CLASSIFICATION PROBLEMS

- Classification problems are an important category of problems in analytics in which the response variable (Y) takes a discrete value.

- The primary objective is to predict the class of a customer (or class probability) based on the values of explanatory variables or predictors.

Classification Problems

- Classification is an important category of problems in which the decision maker would like to classify the case/entity/customers into two or more groups.

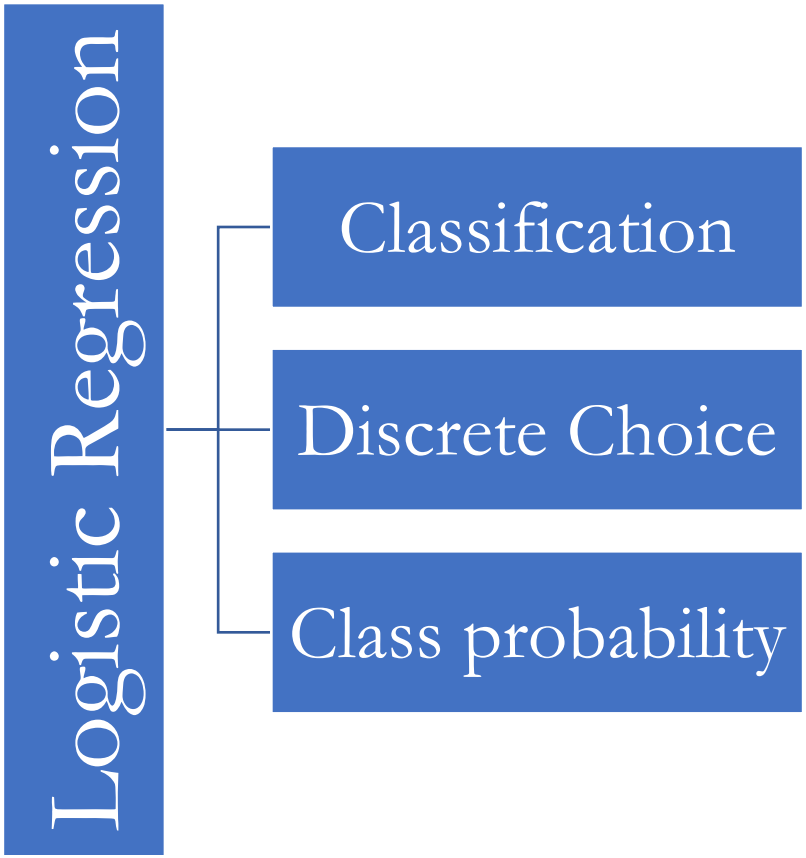
- Examples of Classification Problems:

  - ✔ Customer profiling (customer segmentation)
  - ✔ Customer Churn.
  - ✔ Credit Classification (low, high and medium risk)
  - ✔ Employee attrition.
  - ✔ Fraud (classification of transaction to fraud/no-fraud)
  - ✔ Stress levels
  - ✔ Text Classification (Sentiment Analysis)
  - ✔ Outcome of any binomial and multinomial experiment.

Challenging Classification Problems

- Ransomware
- Anomaly Detection
- Image Classification (Medical Devices, Satellite images)
- Text Classification

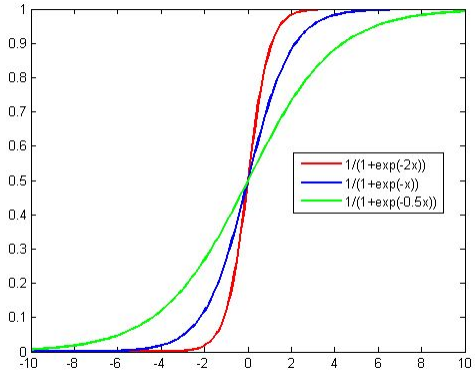
Logistic Regression - Supervised Learning Algorithm


Logistic Regression
Classification
Discrete Choice
Class probability

Logistic Regression - Introduction

- The name logistic regression emerges from logistic distribution function.

$$\frac{e^Z}{1+e^Z}$$

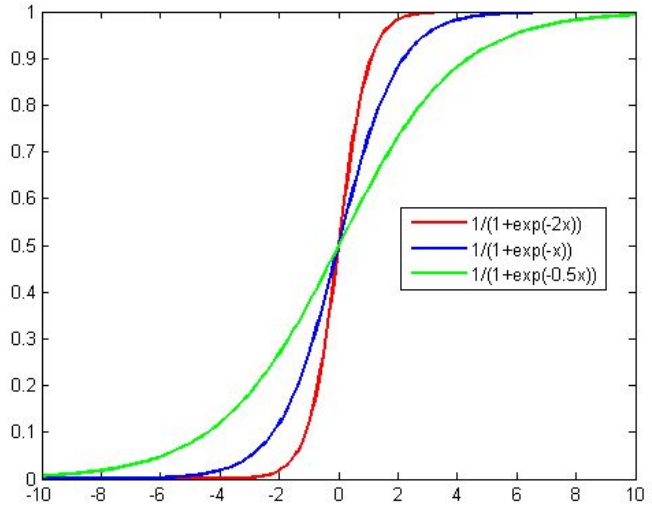


- Mathematically, logistic regression attempts to estimate conditional probability of an event (or class probability).

Logistic Distribution

$$f(t) = \frac{e^Z}{\sigma\left(1+e^Z\right)^2}$$

$$F(Z) = \frac{e^Z}{1+e^Z}$$



It is a symmetrical distribution (density function)

F(Z) is S-shaped curve

Logistic Regression

- Logistic regression models estimate how probability of an event may be affected by one or more explanatory variables.

- Logistic regression is a technique used for predicting "**class probability**", that is the probability that the case belongs to a particular class.

Binomial & Multinomial Logistic Regression

- Binomial (or binary) logistic regression is a model in which the dependent variable is dichotomous.
- In multinomial logistic regression model, the dependent variable can take more than two values.
- The independent variables may be of any type.

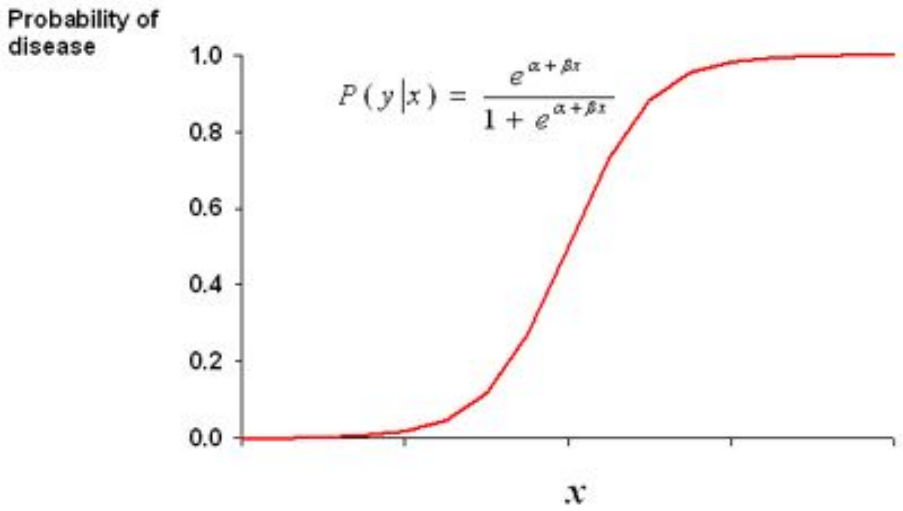
Logistic Function (Sigmoidal function)

$$P(Y=1)=\pi(z)=\frac{1}{1+e^{-z}}=\frac{e^{z}}{1+e^{z}}$$

$$z=\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n$$

Logistic Regression with one Explanatory Variable

$$P(Y=1 \mid X=x) = \pi(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$$



- b = 0 implies that P(Y | x) is same for each value of x
- b > 0 implies that P(Y | x) is increases as the value of x increases
- b < 0 implies that P(Y | x) is decreases as the value of x increases

Generalized Linear Model (GLM)

- Generalization of linear regression model.
- In GLM the error distribution of outcome variable can be other than normal.
- A function of response variable will have a linear relationship with the predictor. The function is called the link function.

Logistic Transformation

- The logistic regression model is given by:

$$\pi_i = \frac{e^{(\beta_0 + \beta_1 X_i)}}{1 + e^{(\beta_0 + \beta_1 X_i)}}$$

$$\frac{\pi_i}{1 - \pi_i} = e^{(\beta_0 + \beta_1 X_i)}$$

$$\ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 X_i$$

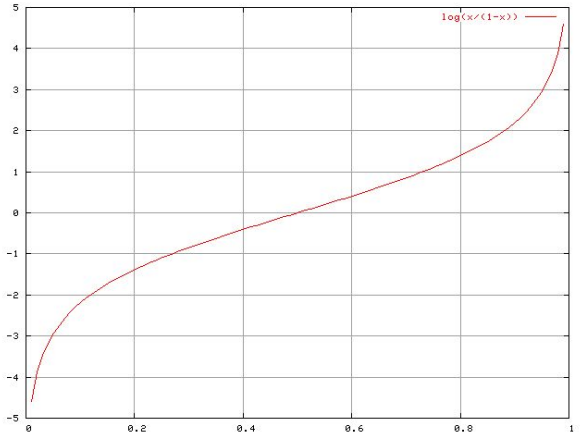
Function with linear properties (Link Function)

Logit Function

- The logit function is the logarithmic transformation of the logistic function. It is defined as the natural logarithm of odds.

- Logit of a variable π (with value between 0 and 1) is given by:

$$Logit(\pi) = \ln(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x$$



Parameter Estimation in Logistic Regression (Maximum Likelihood Estimate)

Likelihood function for Binary Logistic Function

- Probability density function for binary logistic regression is given by:

$$f(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

$$L(\beta) = f(y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

$$\ln(L(\beta)) = \sum_{i=1}^{n} y_i \ln[\pi(x_i)] + \sum_{i=1}^{n} (1-y_i)[\ln(1-\pi_i(x_i))]$$

Likelihood function for Binary Logistic Function

$$\ln\left[L(\beta)\right] = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} \ln(1 + \exp(\beta_0 + \beta_1 x_i))$$

Estimation of LR parameters

$$\frac{\partial \ln(L(\beta_0, \beta_1))}{\partial \beta_0} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} = 0$$

$$\frac{\partial \ln(L(\beta_0, \beta_1))}{\partial \beta_1} = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \frac{x_i \exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} = 0$$

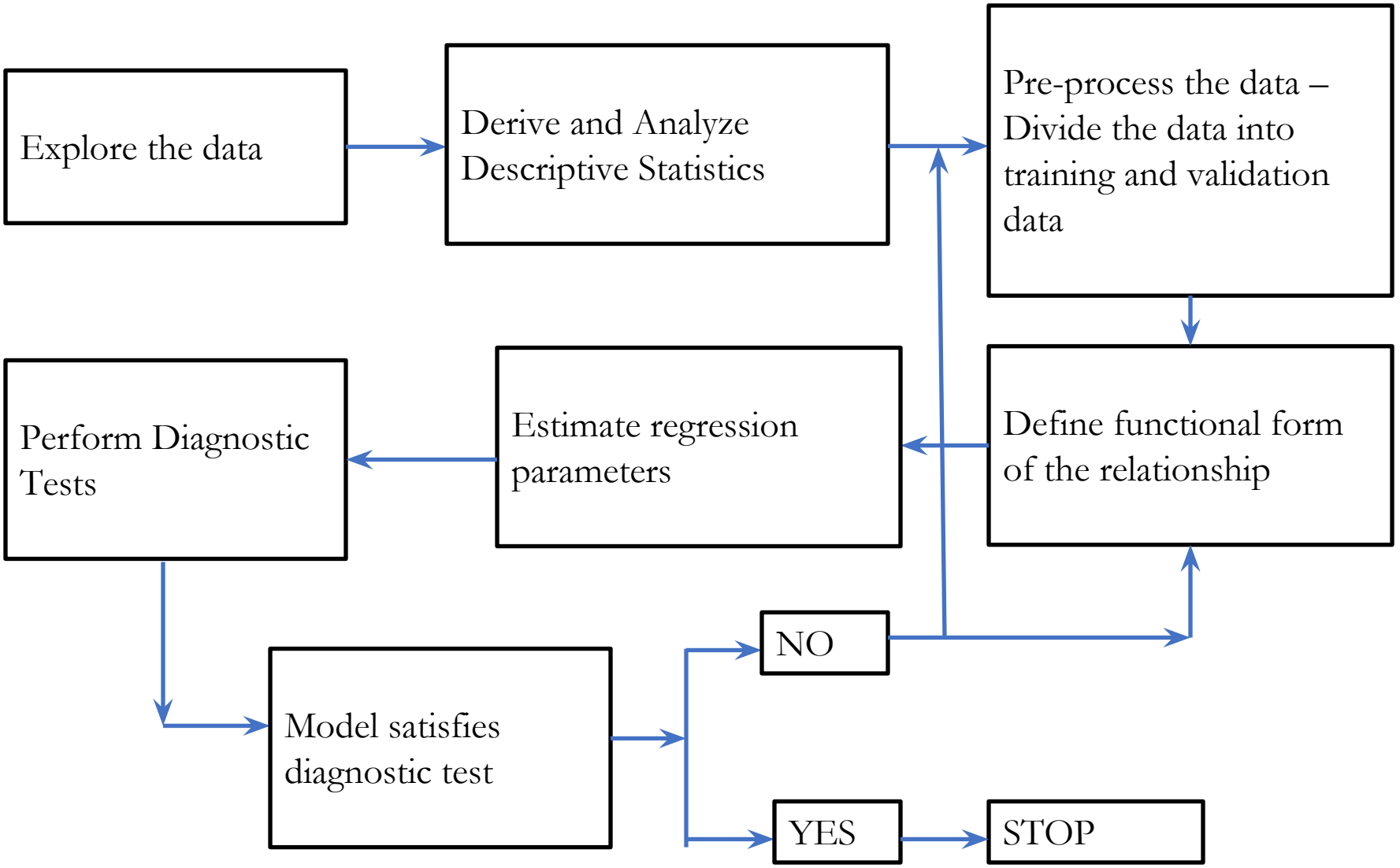
The above system of equations are solved iteratively to estimate β_0 and β_1

Limitations of MLE

- Maximum likelihood estimator may not be unique or may not exist.

- Closed form solution may not exist for many cases, one may have to use iterative procedure to estimate the parameter values.

Logistic Regression Model Development

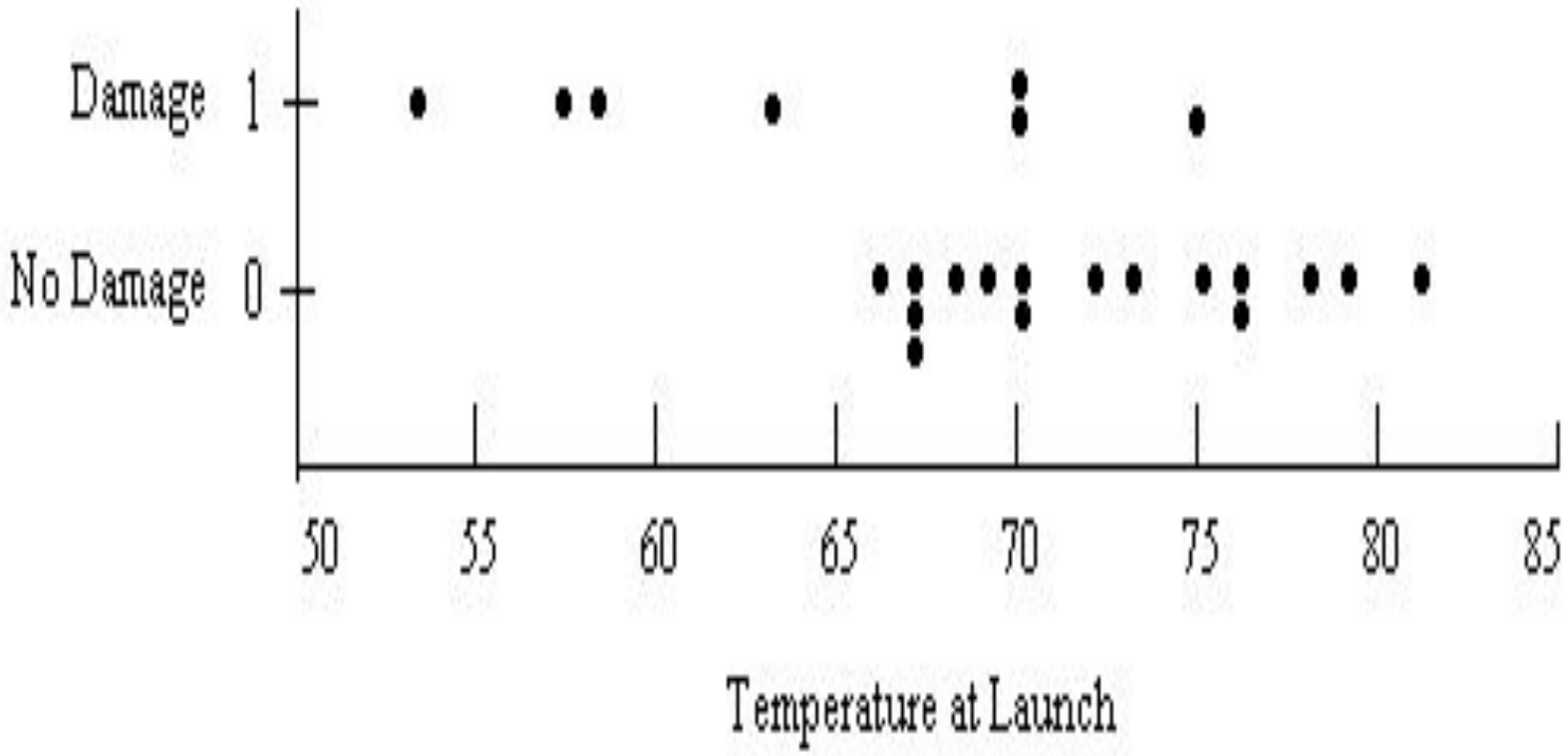


Example 1: Challenger Crash

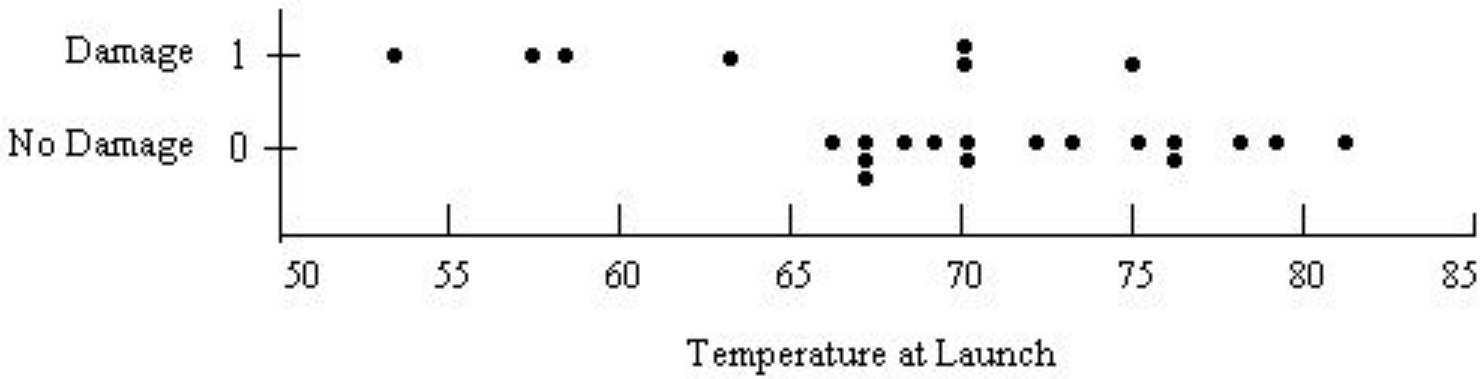
Challenger Data

Flt	Temp	Damage
STS-1	66	No
STS-2	70	Yes
STS-3	69	No
STS-4	80	No
STS-5	68	No
STS-6	67	No
STS-7	72	No
STS-8	73	No
STS-9	70	No
STS-41B	57	Yes
STS-41C	63	Yes
STS-41D	70	Yes

Flt	Temp	Damage
STS-41G	78	No
STS-51-A	67	No
STS-51-C	53	Yes
STS-51-D	67	No
STS-51-B	75	No
STS-51-G	70	No
STS-51-F	81	No
STS-51-I	76	No
STS-51-J	79	No
STS-61-A	75	Yes
STS-61-B	76	No
STS-61-C	58	Yes



Challenger launch temperature vs damage data



Challenger launch temperature vs damage data

Logistic Regression of challenger data

Let:

- $Y_i = 0$ denote no damage
- $Y_i = 1$ denote damage to the O-ring
- $P(Y_i = 1) = \Pi_i$ and $P(Y_i = 0) = 1 - \Pi_i$.
- We have to estimate $P(Y_i = 1 | X_i)$.

Variables in the Equation

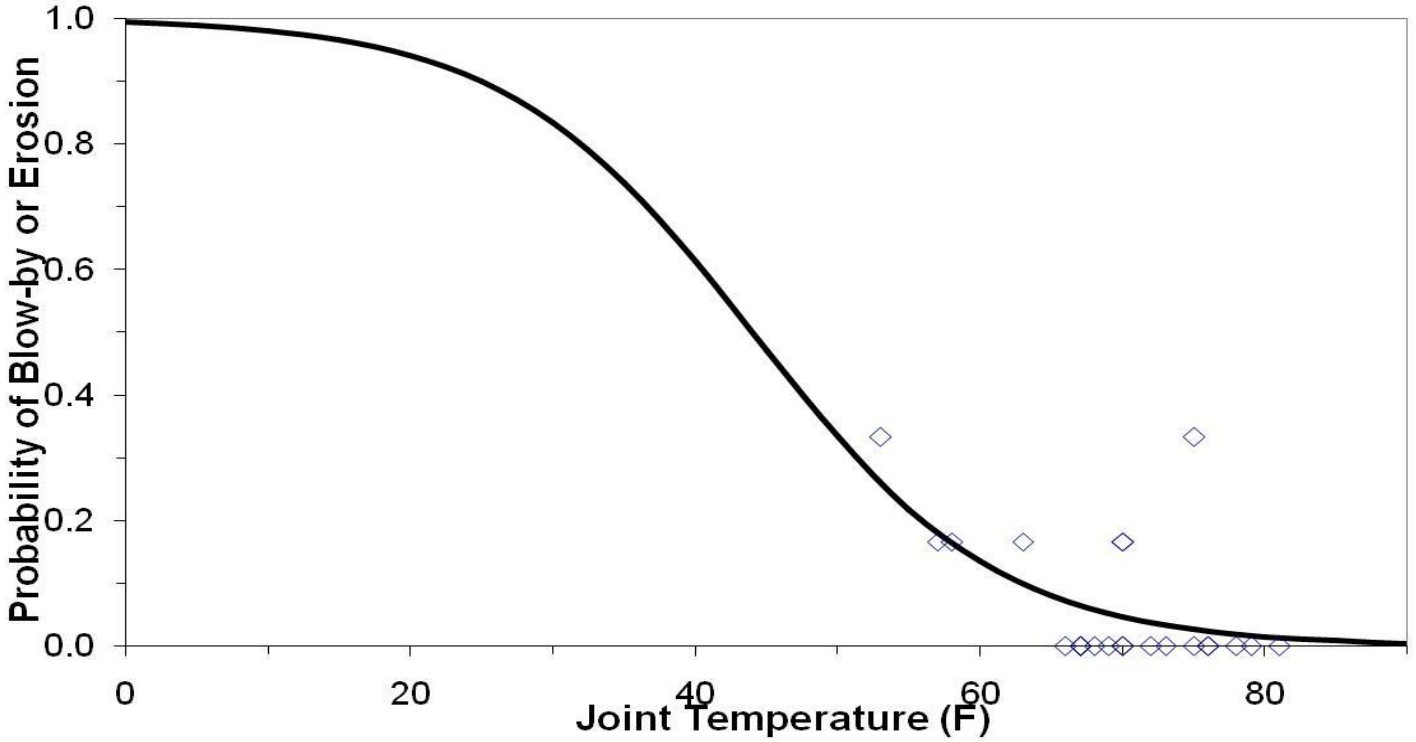
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Launch Temperature	-.236	.107	4.835	1	.028	.790
	Constant	15.297	7.379	4.297	1	.038	4.398E+06

a. Variable(s) entered on step 1: Launch Temperature.

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = 15.297 - 0.236X_i$$

Challenger: Probability of failure estimate

$$\pi_i = \frac{e^{15.297 - 0.236 X_i}}{1 + e^{15.297 - 0.236 X_i}}$$



Classification table from SPSS

Classification Table[a]

Observed			Predicted		
			Damage to O-ring		
			0	1	Percentage Correct
Step 1	Damage to O-ring	0	17	0	100.0
		1	3	4	57.1
	Overall Percentage				87.5

a. The cut value is .500

Case: German Credit Rating

Data Source: University of California, Irvine Machine Learning Repository

Link:
https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29

Model $$P(Y=1)=\frac{e^{Z}}{1+e^{Z}}$$

$$Z=\beta_0+\beta_1\times \text{Duration}$$

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Duration	.036	.006	33.066	1	.000	1.037
	Constant	-1.635	.161	103.092	1	.000	.195

a. Variable(s) entered on step 1: Duration.

Z = -1.653 + 0.036 x Duration

P-value < 0.05

Classification Table

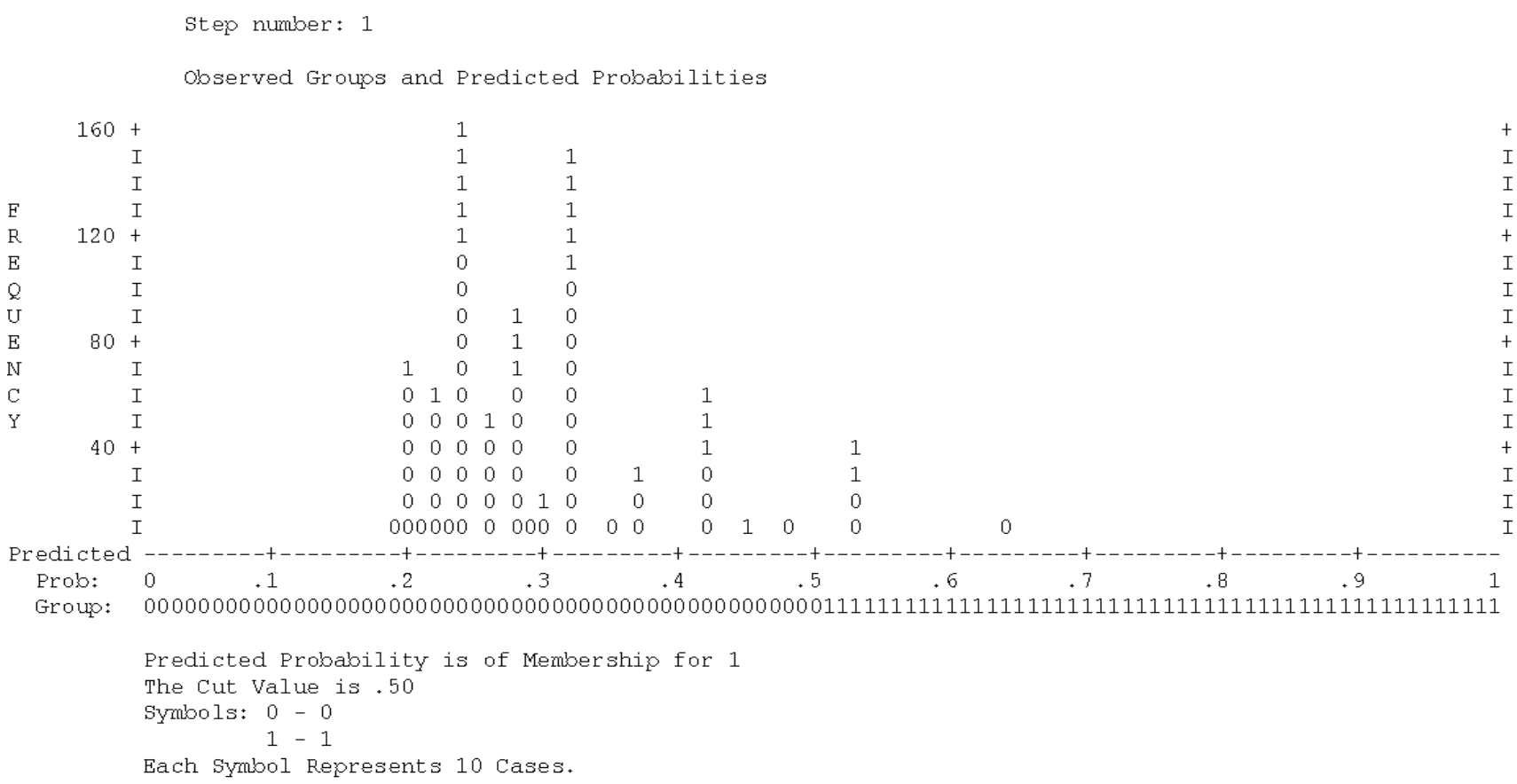
Classification Table[a]

Observed			Predicted		
			Credit Rating		Percentage Correct
			0	1	
Step 1	Credit Rating	0	535	26	95.4
		1	211	28	11.7
	Overall Percentage				70.4

a. The cut value is .500

Classification cut-off probability = 0.5

Classification Plot



Credit Class Vs Marital Status

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	MaritalStatusFD	-.344	.351	.957	1	.328	.709
	MaritalStatusMM	-.705	.420	2.817	1	.093	.494
	MaritalStatusSM	-.890	.344	6.680	1	.010	.411
	Constant	-.211	.326	.419	1	.517	.810

a. Variable(s) entered on step 1: MaritalStatusFD, MaritalStatusMM, MaritalStatusSM.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	MaritalStatusSM	-.518	.156	11.091	1	.001	.596
	Constant	-.584	.109	28.411	1	.000	.558

a. Variable(s) entered on step 1: MaritalStatusSM.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	MaritalStatusSM	-.518	.156	11.091	1	.001	.596
	Constant	-.584	.109	28.411	1	.000	.558

a. Variable(s) entered on step 1: MaritalStatusSM.

$$P(Y=1 \mid \text{Single Male}) = \frac{e^{-1.102}}{1+\exp^{-1.102}} = 0.2494$$

$$P(Y=1 \mid \text{Other Marital Status}) = \frac{e^{-0.584}}{1+\exp^{-0.584}} = 0.3580$$

Classification table from SPSS

Classification Table[a]

Observed			Predicted		
			Credit Rating		Percentage Correct
			0	1	
Step 1	Credit Rating	0	328	233	58.5
		1	109	130	54.4
	Overall Percentage				57.3

a. The cut value is .300

Accuracy Paradox

- Assume an example of insurance fraud. Past data has revealed that out of 1000 claims in the past, 950 are true claims and 50 are fraudulent claims.

- The classification table using a logistic regression model is given below:

Observed	Predicted		% accuracy
	0	1	
0	900	50	94.73%
1	5	45	90.00%

The overall accuracy is 94.5%. Classifying all of them as true claims will give 95% accuracy!

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	@0DM	1.956	.227	74.440	1	.000	7.071
	lessthan200DM	1.704	.225	57.499	1	.000	5.494
	over200DM	.785	.373	4.441	1	.035	2.193
	Constant	-2.061	.179	131.939	1	.000	.127

a. Variable(s) entered on step 1: @0DM, lessthan200DM, over200DM.

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = -2.061 + 1.956 \times 0DM + 1.704 \times Lessthan200 + 0.785 \times Over200DM$$

CLASSIFICATION TABLE

Classification Table[a]

Observed			Predicted		
			Credit Rating		Percentage Correct
			0	1	
Step 1	Credit Rating	0	561	0	100.0
		1	239	0	.0
	Overall Percentage				70.1

a. The cut value is .500

Classification cut-off probability

ODDS and ODDS RATIO

ODDS and ODDS RATIO

- ODDS: Ratio of two probability values.

$$odds = \frac{\pi}{1-\pi}$$

- ODDS RATIO: Ratio of two odds.

ODDS RATIO

Assume that X is an independent variable (covariate). The odds ratio, OR, is defined as the ratio of the odds for X = 1 to the odds for X = 0. The odds ratio is given by:

$$OR = \frac{\pi(1)/1-\pi(1)}{\pi(0)/1-\pi(0)}$$

Odds Ratio for Binary Logistic Regression

$$OR = \frac{\pi(1)/1-\pi(1)}{\pi(0)/1-\pi(0)} = e^{\beta_1}$$

If OR = 2, then the event is twice likely to occur when X = 1 compared to X = 0.

Odds ratio approximates the relative risk.

Interpretation of LR coefficients

- β_1 is the change in log-odds ratio for unit change in the explanatory variable.

- β_1 is the change in odds ratio by a factor $\exp(\beta_1)$.

Interpretation of LR coefficients

$$\beta_1 = \ln\left(\frac{\pi(x+1)/(1-\pi(x+1))}{\pi(x)/(1-\pi(x+1))}\right) = \text{Change in ln odds ratio}$$

$$e^{\beta_1} = \frac{\pi(x+1)/(1-\pi(x+1))}{\pi(x)/(1-\pi(x+1))} = \text{Change in odds ratio}$$

LOGISTIC REGRESSION MODEL DIAGNOSTICS

Omnibus tests are generic statistical tests used for checking whether the variance explained by the model is more than the unexplained variance.

The log likelihood function for binary logistic regression model is given by

$$LL = \sum_{i=1}^{n} Y_i \ln[\pi(Z)] + \sum_{i=1}^{n} (1 - Y_i)[\ln(1 - \pi(Z))]$$

Wald's test

Wald's test is used for checking statistical significance of individual predictor variables (equivalent to *t*-test in MLR model). The null and alternative hypotheses for Wald's test are:

H_0: $\beta_i = 0$

H_1: $\beta_i \neq 0$

Wald's test statistic is given by

$$W = \left[\frac{\hat{\beta}_i}{S_e(\hat{\beta}_i)} \right]^2$$

Sensitivity, Specificity and Precision

- The ability of the model to correctly classify positives and negatives are called sensitivity and specificity, respectively.
- The terminologies sensitivity and specificity originated in medical diagnostics.
- In generic case

Sensitivity = P(model classifies Y_i as positive | Y_i is positive)

Sensitivity is calculated using the following equation:

$$\text{Sensitivity} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}}$$

where True Positive (TP) is the number of positives correctly classified as positives by the model and False Negative (TN) is positives misclassified as negative by the model. Sensitivity is also called as **recall**.

Specificity

- ***Specificity*** is the ability of the diagnostic test to correctly classify the test as negative when the disease is not present. That is:

Specificity = *P*(diagnostic test is negative | patient has no disease)

- In general:

Sensitivity = *P*(model classifies Y_i as negative | Y_i is negative)

Specificity can be calculated using the following equation:

$$\text{Specificity} = \frac{\text{True Negative (TN)}}{\text{True Negative (TN)} + \text{False Positive (FP)}}$$

where True Negative (TN) is number of the negatives correctly classified as negatives by the model and False Positive (FP) is number of negatives misclassified as positives by the model.

- The decision maker has to consider the tradeoff between sensitivity and specificity to arrive at an optimal cut-off probability.
- Precision measures the accuracy of positives classified by the model.

Precision = *P*(patient has disease | diagnostic test is positive)

$$\text{Precision} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Positive (FP)}}$$

- **F Score (F Measure)** is another measure used in binary logistic regression that combines both precision and recall and is given by:

$$\text{F} - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

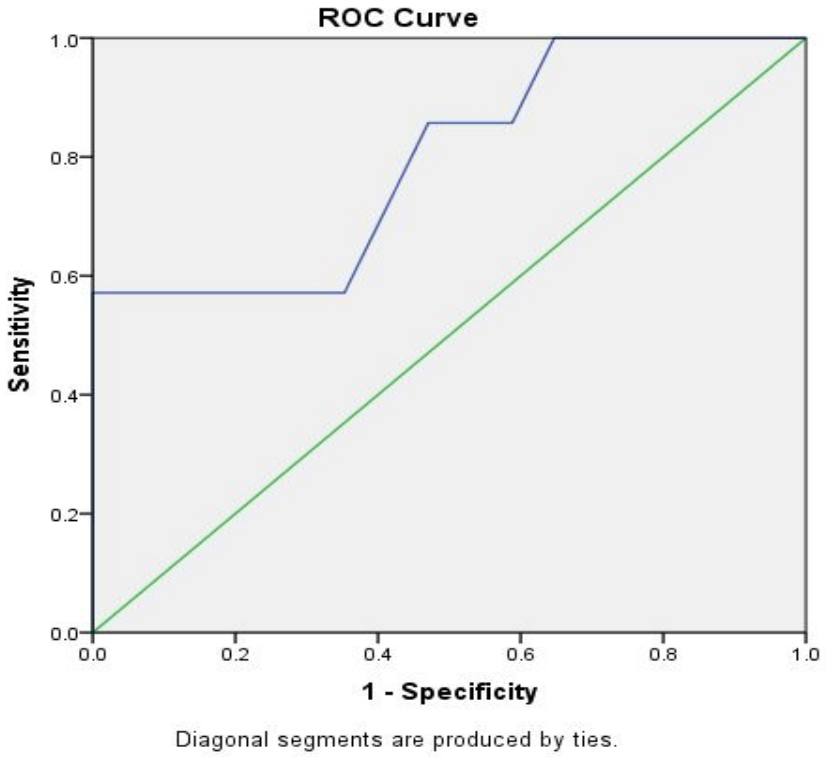
Concordant and Discordant Pairs

- **Discordant Pairs.** A pair of positive and negative observations for which the model has no cut-off probability to classify both of them correctly are called discordant pairs.

- **Concordant Pairs.** A pair of positive and negative observations for which the model has a cut-off probability to classify both of them correctly are called concordant pairs.

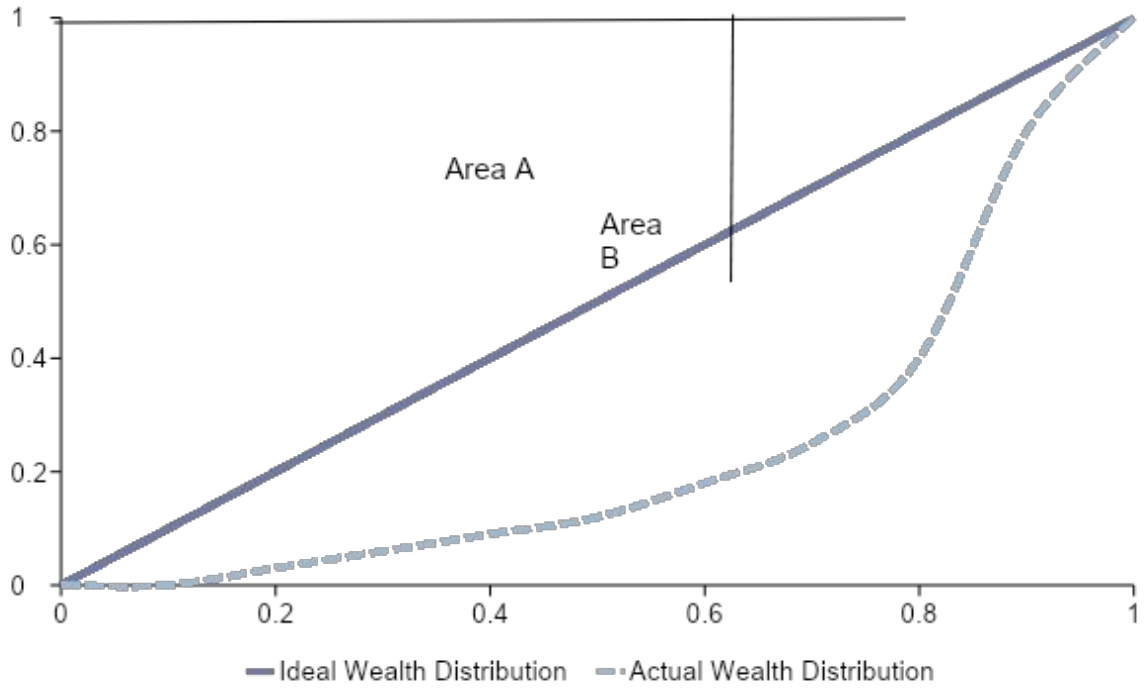
Receiver Operating Characteristics (ROC) Curve

- ROC curve is a plot between sensitivity (true positive rate) in the vertical axis and 1 – specificity (false positive rate) in the horizontal axis.



Area Under ROC Curve (AUC), Lorenz Curve and Gini Coefficient

- Gini coefficient = $\left(\frac{\text{Area A}}{\text{Area A+ Area B}} \right)$



German Credit Rating – Final Model

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 14[n]	Duration	.029	.010	9.079	1	.003	1.030
	@0DM	1.851	.247	56.230	1	.000	6.366
	lessthan200DM	1.551	.245	39.995	1	.000	4.715
	over200DM	.843	.401	4.431	1	.035	2.324
	critical	-.781	.248	9.905	1	.002	.458
	Bankpaid	1.001	.394	6.466	1	.011	2.722
	CreditAmount	.000	.000	4.724	1	.030	1.000
	lessthan100	.801	.225	12.697	1	.000	2.228
	less500	.630	.325	3.762	1	.052	1.878
	SevenYears	-.708	.258	7.559	1	.006	.492
	Install_rate	.338	.090	14.038	1	.000	1.402
	MaritalStatusSM	-.708	.187	14.388	1	.000	.493
	CoapplicantGaurantor	-1.245	.442	7.924	1	.005	.288
	Num_Credits	.359	.180	3.971	1	.046	1.432
	Constant	-4.349	.487	79.694	1	.000	.013

Observed			Predicted		
			Credit Rating		Percentage Correct
			0 (Negative)	1 (positive)	
Step 14	Credit Rating	0 (Negative) 561	507	54	90.4
		1 (Positive) 239	124	115	48.1
	Overall Percentage				77.8

$$\text{Sensitivity} = \left(\frac{TP}{TP + FN}\right) = \left(\frac{115}{115 + 124}\right) = 48.1$$

$$\textit{Specificity} = \left(\frac{TN}{TN + FP}\right) = \left(\frac{507}{507 + 54}\right) = 90.4$$

Sensitivity & Specificity

$$\text{Sensitivity} = \frac{\text{No of true positives}}{\text{Number of true positives} + \text{Number of false negatives}}$$

Sensitivity is the conditional probability that the predicted value of y = 1 given that the observed value is 1 (also known as **recall**)

$$\text{Specificity} = \frac{\text{No of true negatives}}{\text{Number of true negatives} + \text{Number of false positives}}$$

Specificity is the conditional probability that the predicted value of y = 0 given that the observed value is 0

Precision

- Precision measures the ratio of true positive among cases that are classified as positives:

- Precision = True Positive / (True Positive + False Positive)

F Score (or F_1 Score)

- F Score is a measure that combines both precision and recall and is given b

$$\text{F} - \text{Score} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

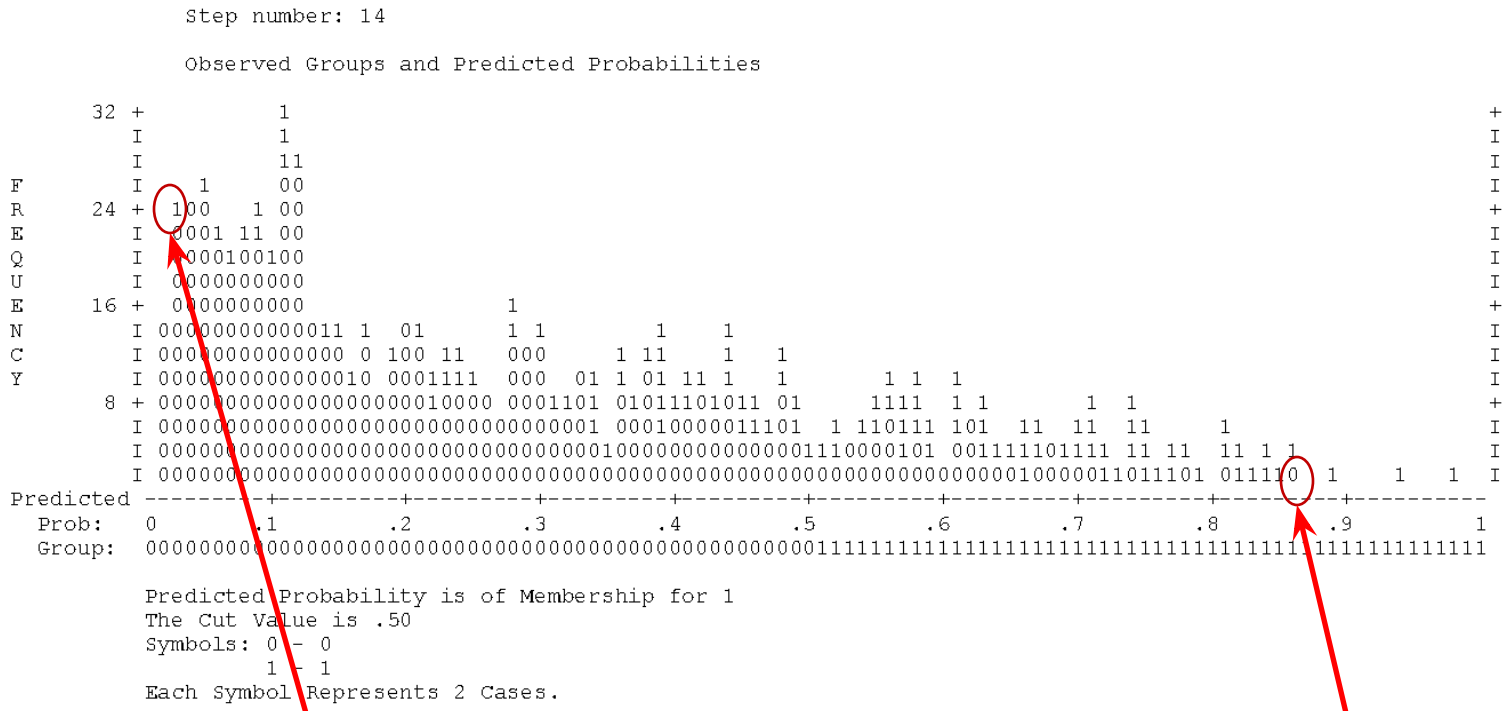
Measures of Classification

Measure	Interpretation
Sensitivity (aka Recall)	P(predicted class is positive | Class is positive)
Specificity	P(predicted class is negative | Class is negative)
Precision	P(class is positive | predicted class is positive)
F-Score	Harmonic mean of Precision and Recall

Concordant and Discordant Pairs

- Divide the dataset into positives (y=1) and negatives (y=0).
- For a randomly chosen positive and negative, if the probability of positive (obtained using logistic regression model) is greater than probability of negative then such pairs are called concordant pairs.
- For a randomly chosen positive and negative, if the probability of positive is less than probability of negative then such pairs are called discordant pairs.
- Area under the ROC curve is the proportion of concordant pairs in the dataset.

Classification Plot



The model predicts less than 5% chance of default

The model predicts more than 85% chance of default

Receiver Operating Characteristics (ROC) Curve

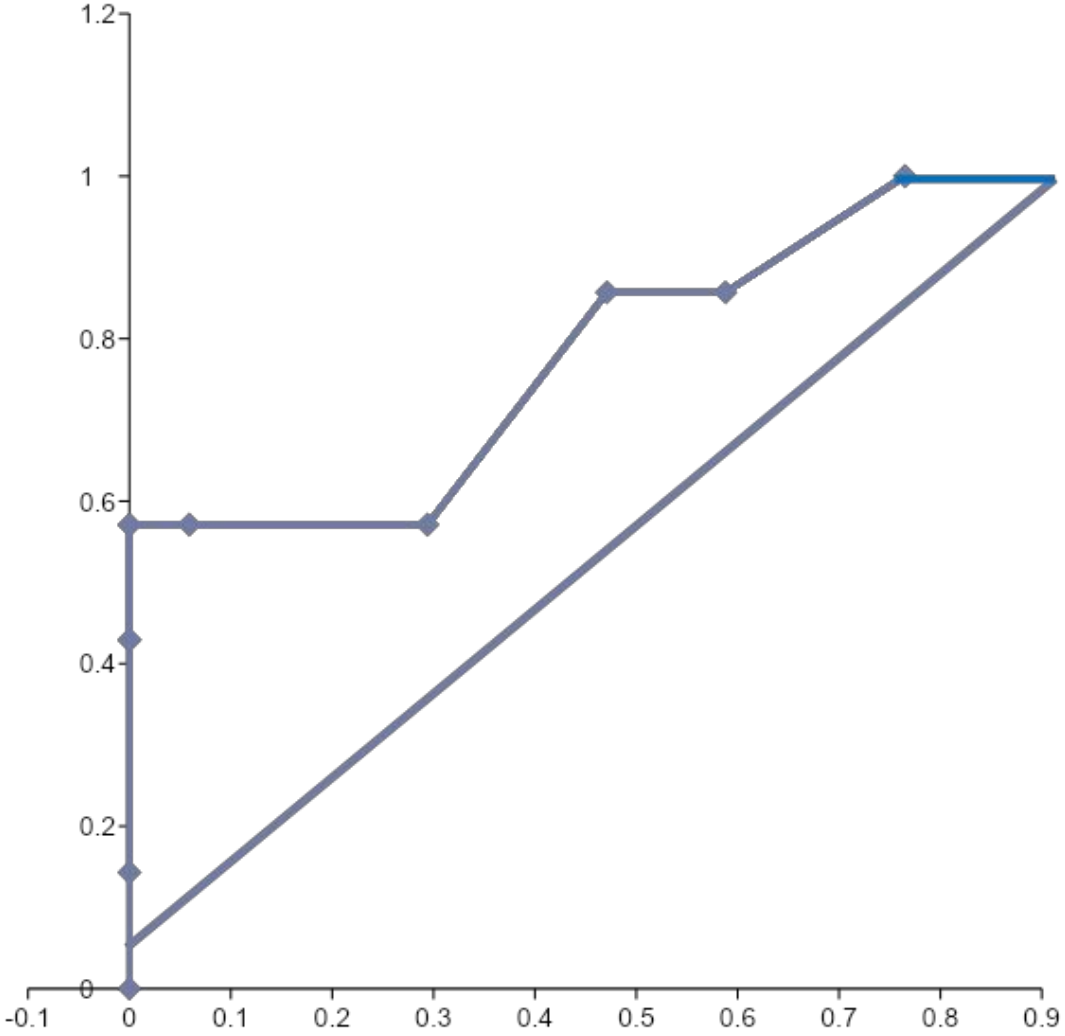
- ROC curve plots the true positive ratio (right positive classification) against the false positive ratio (1-specificity) and compares it with random classification.

- The higher the area under the ROC curve, the better the prediction ability.

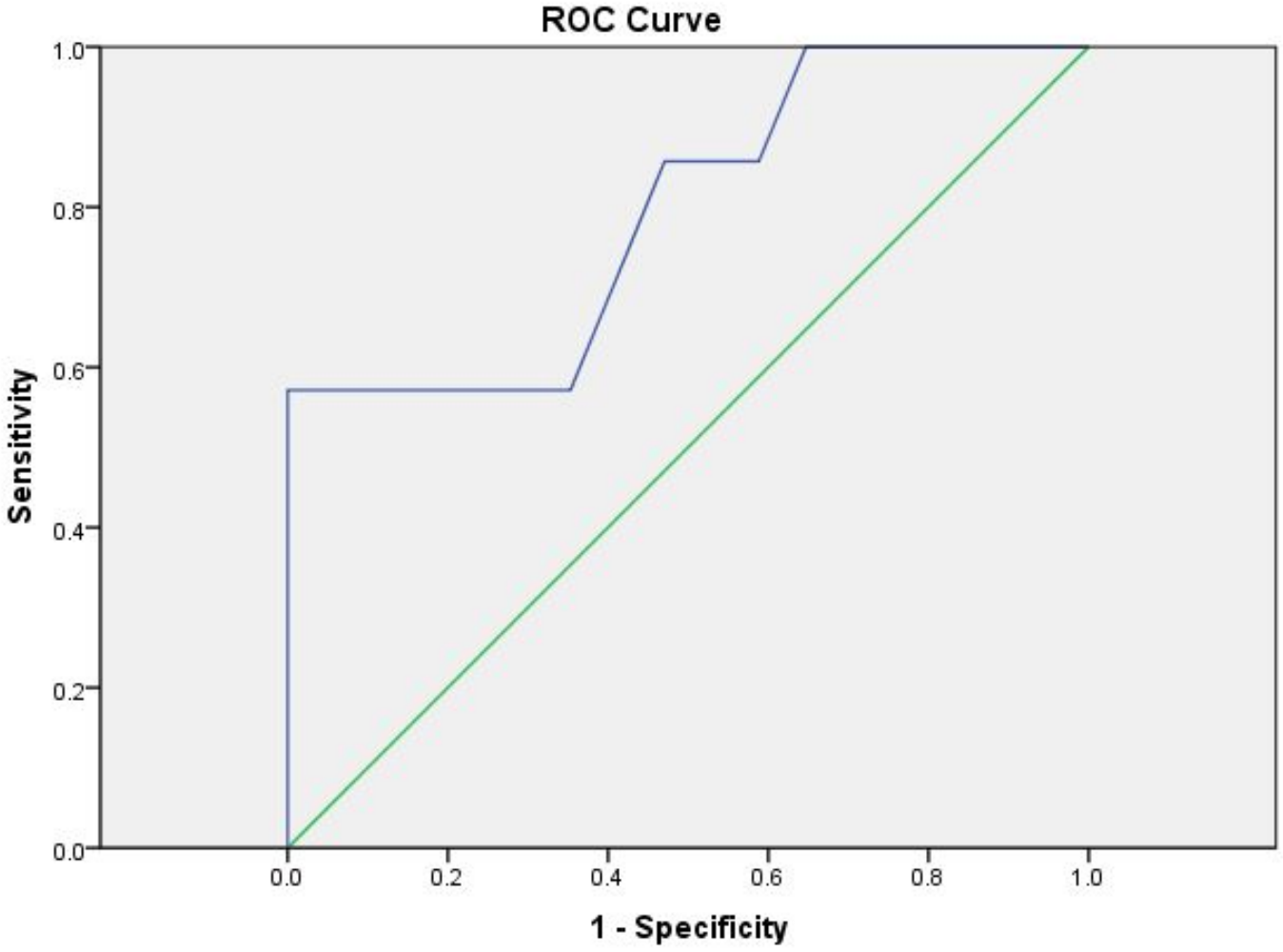
Challenger Example: Sensitivity Vs 1-Specificity (true positive Vs false positive)

Cut off Value	Sensitivity	Specificity	1-specificity
0.05	1	0.235	0.765
0.1	0.857	0.412	0.588
0.2	0.857	0.529	0.471
0.3	0.571	0.706	0.294
0.4	0.571	0.941	0.059
0.5	0.571	1	0
0.6	0.571	1	0
0.7	0.429	1	0
0.8	0.429	1	0
0.9	0.143	1	0
0.95	0	1	0

ROC Curve – Challenger Example



ROC Curve



Diagonal segments are produced by ties.

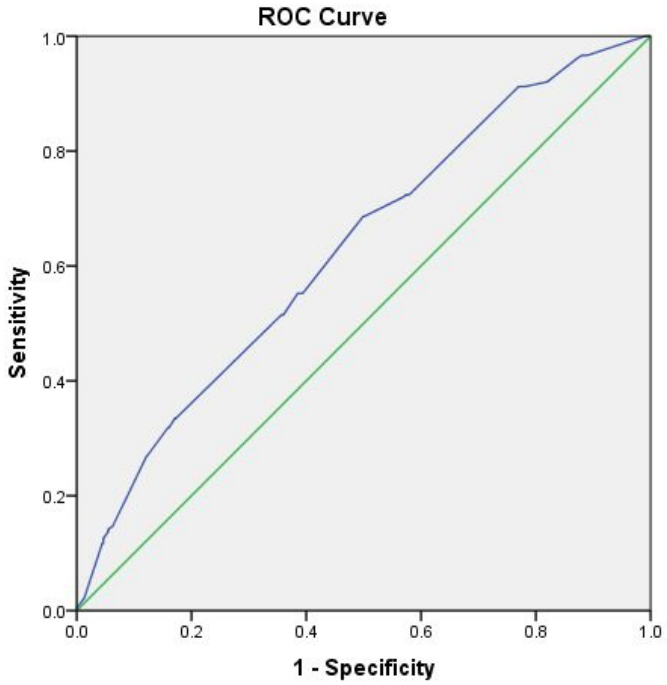
Area Under the Curve

Test Result Variable(s):Predicted probability

Area
.794

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

ROC and Area Under ROC



Diagonal segments are produced by ties.

German Credit Rating with Duration as covariate

Area Under the Curve

Test Result Variable(s): Predicted probability

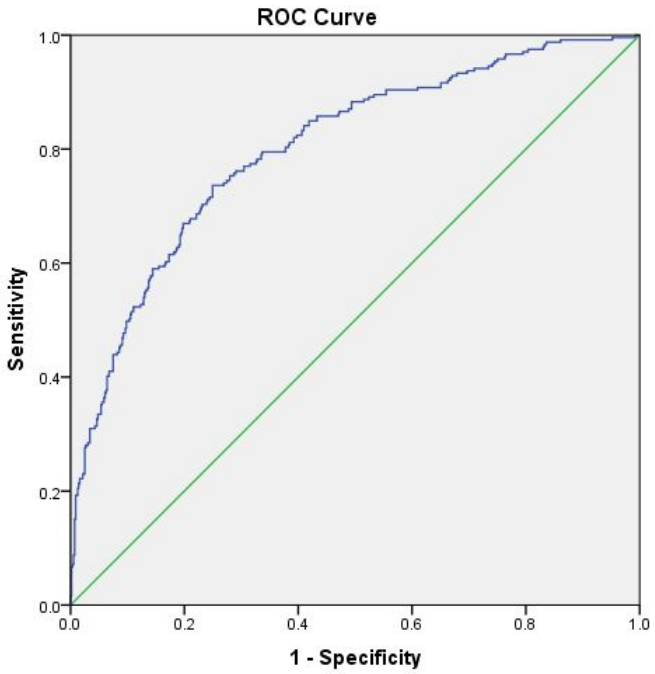
Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.629	.021	.000	.587	.670

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

ROC and Area Under ROC



German Credit Rating with after inclusion of all variables

Area Under the Curve

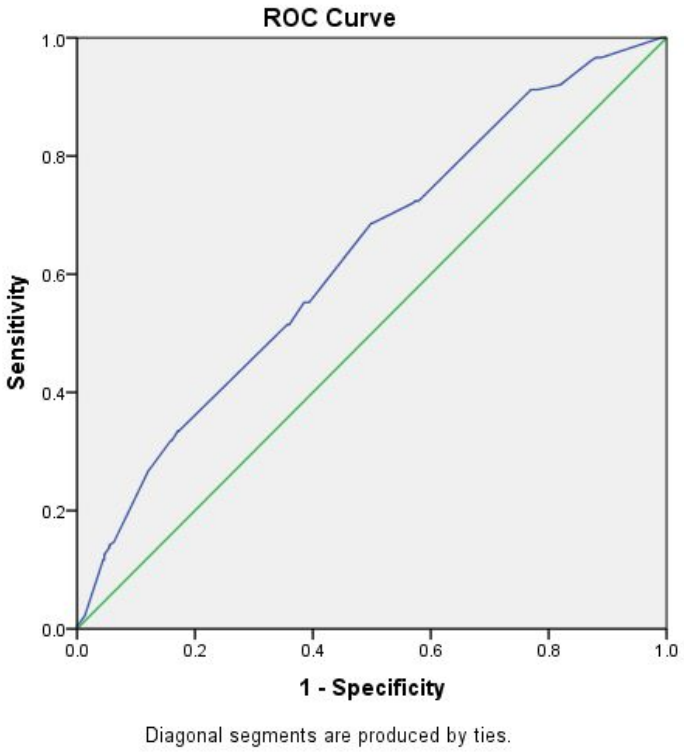
Test Result Variable(s): Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.801	.017	.000	.768	.835

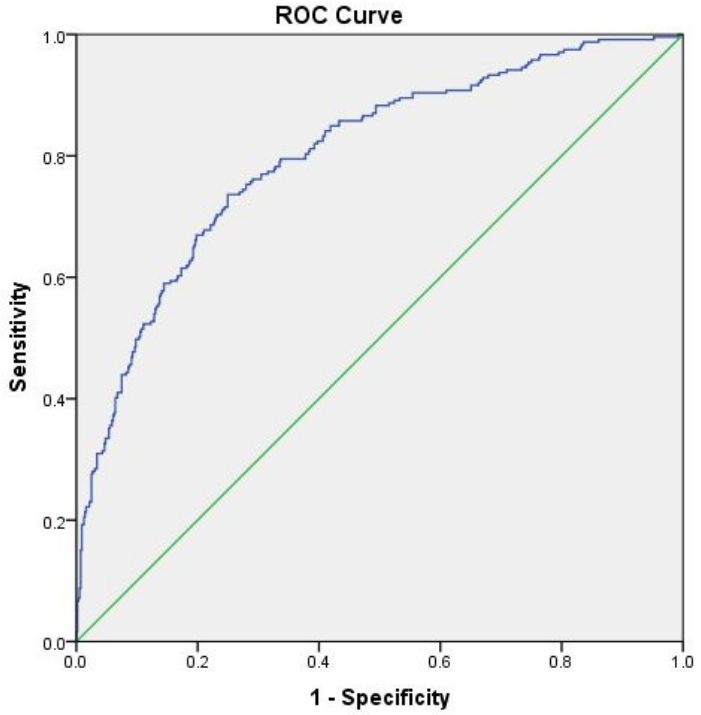
a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Area Under the ROC Curve



AUC = 0.629



AUC = 0.801

Area Under the ROC Curve

- Area under the ROC (AUC) curve is interpreted as the probability that the model will rank a randomly chosen positive higher than randomly chosen negative.

- If n1 is the number of positives (1s) and n2 is the number of negatives (0s), then the area under the ROC curve is the proportion of cases in all possible combinations of (n1, n2) such that n1 will have higher probability than n2.

AUC = P (Random Positive Observation) > P(Random Negative Observation)

Area Under the ROC Curve

Area Under the ROC Curve (AUC) is a measure of the ability of the logistic regression model to discriminate positives and negatives correctly.