Session 4
Classification Methods I

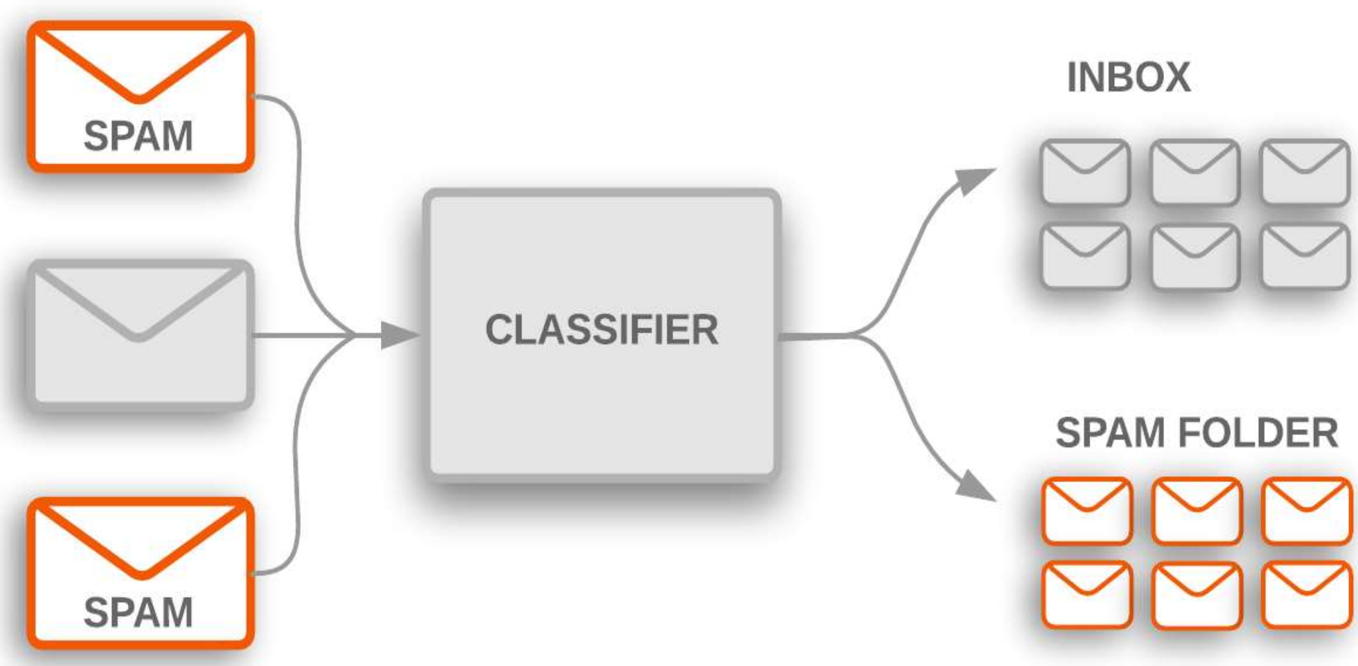
PSYC 560
Heungsun Hwang

Classification

- Classification is a process of explaining or predicting a nominal variable with multiple response categories, classes or labels
  - Assigns an observation to a category

- Popular classification methods or *classifiers*:
  - Logistic regression
  - Discriminant analysis
  - Naïve Bayes
  - K-nearest neighbors

Classification Problems

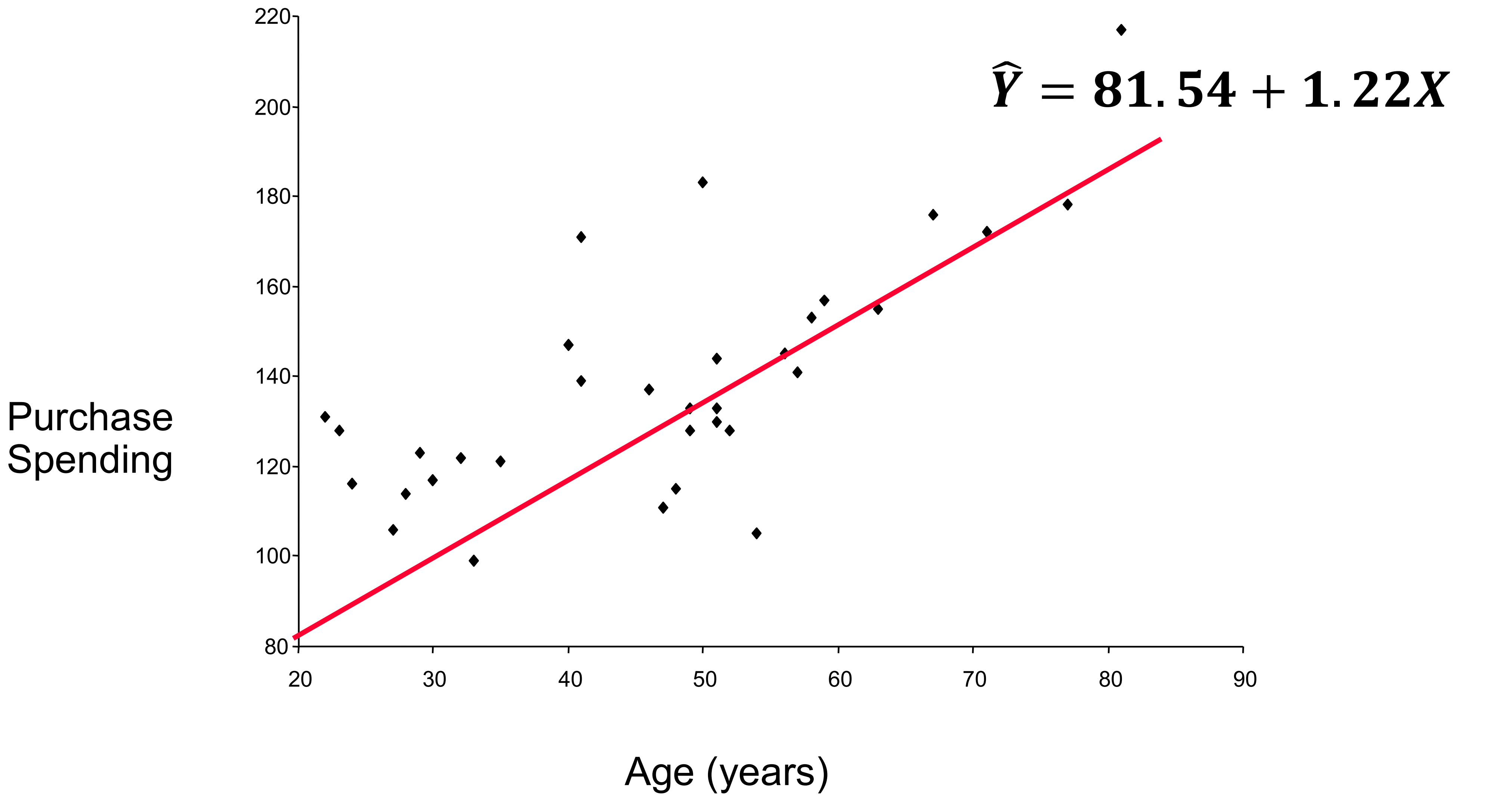


https://medium.com/@naveeen.kumar.k/naive-bayes-spam-detection-7d087cc96d9d

Linear Regression

Table 1. Age and purchase spending ($) among 33 adult women

Age	$	Age	$	Age	$
22	131	41	139	52	128
23	128	41	171	54	105
24	116	46	137	56	145
27	106	47	111	57	141
28	114	48	115	58	153
29	123	49	133	59	157
30	117	49	128	63	155
32	122	50	183	67	176
33	99	51	130	71	172
35	121	51	133	77	178
40	147	51	144	81	217



$\widehat{Y} = 81.54 + 1.22X$
220
200
180
160
140
120
100
80
20
30
40
50
60
70
80
90
Purchase Spending
Age (years)

Linear Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_P X_P + e$$

- Intercept (β_0):
  - Average value of Y when $X_p = 0$
- Slope (β_p):
  - Amount by which y changes on average when X_j changes by one unit, holding all the other X_ps constant.
- Regression coefficients are typically estimated via least squares.

What is Logistic Regression?

- Used when you have a binary response:
  - Yes – no
  - Positive – negative
  - Good credit – bad credit
  - Buyer – not buyer
  - Left – stayed
  - Dead – alive

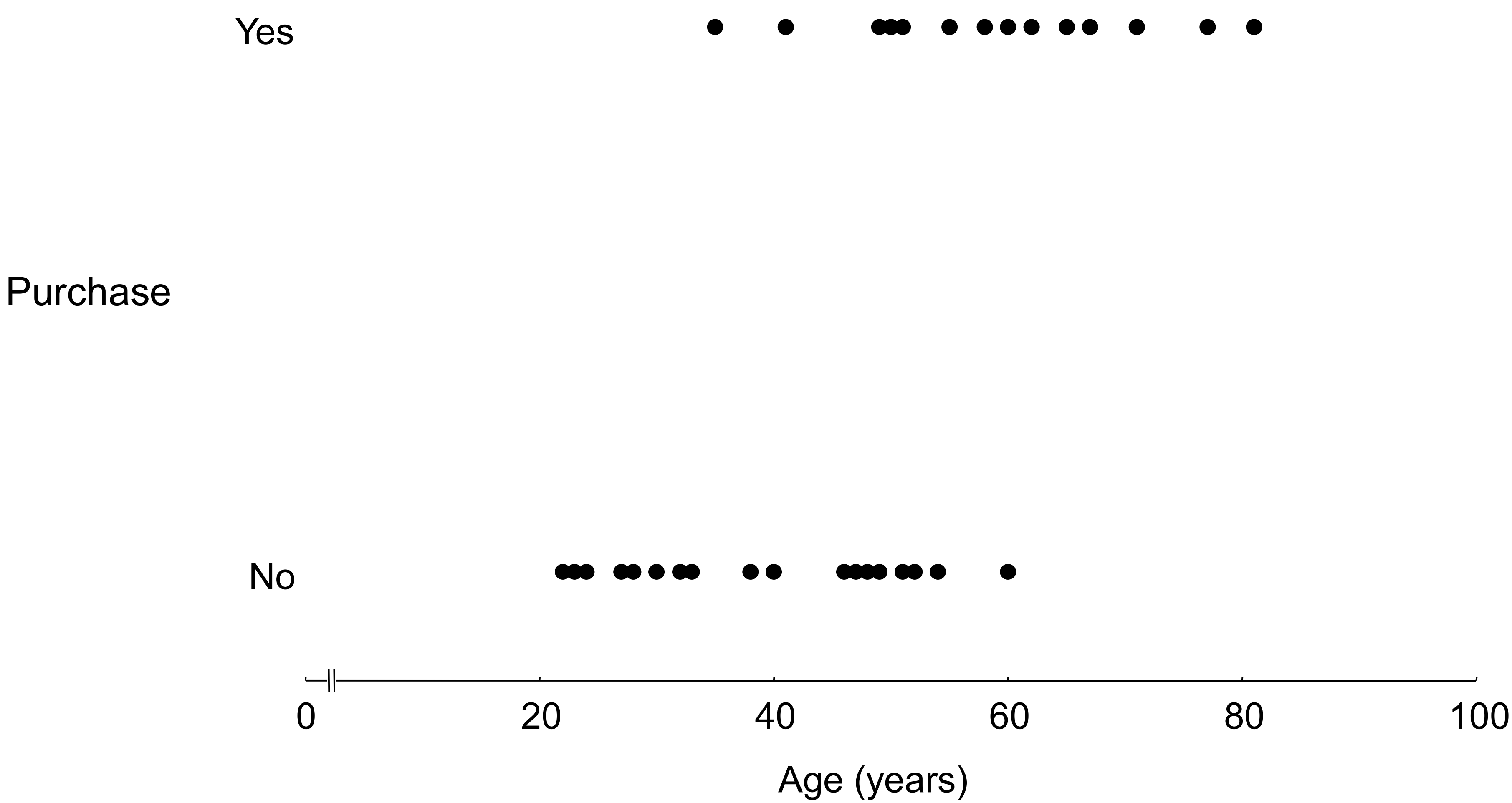
Logistic Regression

Table 2. Age (x) and CD Purchase (yes = 1, no = 0)

Age	CD	Age	CD	Age	CD
22	0	40	0	54	0
23	0	41	1	55	1
24	0	46	0	58	1
27	0	47	0	60	1
28	0	48	0	60	0
30	0	49	1	62	1
30	0	49	0	65	1
32	0	50	1	67	1
33	0	51	0	71	1
35	1	51	1	77	1
38	0	52	0	81	1

Logistic Regression

Data from Table 2



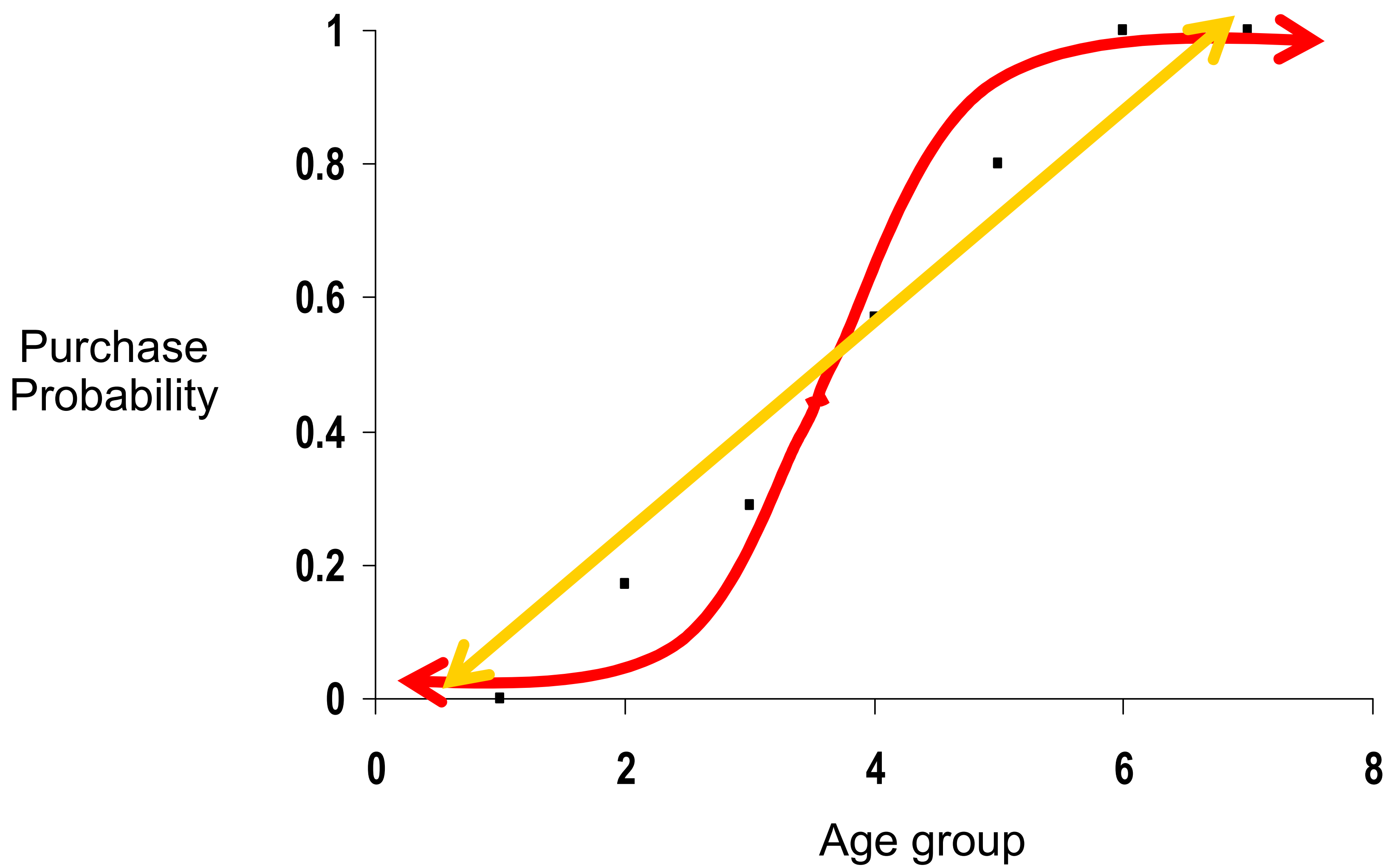
Logistic Regression

Table 3. Probability of CD purchase per age group

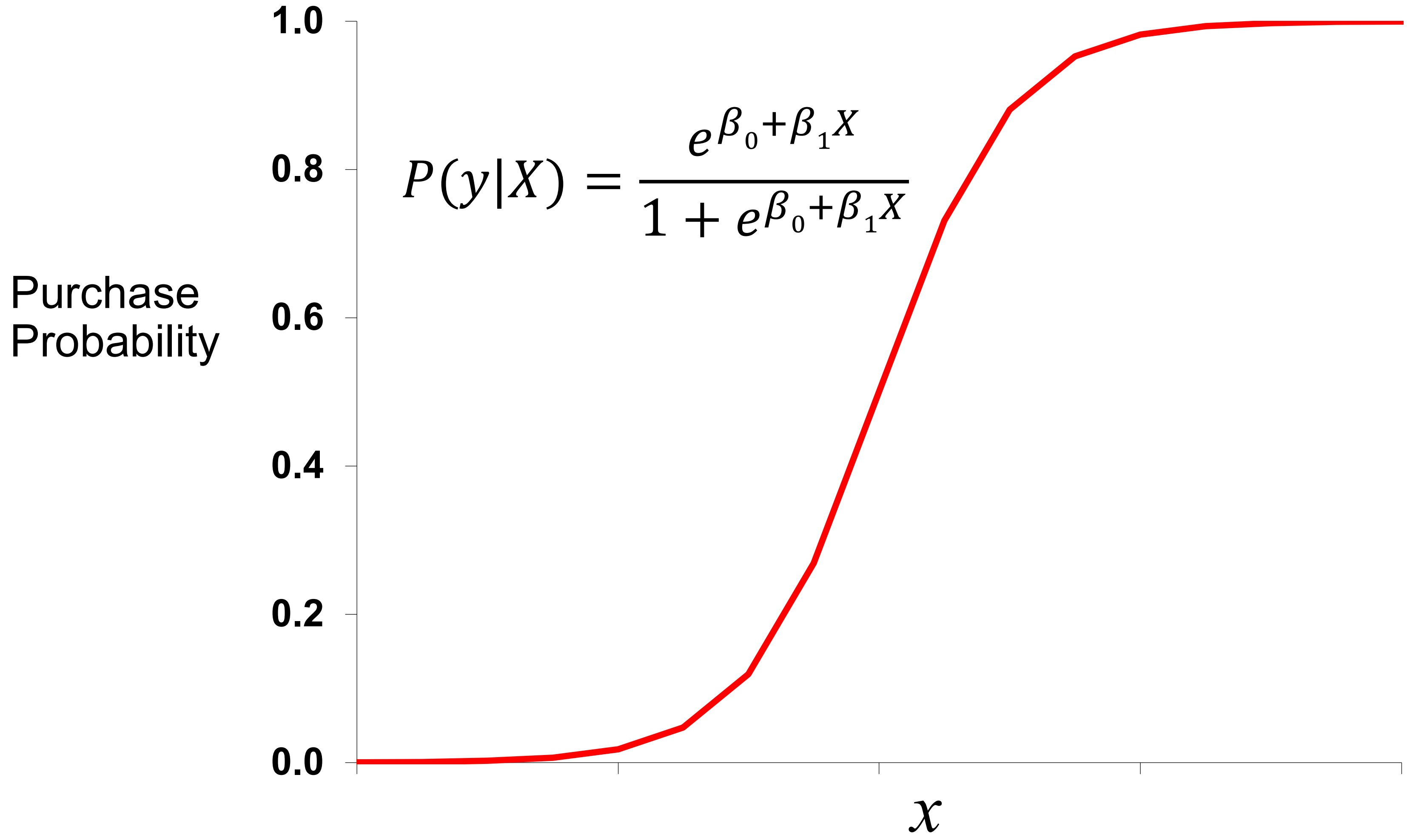
		CD Purchase	
Age group	# in group	#	prob
20 - 29	5	0	0
30 - 39	6	1	.17
40 - 49	7	2	.29
50 - 59	7	4	.57
60 - 69	5	4	.80
70 - 79	2	2	1.0
80 - 89	1	1	1.0

$p(y = 1 | x = \text{age group})$

Logistic Regression



Logistic Function



Logistic Transformation

$$P(y|X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

$$\underbrace{ln\left[\frac{P(y|X)}{1 - P(y|X)}\right]}_{\text{logit of } P(y|x)} = \beta_0 + \beta_1 X$$

$$f(X) = \beta_0 + \beta_1 X$$

Logistic Regression Model

- The statistical model for logistic regression is

$$ln\left[\frac{P(y|X)}{1-P(y|X)}\right] = \beta_0 + \beta_1 X,$$

where

$$P(y|X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Multiple Logistic Regression

- More than one predictor
  - Can be discrete or continuous

$$ln\left[\frac{P(y|X)}{1-P(y|X)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_P X_P$$

- Nominal predictors are dummy coded.
- We typically use a method called maximum likelihood to estimate the coefficients from the training data.

Interpreting Coefficients

- Odds: The ratio of the proportions for two possible outcomes. If p is the proportion for one outcome, then 1-p is the proportion for the other outcome.

$$\text{Odds} = \frac{p}{1-p}$$

  - For example, p = the proportion of binge drinkers in college = .2, and 1-p = the proportion of students who are not binge drinkers = .8. Then, the odds of a college student being a binge drinker are .25 (or ¼).
  - The odds can take on any value between 0 and ∞. The larger, the higher probability of y = 1.

Interpreting Coefficients

- For the binge-drinking example, let's consider a predictor (X): Sex (1 = men & 0 = women).
- The log odds for men:

$$\ln\left(\frac{P}{1-P}\right)_{M} = \beta_0 + \beta_1$$

- The log odds for women:

$$\ln\left(\frac{P}{1-P}\right)_{W} = \beta_0$$

Interpreting Coefficients

- The slope β_1 indicates the difference between the log odds for men and women.

$$\ln\left(\frac{P}{1-P}\right)_M - \ln\left(\frac{P}{1-P}\right)_W = \beta_1$$

- This can be re-expressed as Odds Ratio (OR)

$$\frac{\text{Odds}_m}{\text{Odds}_w} = e^{\beta_1}$$

Interpreting Coefficients

- For example,

$$\frac{\textbf{Odds}_m}{\textbf{Odds}_w} = 1.4 \text{ or } \textbf{Odds}_m = 1.4 \times \textbf{Odds}_w$$

Then, the odds for men are 1.4 times for the odds for women.

- Exp(β_p) = Odds Ratio (OR): Change in the odds by one unit increase in X_p with all the other X's constant.

Interpreting Coefficients

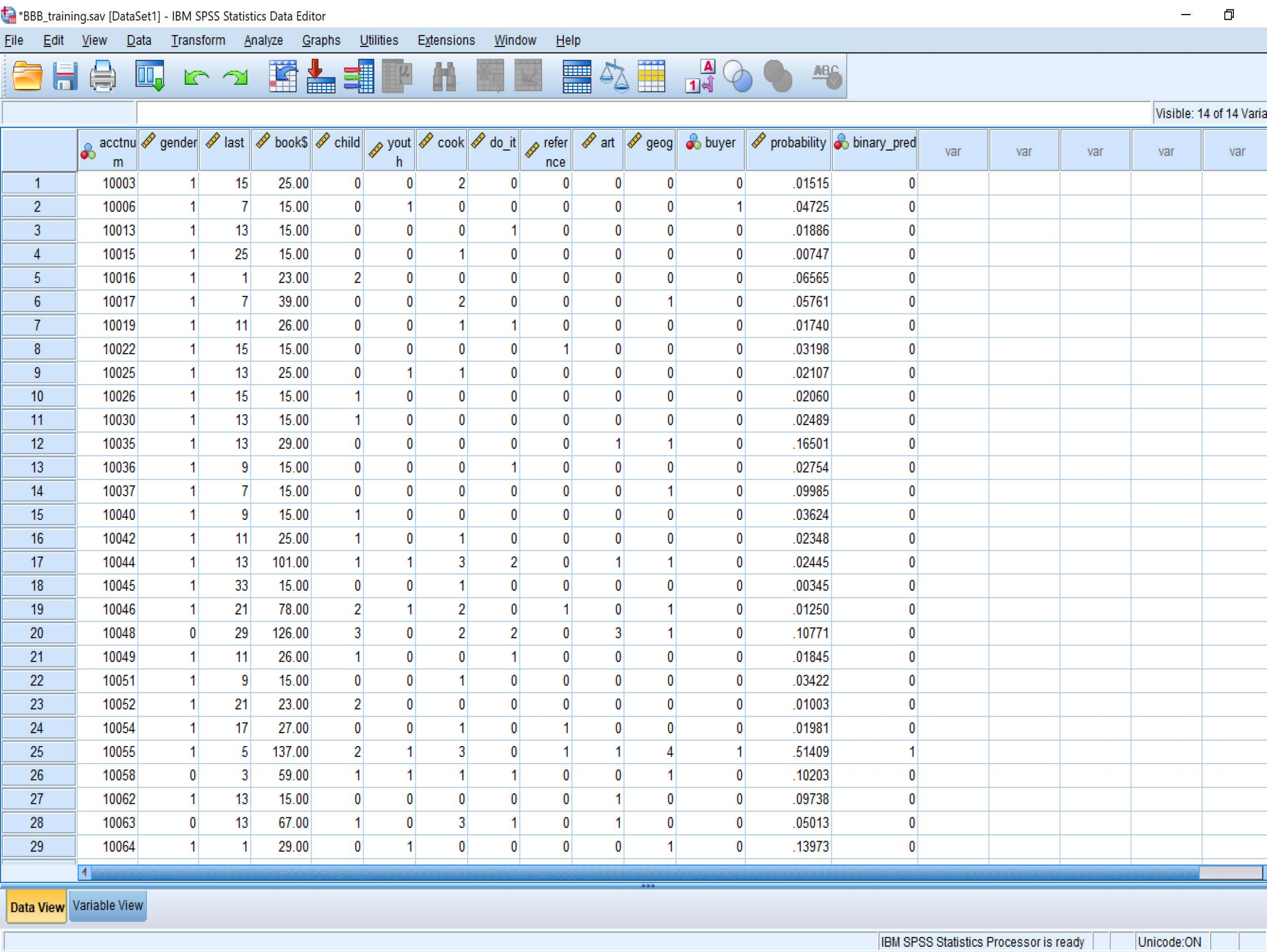
- Exp(β_p):
  - indicates how the odds that y = 1 will change if X_p increases by one unit with all the other Xs constant.
  - positive $\beta_p \rightarrow \exp(\beta_p) > 1$ (= odds increased)
  - negative $\beta_p \rightarrow \exp(\beta_p) < 1$ (= odds decreased)

- The statistical significance of an individual coefficient.
  - Wald test: $W_p = \dfrac{\hat{\beta}_p^2}{SE(\hat{\beta}_p)^2}$

Estimating Probabilities

- Once the coefficients are estimated, it is simple to compute the probability of y = 1 for any given X values in a training or test sample. For example,

$$\hat{P}(y|X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}}$$



	acctnum	gender	last	book$	child	youth	cook	do_it	reference	art	geog	buyer	probability	binary_pred	var	var	var	var	var
1	10003	1	15	25.00	0	0	2	0	0	0	0	0	.01515	0					
2	10006	1	7	15.00	0	1	0	0	0	0	0	1	.04725	0					
3	10013	1	13	15.00	0	0	0	1	0	0	0	0	.01886	0					
4	10015	1	25	15.00	0	0	1	0	0	0	0	0	.00747	0					
5	10016	1	1	23.00	2	0	0	0	0	0	0	0	.06565	0					
6	10017	1	7	39.00	0	0	2	0	0	0	1	0	.05761	0					
7	10019	1	11	26.00	0	0	1	1	0	0	0	0	.01740	0					
8	10022	1	15	15.00	0	0	0	0	1	0	0	0	.03198	0					
9	10025	1	13	25.00	0	1	1	0	0	0	0	0	.02107	0					
10	10026	1	15	15.00	1	0	0	0	0	0	0	0	.02060	0					
11	10030	1	13	15.00	1	0	0	0	0	0	0	0	.02489	0					
12	10035	1	13	29.00	0	0	0	0	0	1	1	0	.16501	0					
13	10036	1	9	15.00	0	0	0	1	0	0	0	0	.02754	0					
14	10037	1	7	15.00	0	0	0	0	0	0	1	0	.09985	0					
15	10040	1	9	15.00	1	0	0	0	0	0	0	0	.03624	0					
16	10042	1	11	25.00	1	0	1	0	0	0	0	0	.02348	0					
17	10044	1	13	101.00	1	1	3	2	0	1	1	0	.02445	0					
18	10045	1	33	15.00	0	0	1	0	0	0	0	0	.00345	0					
19	10046	1	21	78.00	2	1	2	0	1	0	1	0	.01250	0					
20	10048	0	29	126.00	3	0	2	2	0	3	1	0	.10771	0					
21	10049	1	11	26.00	1	0	0	1	0	0	0	0	.01845	0					
22	10051	1	9	15.00	0	0	1	0	0	0	0	0	.03422	0					
23	10052	1	21	23.00	2	0	0	0	0	0	0	0	.01003	0					
24	10054	1	17	27.00	0	0	1	0	1	0	0	0	.01981	0					
25	10055	1	5	137.00	2	1	3	0	1	1	4	1	.51409	1					
26	10058	0	3	59.00	1	1	1	1	0	0	1	0	.10203	0					
27	10062	1	13	15.00	0	0	0	0	0	1	0	0	.09738	0					
28	10063	0	13	67.00	1	0	3	1	0	1	0	0	.05013	0					
29	10064	1	1	29.00	0	1	0	0	0	0	1	0	.13973	0					

Performance Checking: Classification Table/Confusion Matrix

- One way of assessing the performance of a model is to look at the Classification Table or Confusion Matrix.
  - This is a 2 x 2 table which shows how correctly a model predicts the outcome category of cases.
  - The columns of the table are the two predicted values of the DV, while the rows are the two observed (actual) values of the DV.
  - In a perfect model, all cases will be on the diagonal and the overall percent correct will be 100%.
  - Note that this table should not be used as a goodness-of-fit measure because it ignores actual predicted probabilities and instead use dichotomized predictions based on a cutoff (e.g., .5).

Performance Checking: Classification Table/Confusion Matrix

Observed		Predicted: No (0)	Predicted: Yes (1)	Total
	No (0)	True Negative (TN)	False Positive (FP)	N
	Yes (1)	False Negative (FN)	True Positive (TP)	P
	Total	N*	P*	

Performance Checking: Classification Table/Confusion Matrix

- Two types of correct classification
  - Sensitivity: the percentage of Y = 1 (yes) cases that are correctly identified
    - 100*(TP/P)
  - Specificity: the percentage of Y = 0 (no) cases that are correctly identified
    - 100*(TN/N)

Decision (sample information)

		Retain H_0	Reject H_0
True state (population information)	H_0 true	Correct Decision	**Type I Error**
	H_0 false	**Type II Error**	Correct Decision

- α = probability of committing Type I error
- β = probability of committing Type II error
- 1- β (power) = probability of correctly rejecting a false H_0

Performance Checking: Classification Table/Confusion Matrix

		Predicted		
		No (0)	Yes (1)	Total
Observed	No (0)	True Negative (TN)	False Positive (FP)	N
	Yes (1)	False Negative (FN)	True Positive (TP)	P
	Total	N*	P*	

Performance Checking: Classification Table/Confusion Matrix

Name	Definition	Synonyms
False Positive Rate	FP/N	Type I error, 1 – Specificity
True Positive Rate	TP/P	Power, Sensitivity, Recall
Positive Predictive Value	TP/P*	Precision, 1 – False Discovery Proportion
Negative Predictive Value	TN/N*	

Performance Checking: Classification Table/Confusion Matrix

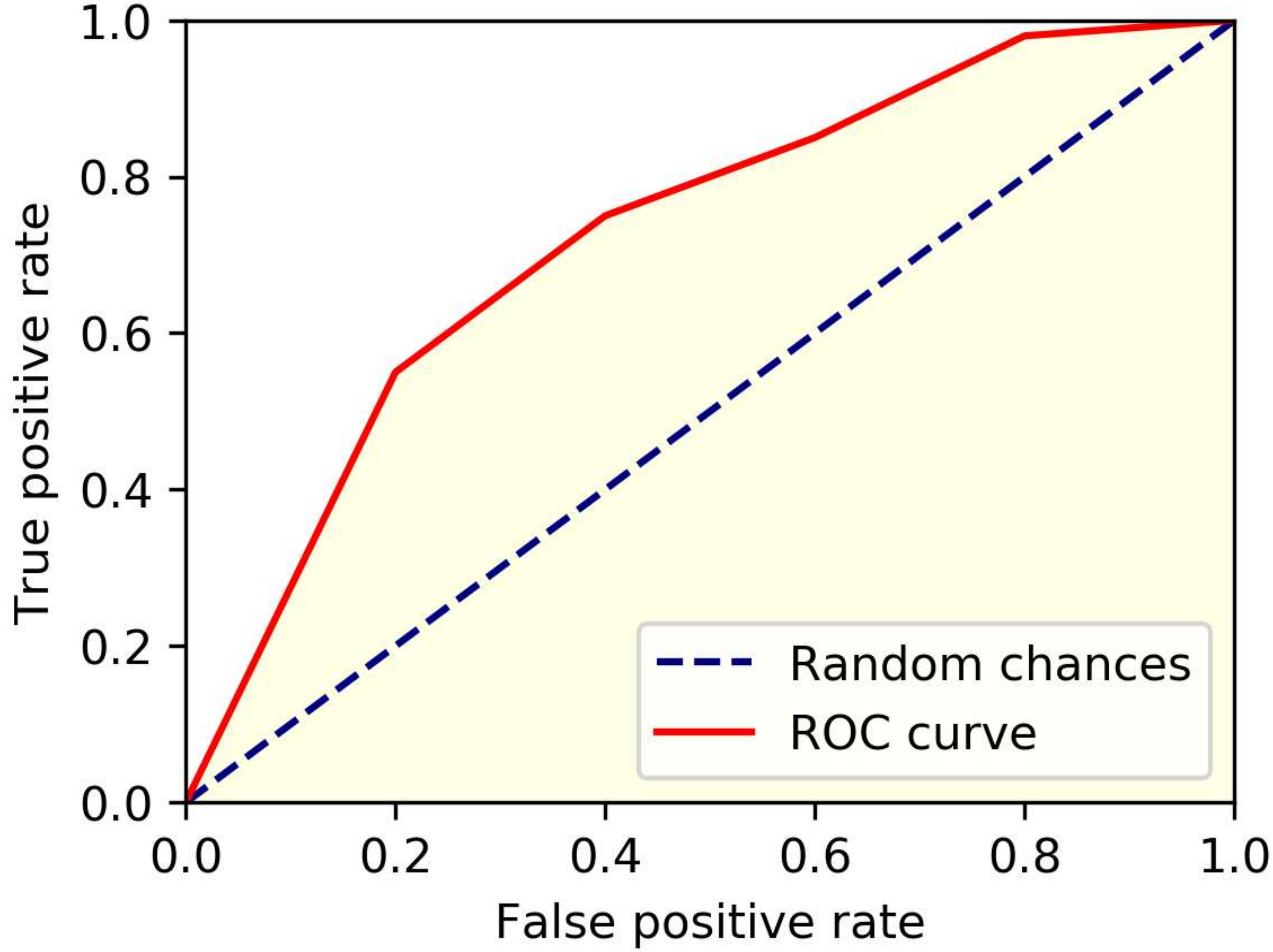
Name	Definition	Synonyms
Accuracy	(TP + TN) / (TP + TN + FP + FN)	
Misclassification	(FP + FN) / (TP + TN + FP + FN)	

Performance Checking: ROC Curve & AUC

- By default, an observation is assigned to class 1 if $p(y = 1|X = x) > .5$. That is, a threshold of 50% for the probability is used.

- The ROC (Receiver Operating Characteristics) curve is a popular graphic for simultaneously displaying two types of classification for all possible thresholds.
  - True Positive Rate vs. False Positive Rate

Performance Checking: ROC Curve & AUC



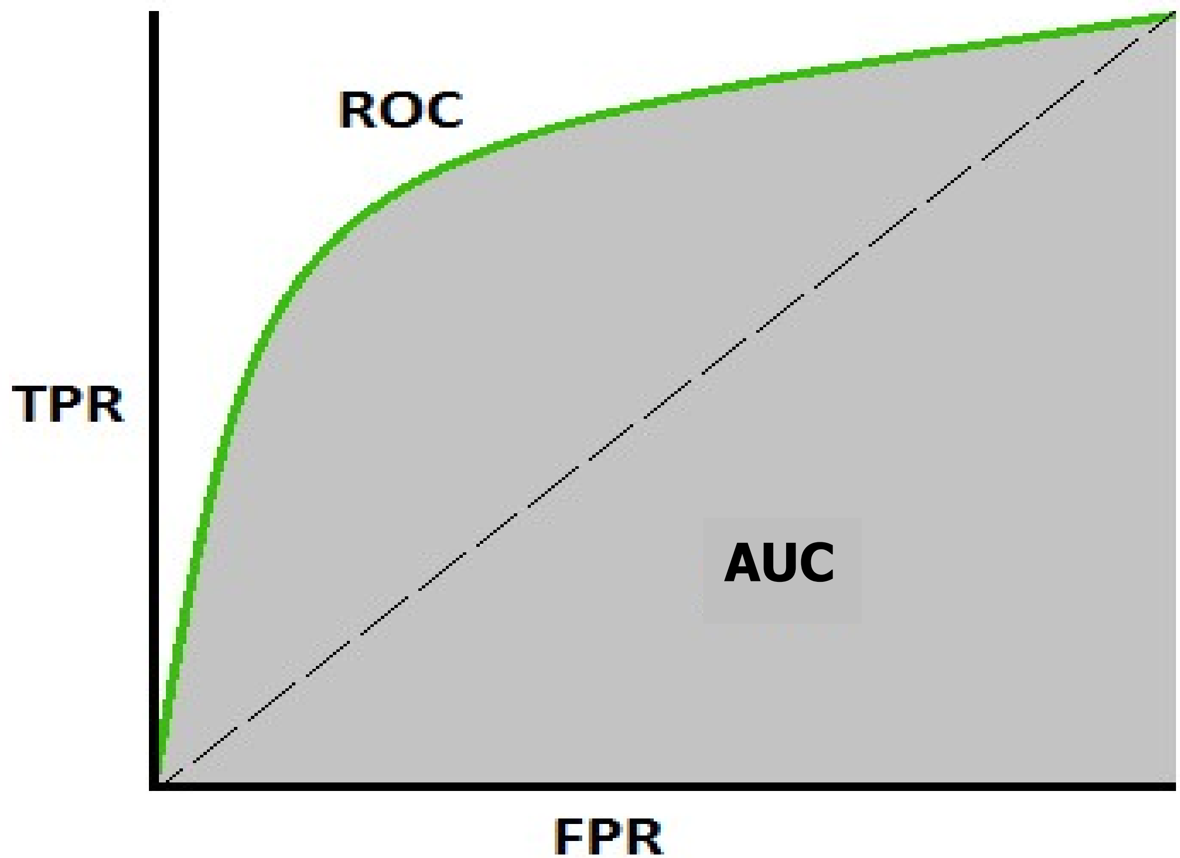
https://towardsdatascience.com/a-simple-explanation-of-the-roc-curve-and-auc-64db32d75541

Performance Checking: ROC Curve & AUC

- An ideal ROC curve will hug the top left corner, indicating a high true positive rate and a low false positive rate.

- The overall performance of a classifier, summarized over all possible thresholds, is given by the Area Under the Curve (AUC).
  - The larger the AUC, the better the classifier.
  - If a classifier's AUC = .5, it performs no better than chance.

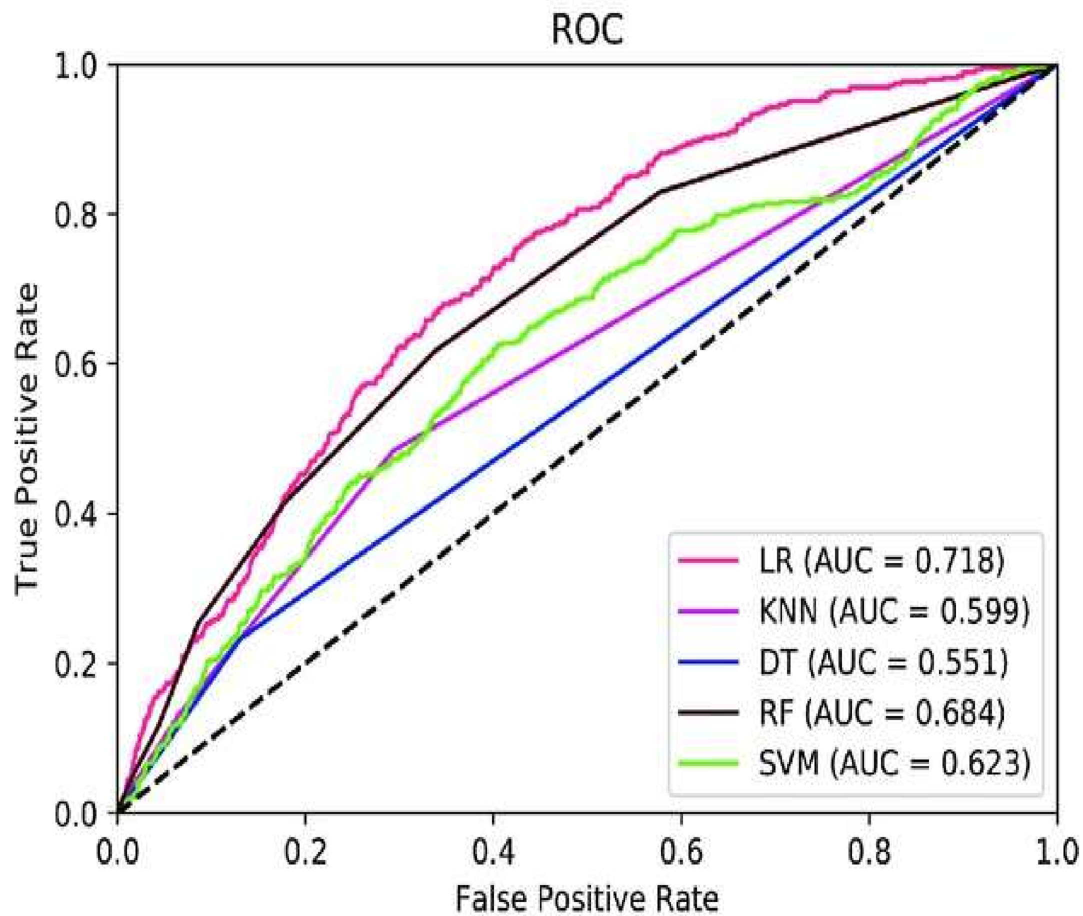
Performance Checking: ROC Curve & AUC



https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Performance Checking: ROC Curve & AUC

- ROC curves (and AUC values) are useful for comparing different classifiers.



Park H, Kim K. Comparisons among Machine Learning Models for the Prediction of Hypercholestrolemia Associated with Exposure to Lead, Mercury, and Cadmium. *International Journal of Environmental Research and Public Health*. 2019; 16(15):2666. https://doi.org/10.3390/ijerph16152666

Example: Logistic Regression

- The BookBinder Book Club data (BBB_training.csv & BBB_test.csv)
  - DV:
    - Buyer: Bought "Art History of Florence?"
  - Predictors:
    - Gender: 0 = male, 1 = female
    - Last : Months since last purchase
    - Book: Total $ spent on books
    - Art: # purchases of Art books
    - Child: # purchases of Children's books
    - Youth: # purchases of Youth books
    - Cook: # purchases of Cookbooks
    - Do_it: # purchases of Do-it-yourself books
    - Reference: # purchases of Reference books
    - Geog: # purchases of Geography books

Example: Logistic Regression

```
Call:
glm(formula = buyer ~ gender + last + book + art + child + youth +
    cook + do_it + reference + geog, family = binomial, data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5014  -0.4092  -0.2734  -0.1791   3.3182

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.342705   0.075834 -17.706  < 2e-16 ***
gender      -0.761378   0.050699 -15.018  < 2e-16 ***
last        -0.096848   0.003962 -24.443  < 2e-16 ***
book        -0.022195   0.012458  -1.782  0.07481 .
art          1.469313   0.157176   9.348  < 2e-16 ***
child        0.027970   0.124555   0.225  0.82232
youth        0.111025   0.124815   0.890  0.37372
cook        -0.031559   0.135510  -0.233  0.81584
do_it       -0.255634   0.143092  -1.787  0.07402 .
reference    0.479709   0.150708   3.183  0.00146 **
geog         0.916020   0.164505   5.568 2.57e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Example: Logistic Regression

Exp($\hat{\beta}$) = OR

		2.5 %	97.5 %
(Intercept)	0.2611383	0.2250304	0.3029367
gender	0.4670225	0.4228475	0.5158256
last	0.9076938	0.9006102	0.9147080
book	0.9780497	0.9544354	1.0022062
art	4.3462469	3.1958410	5.9182464
child	1.0283649	0.8055285	1.3126619
youth	1.1174233	0.8749569	1.4272501
cook	0.9689335	0.7428623	1.2636643
do_it	0.7744253	0.5850183	1.0251891
reference	1.6156037	1.2026832	2.1714549
geog	2.4993232	1.8109789	3.4514501

Example: Logistic Regression

Training sample

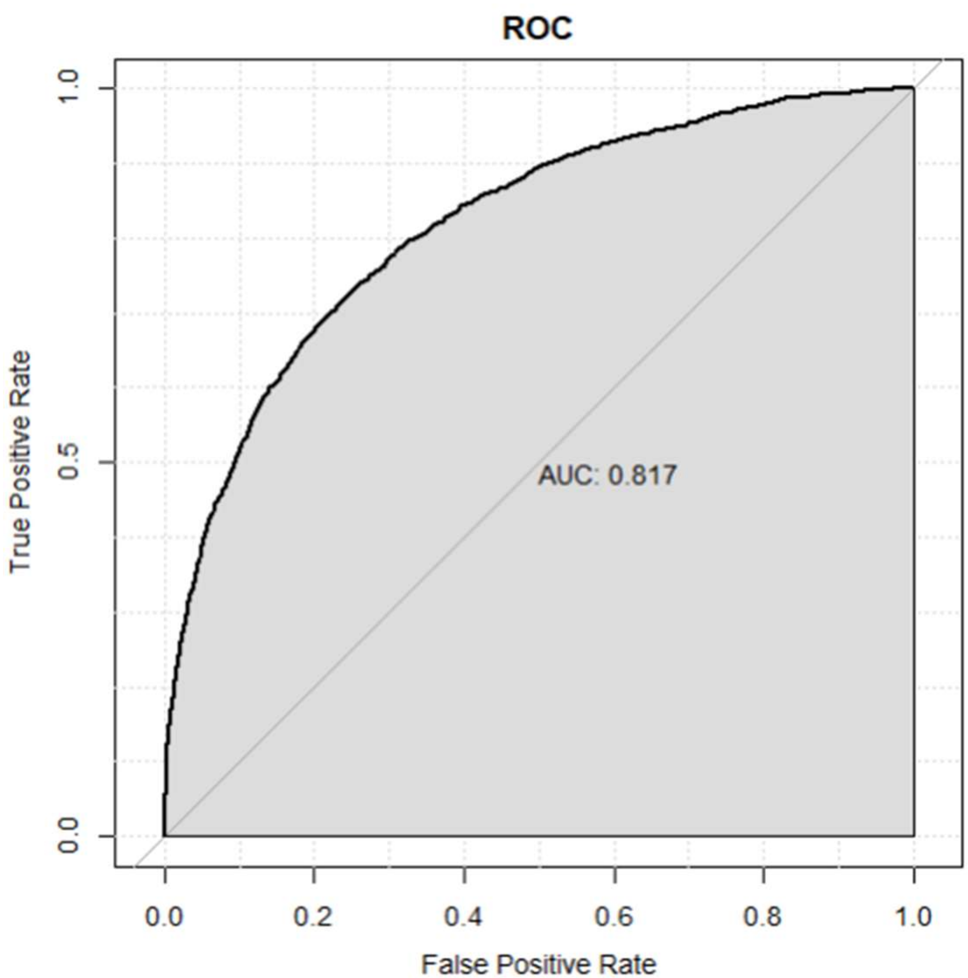
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 22736  1901
         1   190   368

               Accuracy : 0.917
                 95% CI : (0.9135, 0.9204)
    No Information Rate : 0.9099
    P-Value [Acc > NIR] : 3.894e-05

                  Kappa : 0.2331

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.16219
            Specificity : 0.99171
         Pos Pred Value : 0.65950
         Neg Pred Value : 0.92284
             Prevalence : 0.09006
         Detection Rate : 0.01461
   Detection Prevalence : 0.02215
      Balanced Accuracy : 0.57695

       'Positive' Class : 1
```

Test sample

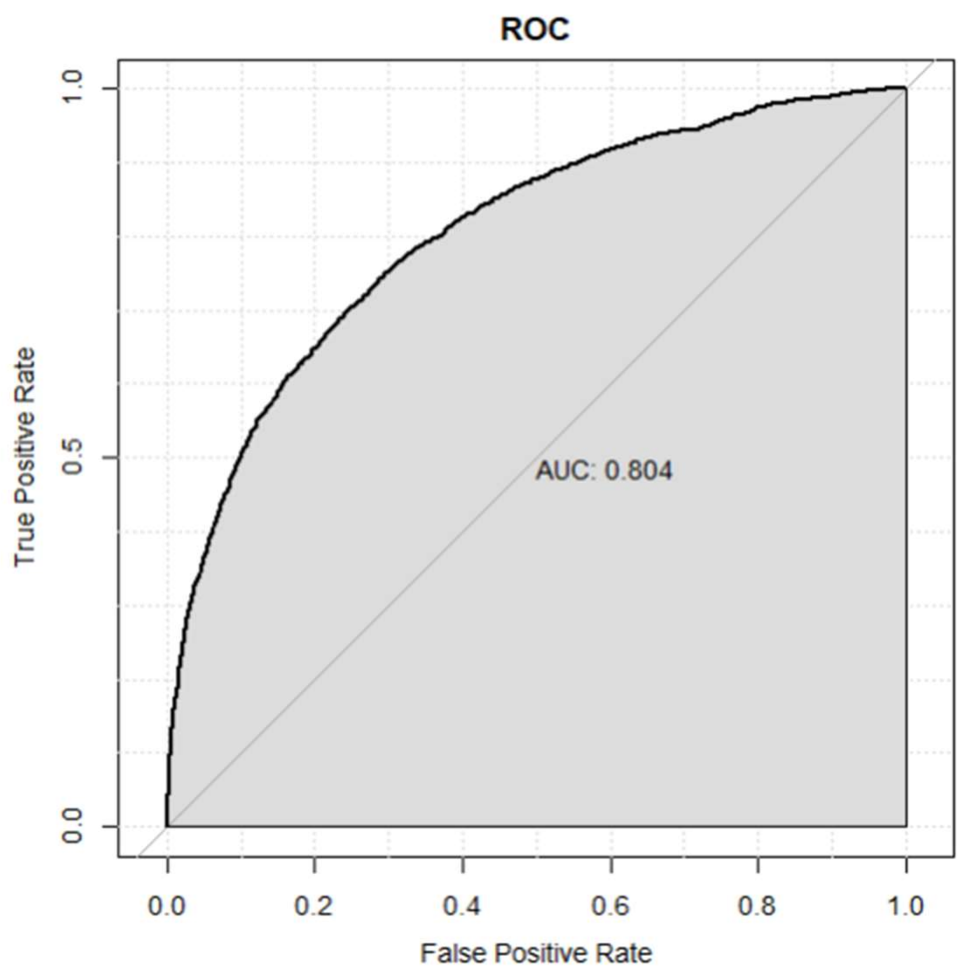
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 22365  1921
         1   187   332

               Accuracy : 0.915
                 95% CI : (0.9115, 0.9185)
    No Information Rate : 0.9092
    P-Value [Acc > NIR] : 0.0006381

                  Kappa : 0.2128

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.14736
            Specificity : 0.99171
         Pos Pred Value : 0.63969
         Neg Pred Value : 0.92090
             Prevalence : 0.09083
         Detection Rate : 0.01338
   Detection Prevalence : 0.02092
      Balanced Accuracy : 0.56953

       'Positive' Class : 1
```

Example: Logistic Regression

Training sample



Test sample



Logistic Regression for > 2 Classes

- We may need to classify a response variable that has more than two classes. For example,

$$Y \begin{cases} 1 = \text{stroke} \\ 2 = \text{drug overdose} \\ 3 = \text{epileptic seizure} \end{cases}$$

- It is straightforward to generalize (binary) logistic regression to more than two classes. This extension is known as multinomial logistic regression (or multiclass logistic regression).