Binary logit Regression, Binary logistic regression (LR) is a regression model where the target variable is binary, that is, it can take only two values, 0 or 1.Logistic regression is easier to implement, interpret, and very efficient to train. If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting. It makes no assumptions about distributions of classes in feature space. It is named 'Logistic Regression' because its underlying technique is quite the same as Linear Regression. The term “Logistic” is taken from the Logit function that is used in this method of classification.

LEFT PANEL (INPUT AREA)

OPERATIONAL ANALYSIS TAB (MAIN PANEL)



LEFT PANEL (INP)

Upload your dataset here

Select your favorable variables required to base the analysis

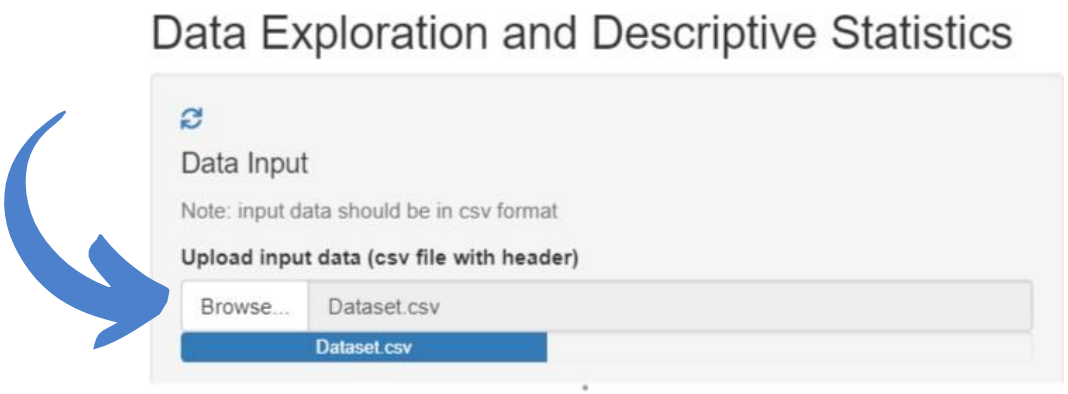
Apply any changes if you want to do.

Select the subsamples or the whole data for testing.

Deal with missing values either drop or immute it.

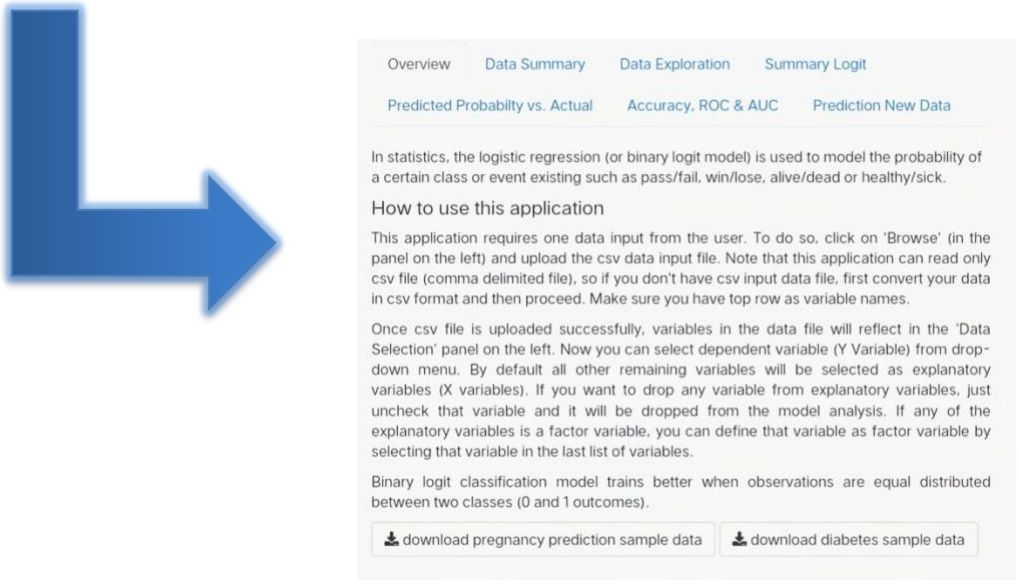
DATA INPUT (UPLOADING DATASET)

- Click on browse
- Select the datafile that is in the form of csv format.(Ex program.csv)
- Browse the file and select the data to train your model for prediction.
- Top rows of the dataset should be of ‘variable names’.



OVERVIEW TAB

This tab provides you with relevant study resources, tutorials, sample datasets and a short overview to start with, which helps you understand and comprehend your data correctly. This tab also provides you the basic idea about regression analysis , gives sample data and provides the description about regression.



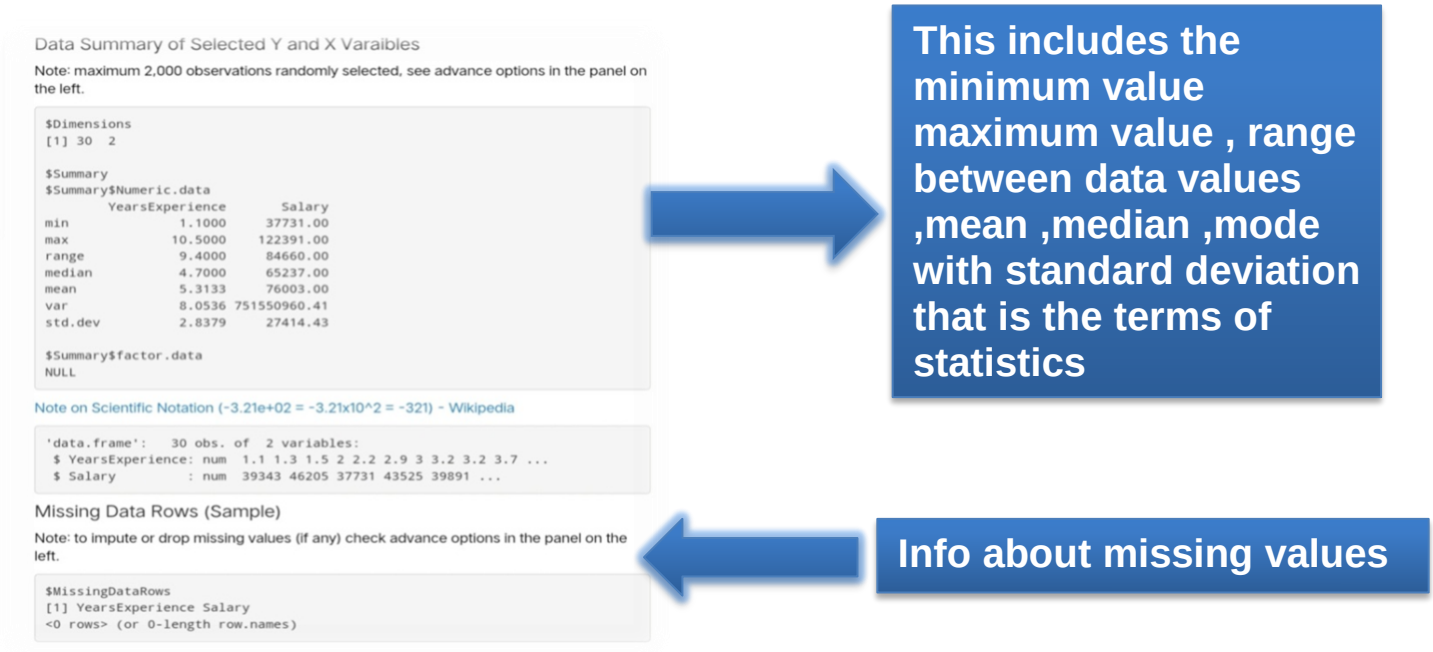
DATA SUMMARY TAB

04

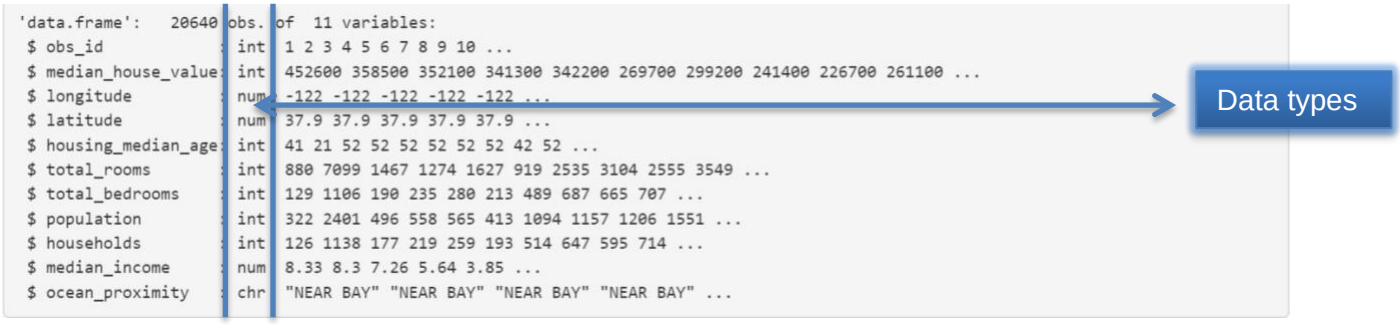
It is very important to understand our data completely to infer meaningful insights and to get an overview of all the data points as a whole, but it is quite impossible to analyze thousand data points manually.

The **'Data Summary'** option enables you to get a comprehensive evaluation through statistical measures that help us form the basis of our analysis.

It will display all the 'descriptive analytics' measures including mean, median, standard deviation, variance etc. for all the data variables present in the dataset. we can review the uploaded data and the contents of it, A brief summary of the data can be seen it includes range of data values, minimum and maximum value missing and null values etc.



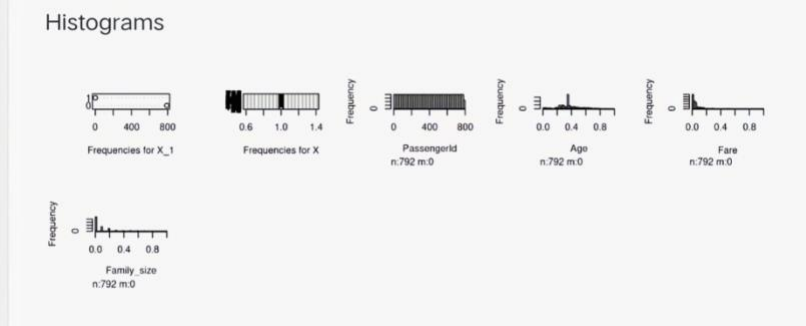
It also segregates dataset variables into respective data types, such as integer, whole numbers, character etc.

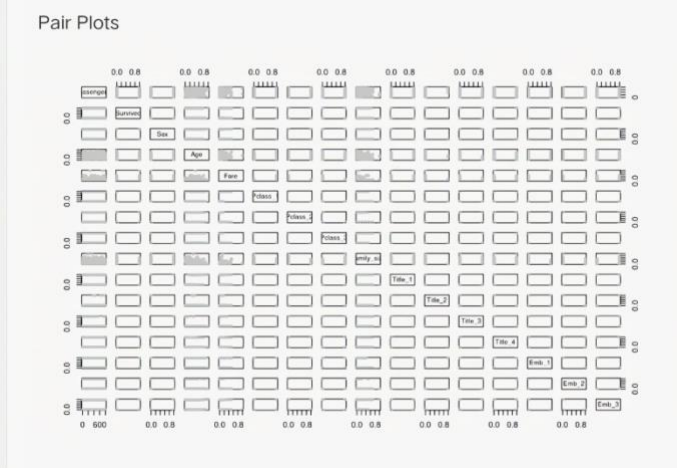


Use the left panel to transform selected variables as per the requirement of analysis ,correspondingly the data summary will also change.

DATA EXPLORATION TAB

Data Exploration is the representation of data through use of common graphics, such as charts, plots, infographics, and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.





1)HISTOGRAM: A histogram is a graph that shows the frequency of numerical data using rectangles.

The height of a rectangle (the vertical axis) represents the distribution frequency of a variable (the amount, or how often that variable appears). Histograms give a rough sense of the density of the underlying distribution of the data, and often for density estimation,estimating the probability density function of the underlying variable. The total area of a histogram used for probability density is always normalized to 1. If the length of the intervals on the *x*-axis are all 1, then a histogram is identical to a relative frequency plot.

PAIR PLOTS : A pairplot plot a pairwise relationships in a dataset. The pairplot function creates a grid of Axes such that each variable in data will by shared in the y-axis across a single row and in the x-axis across a single column. That creates plots as shown above Pair plot is used to understand the best set of features to explain a relationship between two variables or to form the most separated clusters. It also helps to form some simple classification models by drawing some simple lines or make linear separation in our data-set.

2)CORRELATION TABLE : A two-way tabulation of the relations between correlates; row headings are the scores on one variable and column headings are the scores on the second variables and a cell shows how many times the score on that row was associated with the score in that column.

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.

Example: A positive correlation is a relationship between two variables in which both variables move in the same direction. Therefore, when one variable increases as the other variable increases, or one variable decreases while the other decreases. An example of positive correlation would be height and weight.

Use the left panel to modify/deal with the outliers identified here.

06

SUMMARY LOGIT TAB

The summary tab provides the overall analysis result. The residual deviance tells us how well the response variable can be predicted by a model with p predictor variables. The lower the value, the better the model is able to predict the value of the response variable. Pearson residuals are defined as the standardized distances between the observed and expected responses, and deviance residuals are defined as the signed square root of the individual contributions to the model deviance. If we use a generalized linear model (GLM) to model the relationship, deviance is a measure of goodness of fit: the smaller the deviance, the better the fit. An important assumption of logistic regression is that the errors (residuals) of the model are approximately normally distributed. The observed values on the response variable cannot be normally distributed themselves, because Y is binary.

CORRELATION PLOT : The correlation coefficient is the specific measure that quantifies the strength of the linear relationship between two variables in a correlation analysis. The coefficient is what we symbolize with the r in a correlation report. A correlation analysis provides information on the strength and direction of the linear relationship between two variables, while a simple linear regression analysis estimates parameters in a linear equation that can be used to predict values of one variable based on the other.

Summary Logistic Regression Model

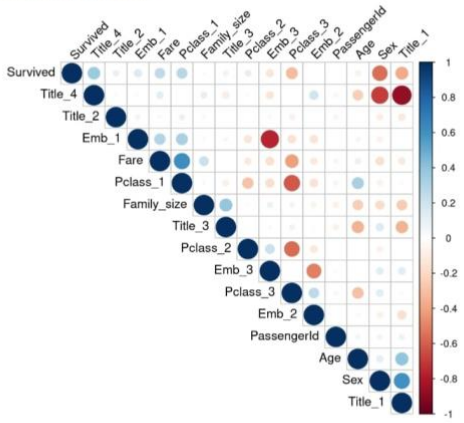
Note on Logistic Regression - Wikipedia

Model predicts log-odds of X = 1

```
Call:
glm(formula = formula, family = binomial, data = mydata())

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.333e-04  -2.100e-08  -2.100e-08  -2.100e-08   2.518e-04

Coefficients: (2 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.625e+01  3.570e+05   0.000    1.000
PassengerId -2.044e-01  3.312e+01  -0.006    0.995
Survived     3.816e+00  1.865e+04   0.000    1.000
Sex         -5.213e+01  1.916e+04  -0.003    0.998
Age         -1.505e+02  2.650e+04  -0.006    0.995
Fare        -3.357e+02  6.954e+04  -0.005    0.996
Pclass_1     7.996e+01  1.700e+04   0.005    0.996
Pclass_2    -2.784e+01  1.949e+04  -0.001    0.999
Pclass_3            NA         NA      NA       NA
Family_size  1.448e+02  6.960e+04   0.002    0.998
Title_1      4.369e+01  4.170e+04   0.001    0.999
Title_2      6.373e+01  1.499e+05   0.000    1.000
Title_3     -2.749e+01  6.160e+04   0.000    1.000
Title_4             NA         NA      NA       NA
Emb_1       -3.231e+01  3.593e+05   0.000    1.000
Emb_2       -1.410e+01  3.587e+05   0.000    1.000
Emb_3       -8.577e+01  3.596e+05   0.000    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.5348e+01  on 791  degrees of freedom
Residual deviance: 1.8442e-07  on 777  degrees of freedom
AIC: 30

Number of Fisher Scoring iterations: 25
```

Correlation Plot

Note on Correlation (Wikipedia)



Estimated confidence interval of beta (X) coefficient under normal distribution

This tab evaluates the correlation coefficients between variables and represents them through a correlation map as shown, where each cell depicts acorrelation between many variables. The size and colorof the circles in each cell depict the degree of correlation, the larger the size and darker the color shade; the higher is the correlation.

07

PREDICTED PROB VS ACTUAL TAB

Logistic Regression is the statistical fitting of an s-curve logistic or logit function to a dataset in order to calculate the probability of the occurrence of a specific categorical event based on the values of a set of independent variables. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes.
Theoretical probability uses math to predict the outcomes. Just divide the favorable outcomes by the possible outcomes. Experimental probability is based on observing a trial or experiment, counting the favorable outcomes, and dividing it by the total number of times the trial was performed.

Predicted Probability : Logistic regression analysis predicts the odds of an outcome of a categorical variable based on one or more predictor variables. A categorical variable is one that can take on a limited number of values, levels, or categories, such as "valid" or "invalid".

Actual Probability : In statistics, the actual value is the value that is obtained by observation or by measuring the available data. It is also called the observed value. The predicted value is the value of the variable predicted based on the regression analysis.

Use the left panel to impute or drop the missing values identified here

08

ACCURACY,ROC & AUC TAB

Accuracy : accuracy = correct_predictions / total_predictions.
Accuracy is the proportion of correct predictions over total predictions. This is how we can find the accuracy with logistic regression: score = LogisticRegression.score(X_test, y_test). The most basic diagnostic of a logistic regression is predictive accuracy. To understand this we need to look at the prediction-accuracy table (also known as the classification table,

hit-miss table, and confusion matrix).
Not only does binary logistic regression allow you to assess how well your set of variables predicts your categorical dependent variable and determine the “goodness-of-fit” of your model as does regular linear regression, but also it provides a summary of the accuracy of the classification of cases,

Confusion Matrix : A confusion matrix of binary classification is a two by two table formed by counting of the number of the four outcomes of a binary classifier. We usually denote them as TP, FP, TN, and FN instead of “the number of true positives”, and so on. Predicted, Theoretical probability uses math to predict the outcomes. Just divide the favorable outcomes by the possible outcomes. It gives information about errors made by the classifier and the types of errors that are being made. It reflects how a classification model is disorganized and confused while making predictions.

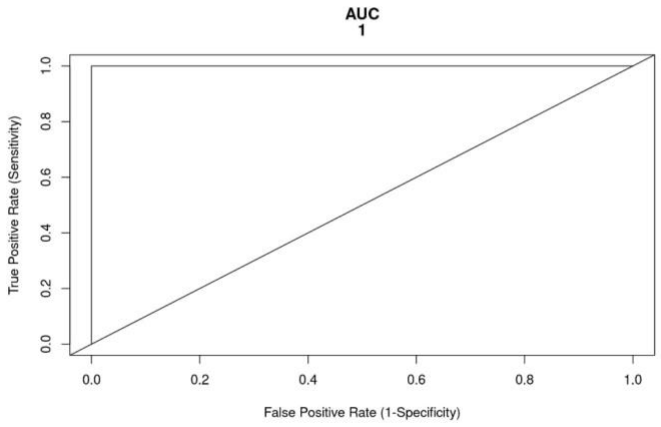
AUC : The Area Under the ROC curve (AUC) is an aggregated metric That evaluates how well a logistic regression model classifies positive and Negative outcomes at all possible cutoffs. It can range from 0.5 to 1, and the larger it is the better.

ROC : ROC curves in logistic regression are used for determining the best cutoff value for predicting whether a new observation is a "failure" (0) or a "success" (1)The ROC curve shows the trade-off between sensitivity (or TPR) and specificity (1 – FPR). Classifiers that give curves closer to the top-left corner indicate a better performance. As a baseline, a random classifier is expected to give points lying along the diagonal (FPR = TPR).

ROC Curve

Note on ROC-AUC (Wikipedia)

ROC curve is created using all cutoff probabilities from 0 to 1. Thus AUC is independent of cutoff, and is often used for model selection (similar to adjusted r-squared in regression)



ROC AND AUC

Confusion Matrix Summary

'Postive' Class = 1 is X = 1

Sensitivity measures how well the model predicts 'Positive' Class = 1, and Specificity measures how well the model predicts 'Positive' Class = 0.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 791   0
         1   0   1

               Accuracy : 1
                 95% CI : (0.9954, 1)
    No Information Rate : 0.9987
    P-Value [Acc > NIR] : 0.3676

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.000000
            Specificity : 1.000000
         Pos Pred Value : 1.000000
         Neg Pred Value : 1.000000
             Prevalence : 0.001263
         Detection Rate : 0.001263
   Detection Prevalence : 0.001263
      Balanced Accuracy : 1.000000

       'Positive' Class : 1
```

ACCURACY