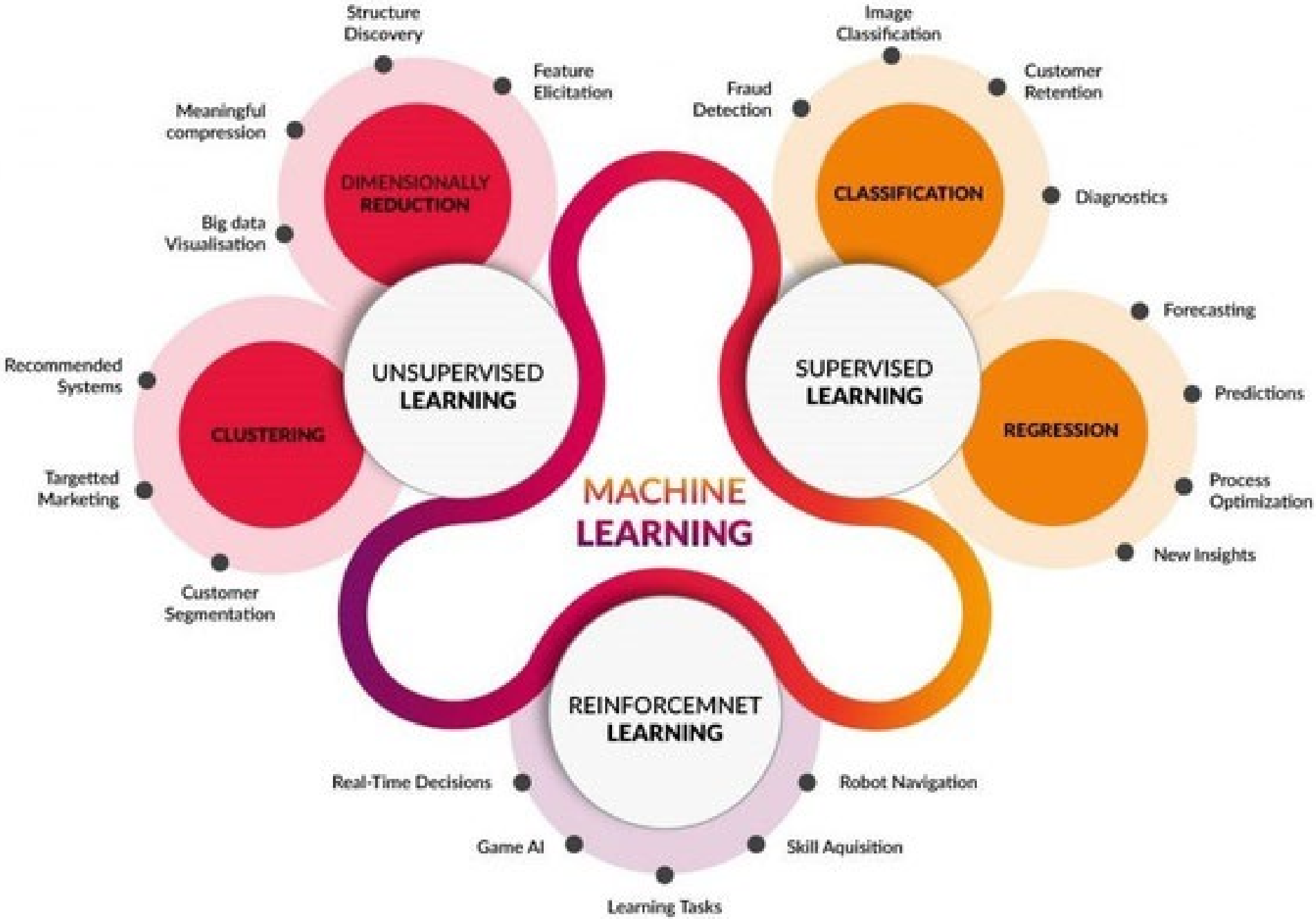

Structure Discovery
Feature Elicitation
Meaningful compression
Big data Visualisation
DIMENSIONALLY REDUCTION
Image Classification
Customer Retention
Fraud Detection
CLASSIFICATION
Diagnostics
Recommended Systems
Targetted Marketing
CLUSTERING
Customer Segmentation
UNSUPERVISED LEARNING
SUPERVISED LEARNING
Forecasting
Predictions
Process Optimization
New Insights
REGRESSION
MACHINE LEARNING
REINFORCEMNET LEARNING
Real-Time Decisions
Robot Navigation
Game AI
Skill Aquisition
Learning Tasks

Business Problems with Classification

Problem 1. MBA students placement

- Expected CTC? (MLR)
- Whether the student will be placed in first week or not?
- Whether the student will be placed or not at all?

- Based on
  - Educational background (Tech Vs Non-tech)
  - CGPA
  - First year GPA
  - CAT Score
  - Etc.

Problem 2. Travel company (yatra, MMT, Cleartrip)

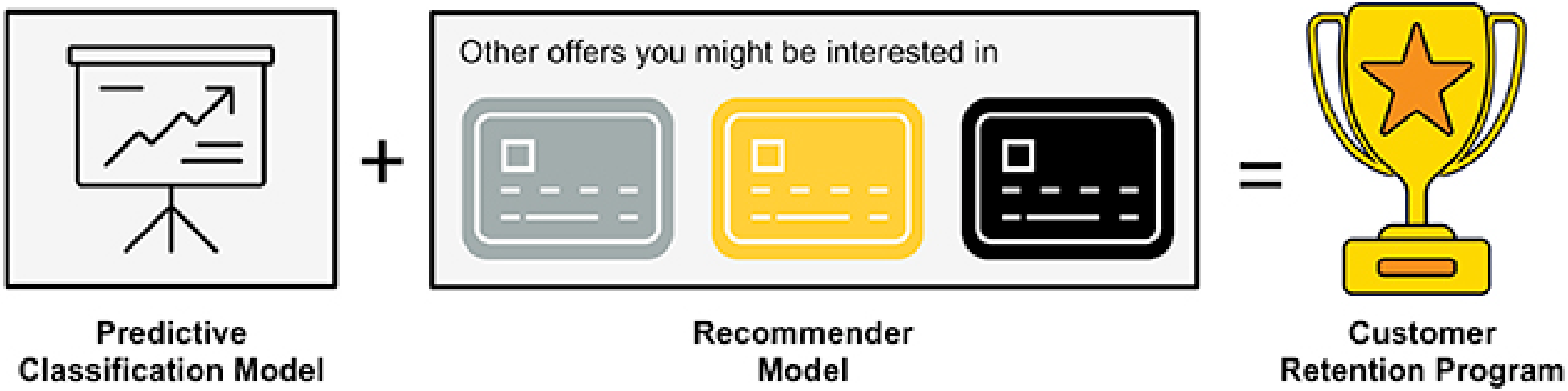
- Product:
  - School children exchange : **[Teacher organised parents paid]**
  - University international immersion program
  - Artistic destination trips or family trips
- Operations:
  - Visa, permits
  - Transfer points facilities, hotel, meals
  - Insurance, Safety and security

yatra

Objective of travel company

- To predict **whether the client will turn back or not**
- Targeted marketing & Promotional offer
- Minimising the cost of marketing

Customer retention

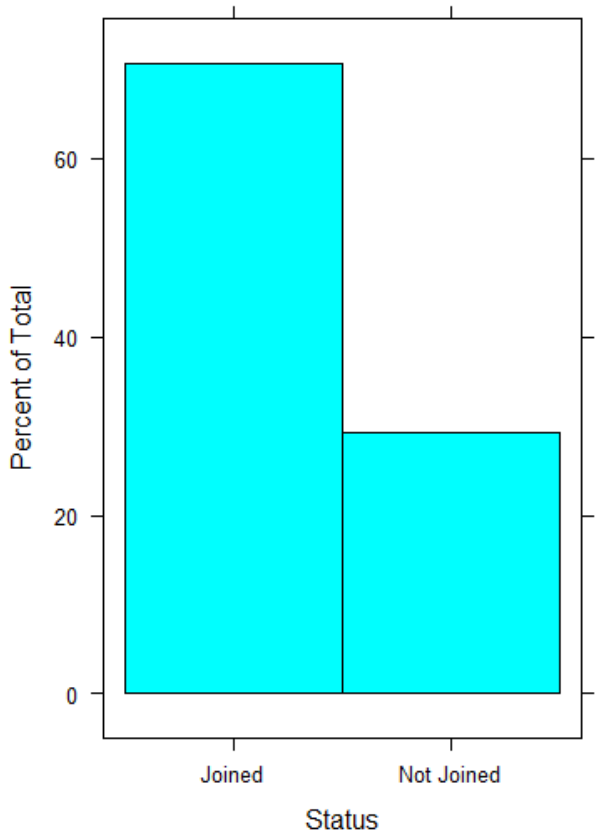


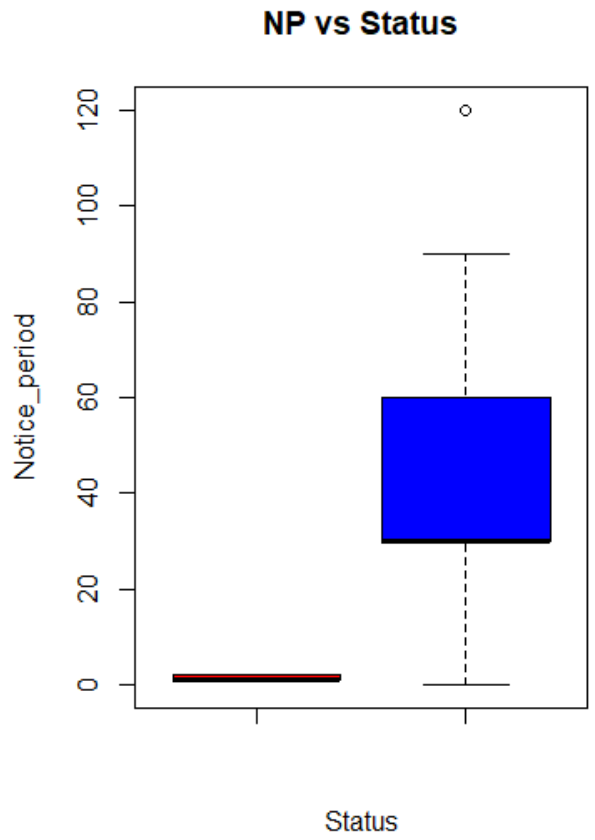
Problem 3. Human Resource uncertainty

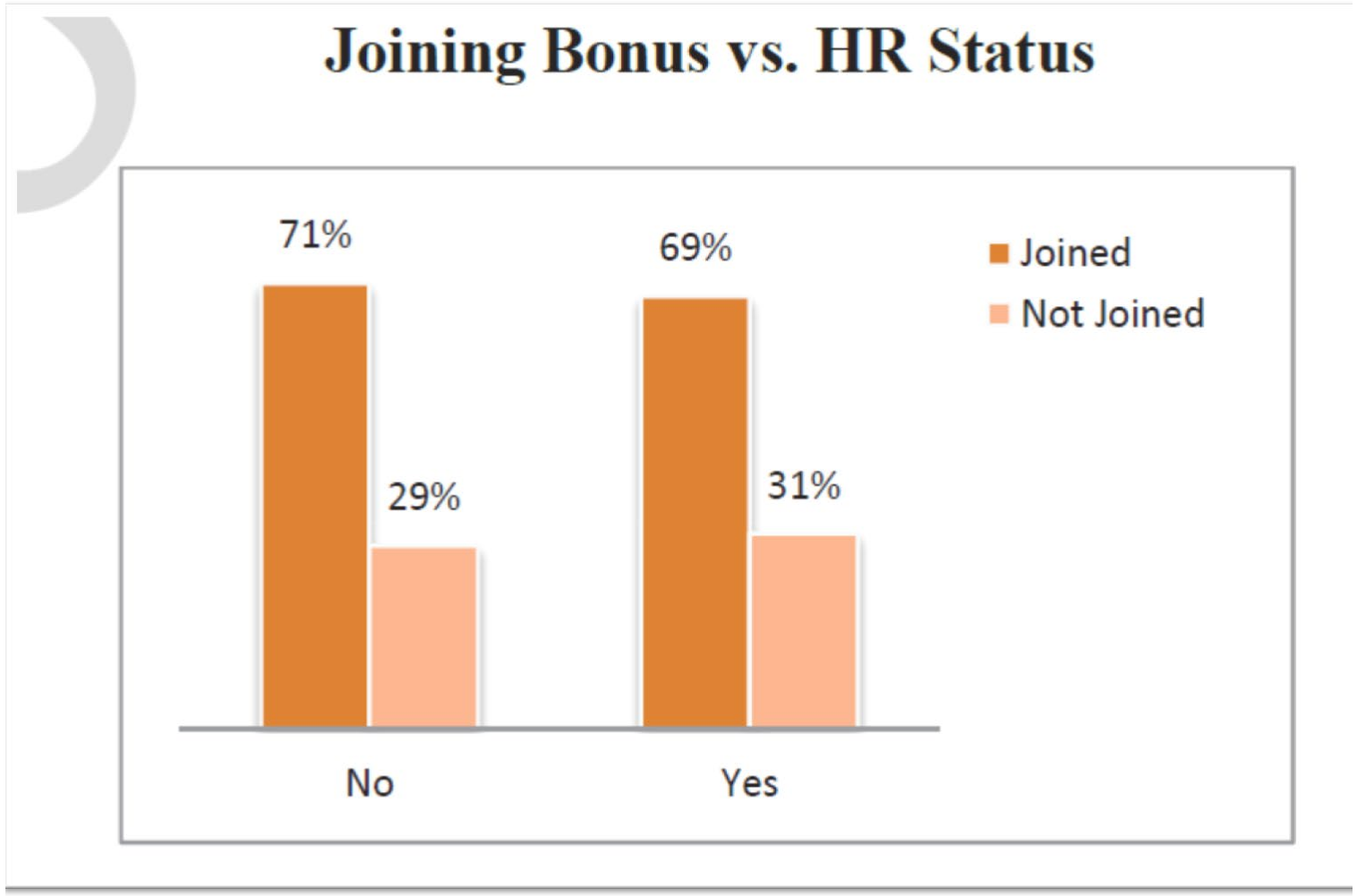
- When the expected compensation is less than X%
- Candidate educational background is of tier 1 institute
- When candidate is relocated from the on city to another
- Notice period
- Extension on joining date
- Joining bonus

Problem 3. Human Resource uncertainty

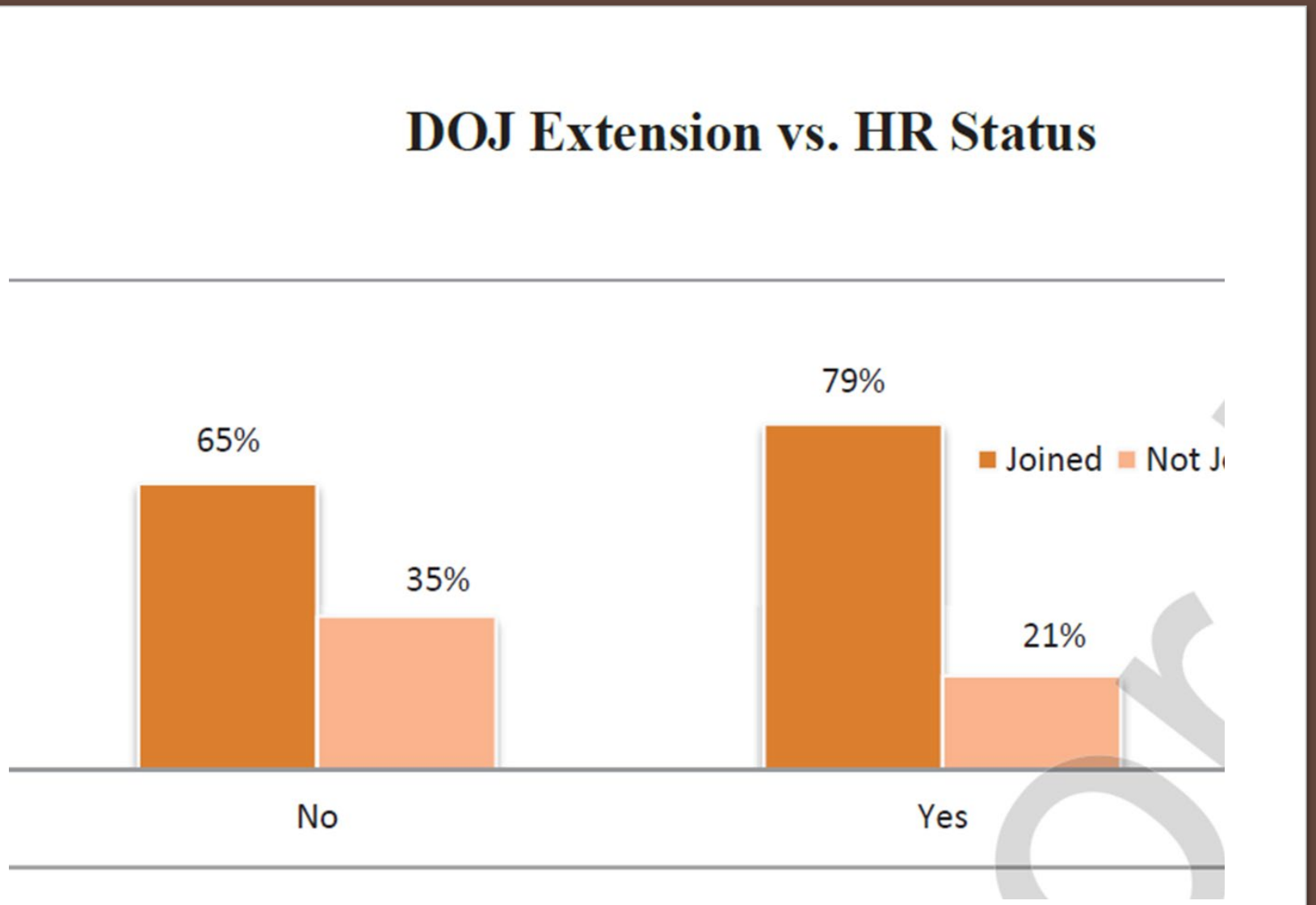
Predicting for candidates who are unlikely to join







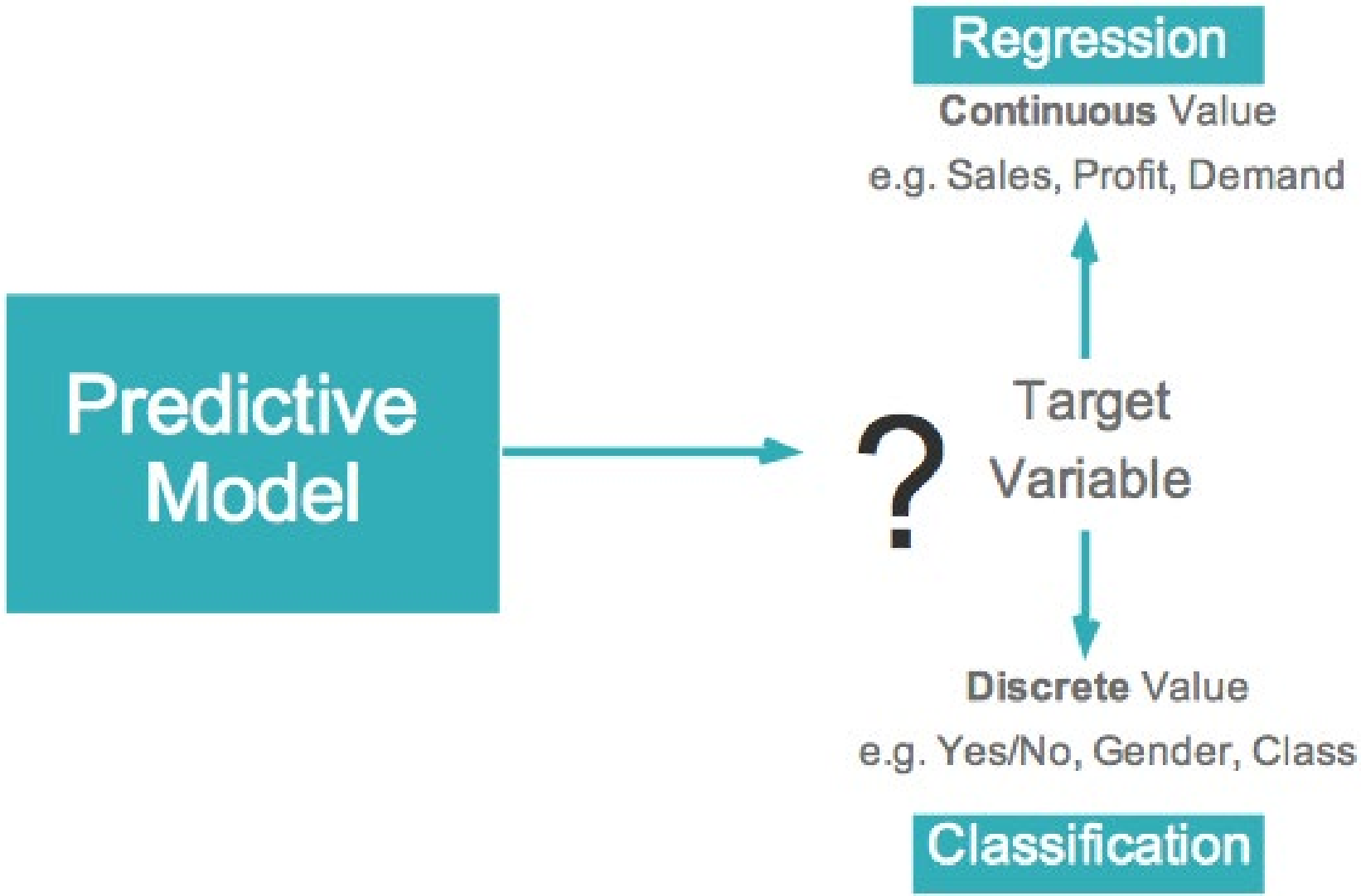
Can not say anything



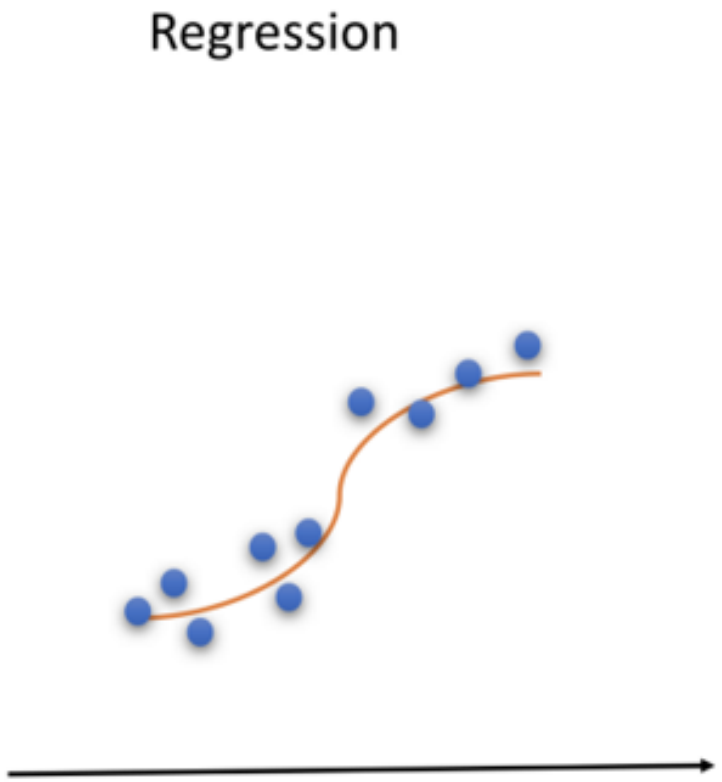
Likely to join with extension

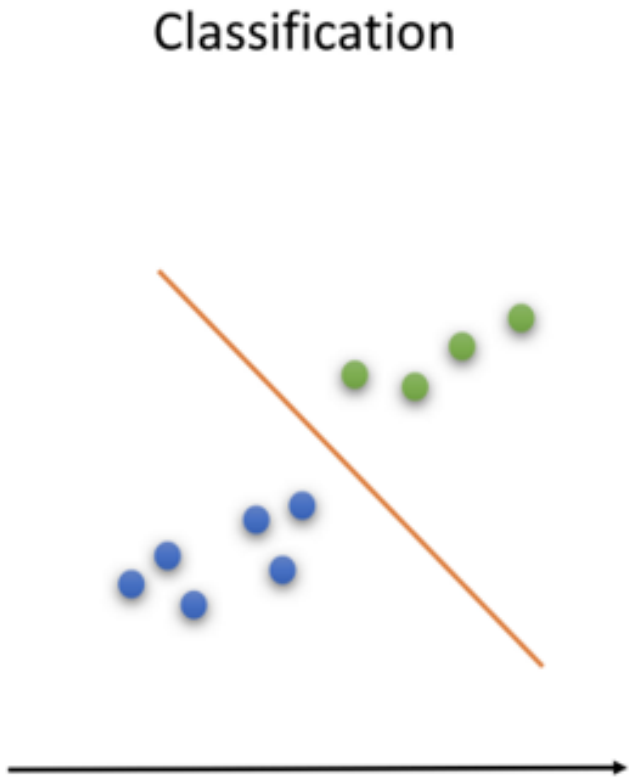
Problem 4. Default Data Set

- A set containing information on ten thousand customers.
- Variables
  - ✓Default: A factor with levels “No” and “Yes” indicating whether the customer defaulted on their debt.
  - ✓Student: A factor with levels “No” and “Yes” indicating whether the customer is a student.
  - ✓Balance: The average balance that the customer has remaining on their credit card after making their monthly payment.
  - ✓Income: Income of customer.
- Objective: We are interested in predicting whether an individual will default on his or her credit card payment.


Regression
Continuous Value
e.g. Sales, Profit, Demand
Predictive Model
?
Target Variable
Discrete Value
e.g. Yes/No, Gender, Class
Classification

Classification Problems in Data Science





Introduction

- The linear regression model assumes that the response variable *Y* is quantitative.
- In many situations, the response variable is instead *qualitative*. For example, eye colour is qualitative, taking qualitative on values blue, brown, or green.
- Often qualitative variables are referred to as *categorical* ; we will use these terms interchangeably.
- Here we study approaches for predicting qualitative responses, a process that is known as *classification*.

Introduction

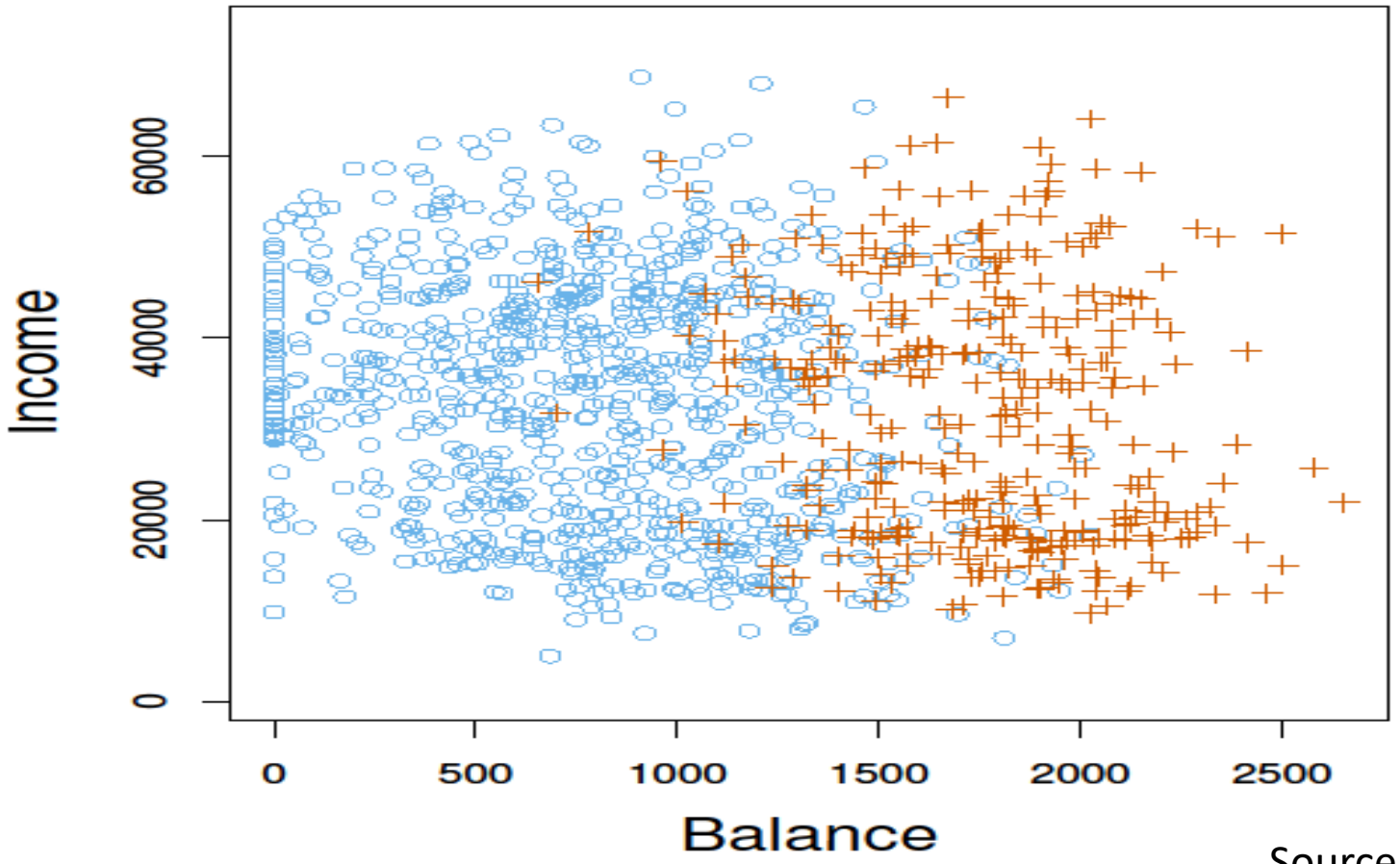
- Predicting a qualitative response for an observation can be referred to as *classifying* that observation, since it involves assigning the observation to a category, or class.
- The methods used for classification generally predict the probability of each of the categories of the qualitative variable, as the basis for making the classification. In this sense they also behave like regression methods.
- **Here we will use logistic regression in modelling a binary response variable.**

Classification Methods

Default Data Set

- A set containing information on ten thousand customers.
- Variables
  - ✓Default: A factor with levels “No” and “Yes” indicating whether the customer defaulted on their debt.
  - ✓Student: A factor with levels “No” and “Yes” indicating whether the customer is a student.
  - ✓Balance: The average balance that the customer has remaining on their credit card after making their monthly payment.
  - ✓Income: Income of customer.
- Objective: We are interested in predicting whether an individual will default on his or her credit card payment.

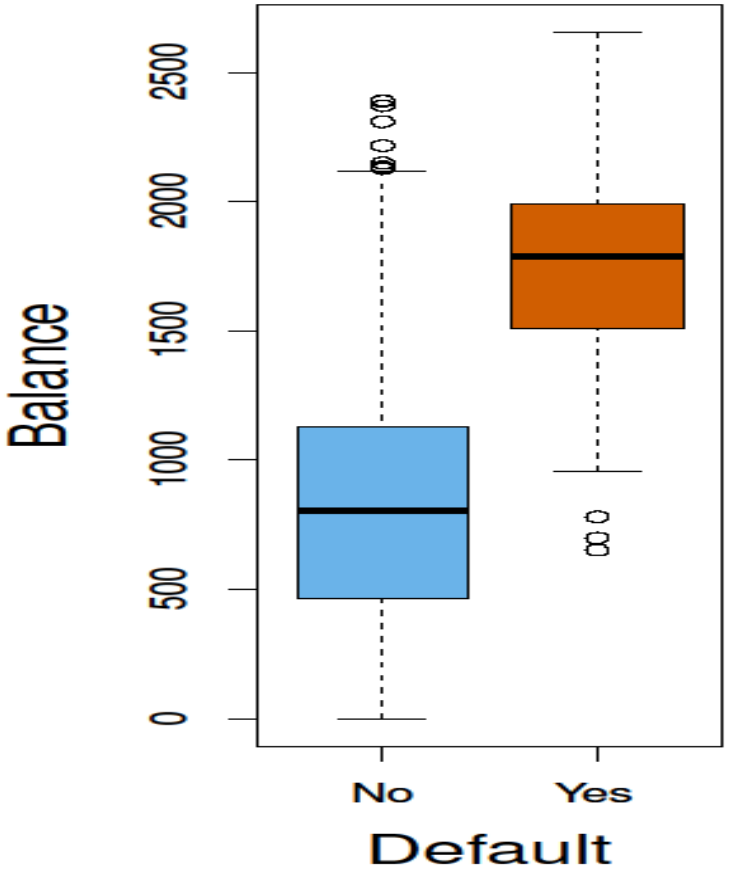
Default Data Set: Exploration Visualization

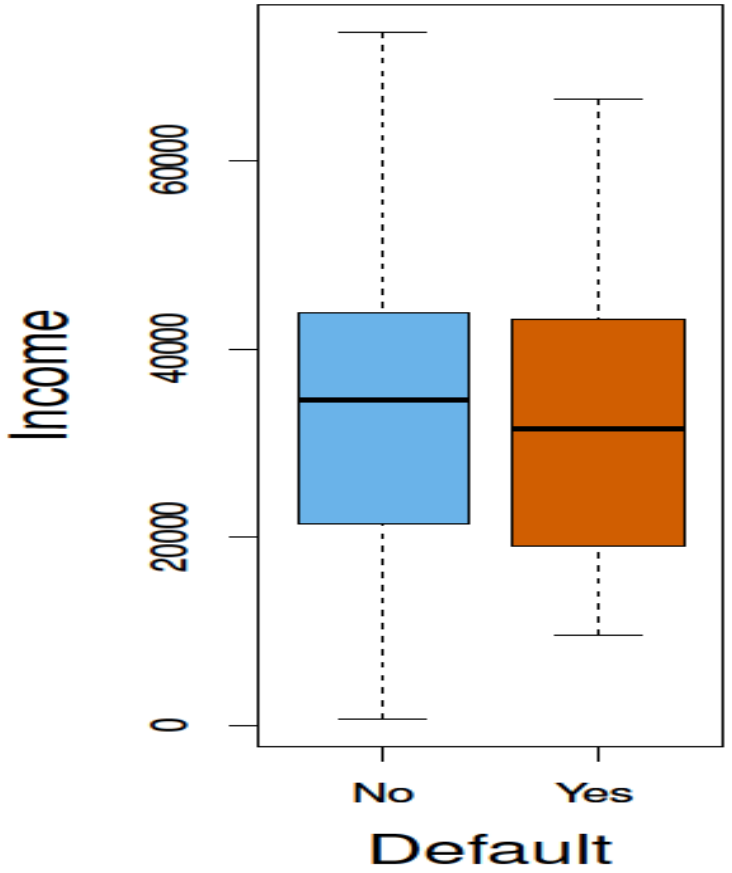


- Who defaulted on their credit card payments are shown in orange
- Who did not are shown in blue

Source: http://www-bcf.usc.edu/~gareth/ISL/

Default Data Set





- Distribution of balance & income split by the default variable
- It seems that there is very pronounced relationship between the predictor balance and the response default.
- However, the relationship between the predictor income and the response default seems to be weak.

Source: http://www-bcf.usc.edu/~gareth/ISL/

Why not Linear Regression?

- For the Default Data set, there are only two possibilities for the variable Default: *No* and *Yes*.
- We could then potentially use the *dummy variable* approach to code the response as follows:

$$Y = \begin{cases} 0 & \text{if Default} = \text{No}; \\ 1 & \text{if Default} = \text{Yes}. \end{cases}$$

- Suppose we have one predictor, i.e., “Balance”.

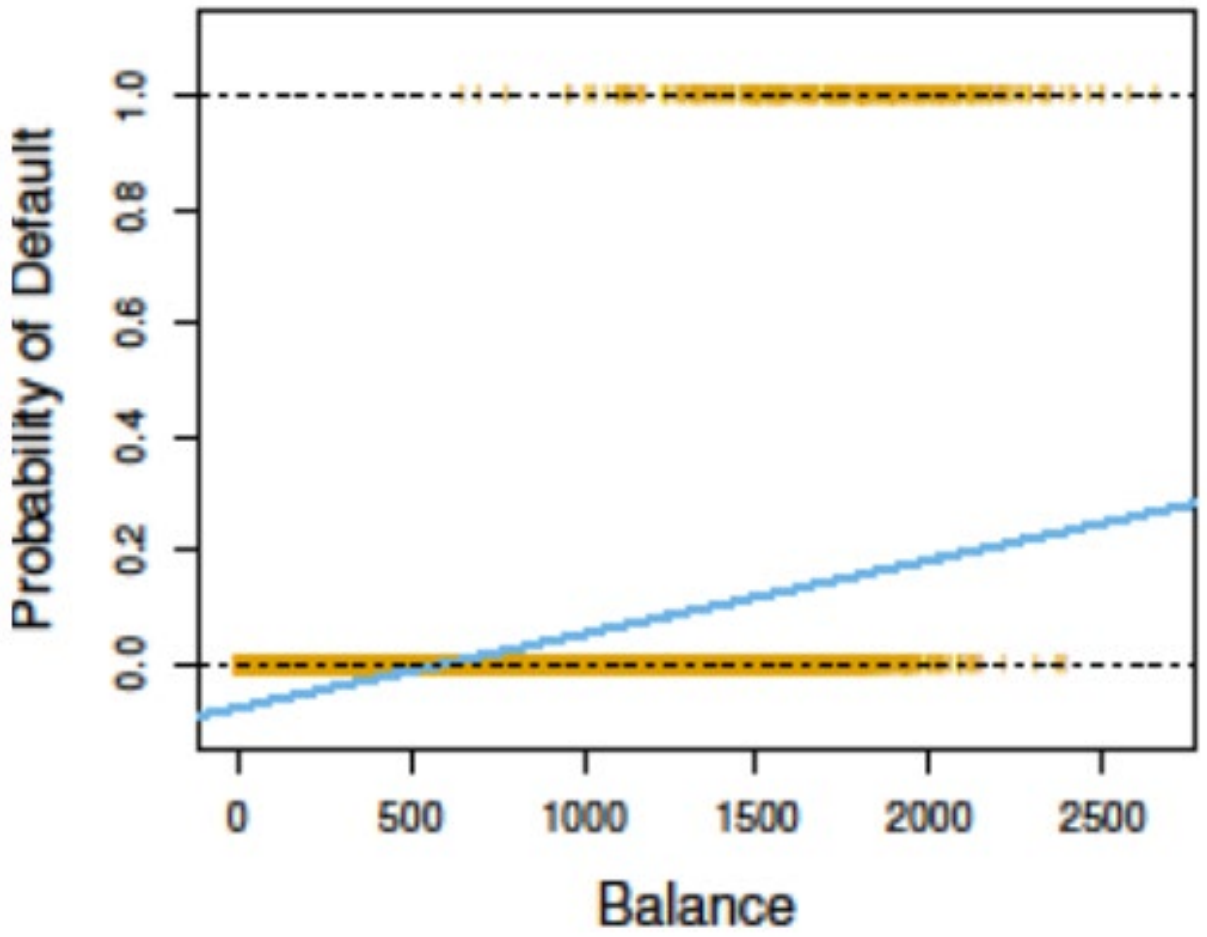
Why not Linear Regression?

- We could then fit a linear regression to this binary response with Balance as the predictor, and predict "Default=Yes" if $\hat{Y} > 0.5$ and "Default=No" otherwise.
- But, the possible values in the left-hand side are 0 and 1, and the predictions are certainly not 0 and 1!

$$\text{Y} = \beta_0 + \beta_1 X.$$

Problem with Linear Regression

- Source: http://www-bcf.usc.edu/~gareth/ISL/



Logistic Regression

Logistic Regression

- Instead of modelling this response $Default$ directly, logistic regression models the *probability* that *Default* belongs to a particular category.
- For the Default data, logistic regression models the probability of default.
- For example, the probability of default given balance can be written as

$$\Pr(\text{Default} = \text{Yes}|\text{Balance}).$$

- The values of $\Pr(\text{Default} = \text{Yes}|\text{Balance})$, which we abbreviate $p(\text{balance})$, will range between 0 and 1.

Logistic Regression

- Then for any given value of balance, a prediction can be made for default.
- For example, one might predict $default = Yes$ for any individual for whom $p(\text{balance}) > 0.5$.
- Alternatively, if a company wishes to be conservative in predicting individuals who are at risk for default, then they may choose to use a lower threshold, such as $p(\text{balance}) > 0.1$.

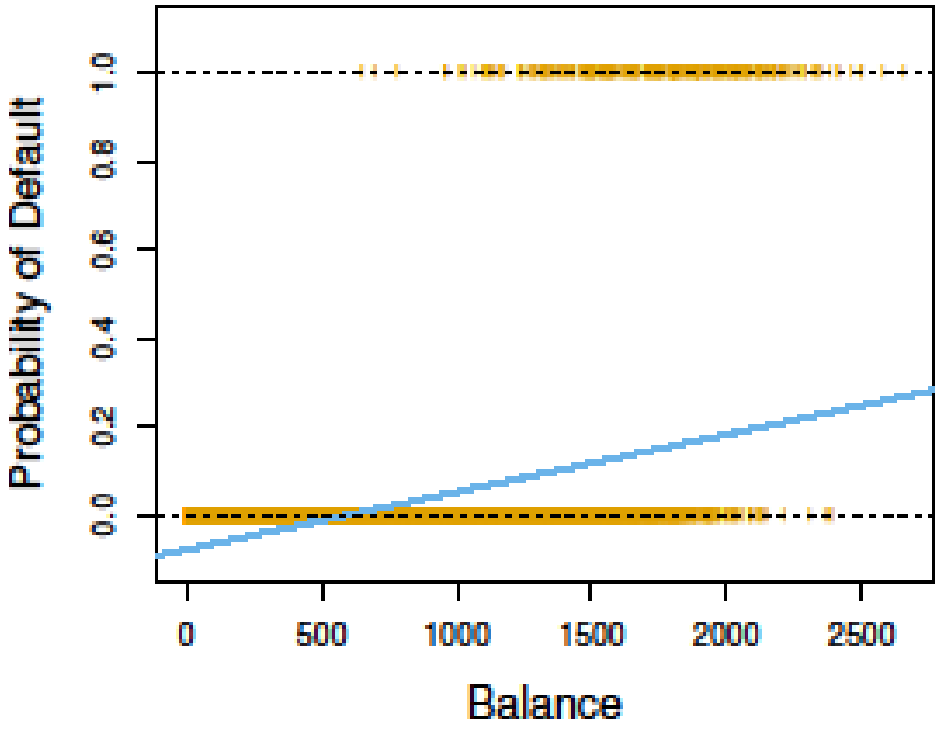
Logistic Model

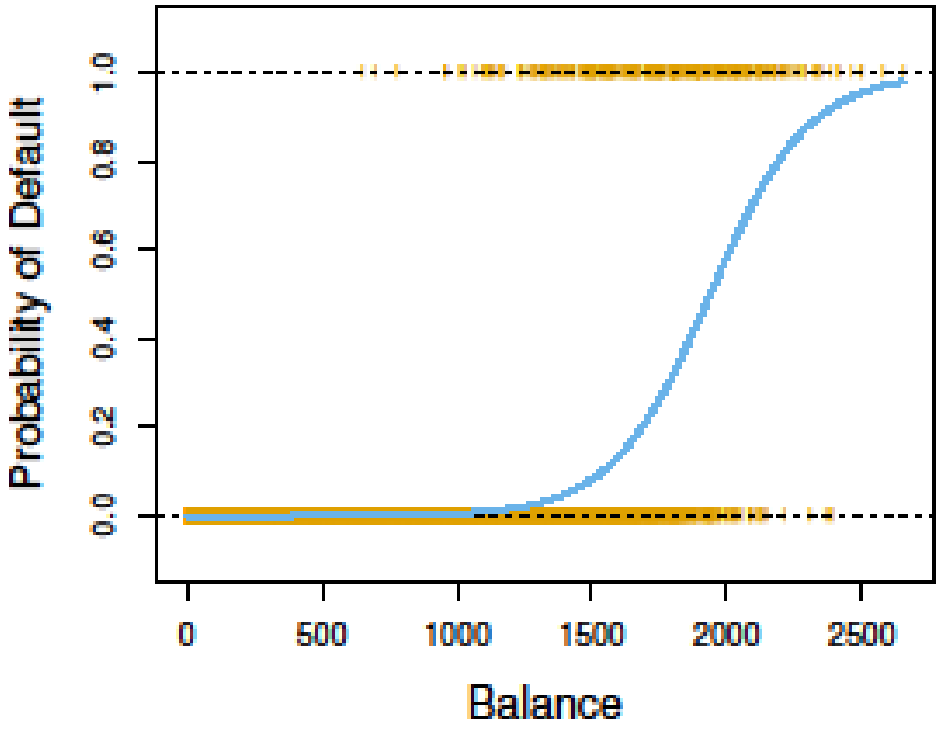
- How should we model the relationship between $p(X) = \Pr(\mathrm{Y} = 1|\mathrm{X})$ and X?
- For convenience we are using the generic 0/1 coding for the response.
- Previously, we talked of using a linear regression model to represent these probabilities:

$$p(X) = \beta_0 + \beta_1 X.$$

- If we use this approach to predict $default = yes$ using balance, then we obtain the model shown in the left-hand panel of the next figure.

Logistic Model





Source: http://www-bcf.usc.edu/~gareth/ISL/

Logistic Model

- Here we see the problem with this approach:
  - ✓for balances close to zero we predict a negative probability of default;
  - ✓if we were to predict for very large balances, we would get values bigger than 1.
- These predictions are not sensible, since of course the true probability of default, regardless of credit card balance, must fall between 0 and 1.

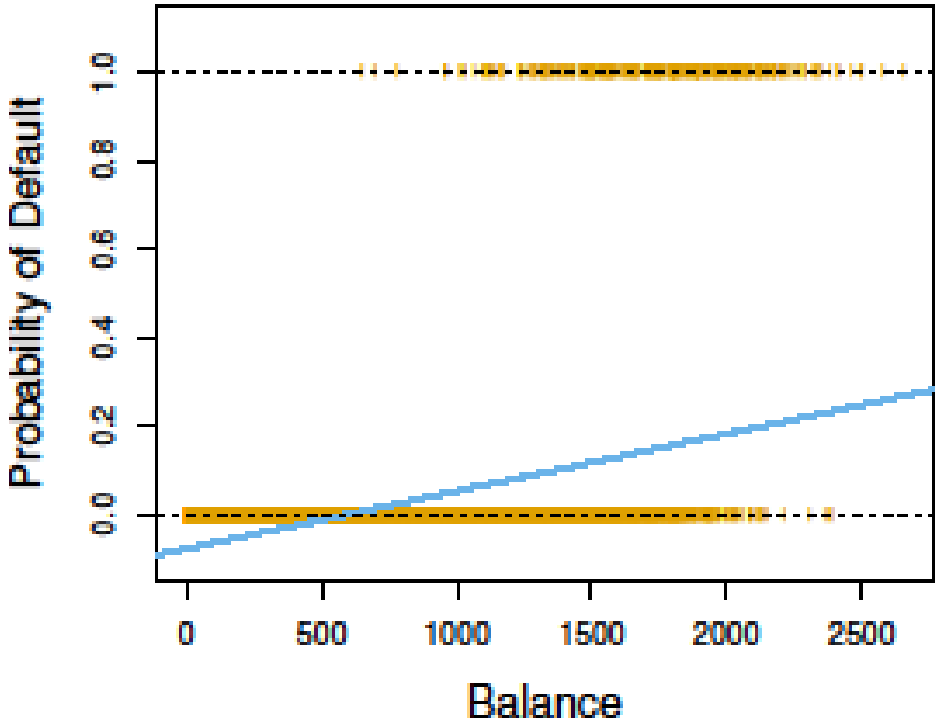
Logistic Model

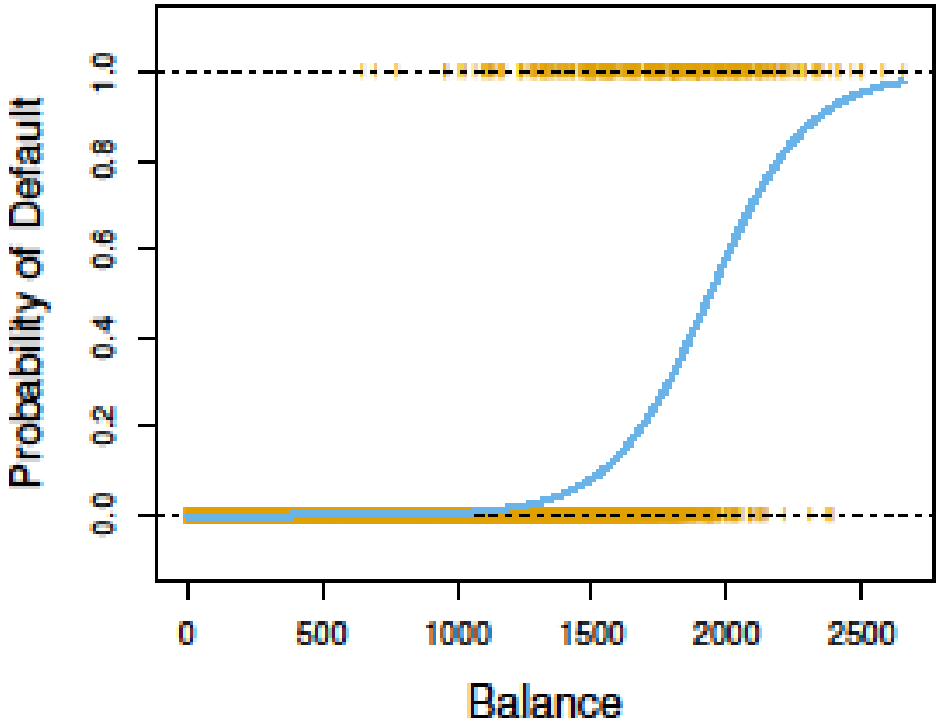
- To avoid this problem, we must model $p(X)$ using a function that gives outputs between 0 and 1 for all values of X.
- Many functions meet this description.
- In logistic regression, we use the *logistic function*,

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \dots\dots\dots\dots (1)$$

- β_0 and β_1 are the parameters of the model.

Logistic Model





Source: http://www-bcf.usc.edu/~gareth/ISL/

Logistic Model

- After a bit of manipulation of Equation (1), we find that

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X} \ldots\ldots\ldots (2)$$

- The quantity $p(X)/[1-p(X)]$ is called the *odds*, and can take on any value odds between 0 and ∞.

Odds

- Values of the odds close to 0 and ∞ indicate very low and very high probabilities of default, respectively.
- For example, on average 1 in 5 people with an odds of 1/4 will default, since $p(X) = 0.2$ implies an odds of $\frac{0.2}{(1-0.2)} = 1/4$.
- Likewise on average nine out of every ten people with an odds of 9 will default, since $p(X) = 0.9$ implies an odds of $0.9/(1 - 0.9) = 9$.
- Odds are traditionally used instead of probabilities in horse-racing, since they relate more naturally to the correct betting strategy.

Logistic Model

- Alternatively, (2) can be written as

$$\ln\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \dots\dots\dots (3).$$

- The left-hand side is called the *log-odds* or *logit*.
- We see that the logistic regression model (1) has a logit that is linear in *X*.

Interpretation of the parameters to make decision

Interpretation of the Parameters

- In a linear regression model, the slope parameters β_1 gives the average change in Y associated with a one-unit increase in X.

- But, in a logistic regression model, increasing X by one unit changes the log odds by β_1 (from Equation (3)).

- Or equivalently, it multiplies the odds by e^{β_1} (from Equation (2)).

Interpretation of the Parameters

- Note that the relationship between $p(X)$ and X in (in Equation (1)) is not a straight line.
- As a result, β_1 does not correspond to the change in $p(X)$ associated with a one-unit increase in X.
- The amount that $p(X)$ changes due to a one-unit change in X will depend on the current value of X.
- But regardless of the value of X, if β_1 is positive then increasing X will be associated with increasing $p(X)$, and if β_1 is negative then increasing X will be associated with decreasing $p(X)$.

Estimating the Parameters

- The coefficients β_0 and β_1 in Equation(1) are unknown, and must be estimated based on the available training data.
- We use method of *maximum likelihood* for this purpose.
- The basic intuition behind using maximum likelihood to fit a logistic regression model is as follows.
- We seek estimates for β_0 and β_1 such that the predicted probability $\hat{p}(x_i)$ of default for each individual, using Equation (1), corresponds as closely as possible to the individual's observed default status.

Estimating the Parameters

- In other words, we try to find β_0 and β_1 such that plugging these estimates into the model for $p(X)$, given in (1), yields a number close to one for all individuals who defaulted, and a number close to zero for all individuals who did not.
- This intuition can be formalized using a mathematical equation called a *likelihood function*:

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})).$$

- The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to *maximize* this likelihood function.

Default Data Set: Estimated Regression Coefficients

	Coefficient	Std. error	Z-statistic	P-value	**Odds Ratio**
Intercept	−10.6513	0.3612	−29.5	<0.0001	
balance	0.0055	0.0002	24.9	<0.0001	1.005

Making Predictions for Decision

Making Predictions

- Once the coefficients have been estimated, it is a simple matter to compute the probability of default for any given credit card balance.
- For example, using the coefficient estimates given in the Table, we predict that the default probability for an individual with a balance of $1, 000 is

$$\hat{p}(X) = \frac{e^{\widehat{\beta_0}+\widehat{\beta_1}x}}{1+e^{\widehat{\beta_0}+\widehat{\beta_1}x}} = 0.00576.$$

- This is below 1%.
- In contrast, the predicted probability of default for an individual with a balance of $2, 000 is much higher, and equals 0.586 or 58.6%.

Qualitative Predictors

- One can use qualitative predictors with the logistic regression model using the dummy variable approach.
- As an example, the Default data set contains the qualitative variable student.
- To fit the model we simply create a dummy variable that takes on a value of 1 for students and 0 for non-students.
- The logistic regression model that results from predicting probability of default from student status can be seen in the next Table.

Qualitative Predictors

	Coefficient	Std. error	z-statistic	p-value	**Odds Ratio**
Intercept	−3.5041	0.0707	−49.55	<0.0001	
Student[Yes]	0.4049	0.1150	3.52	0.0004	1.499

Interpreting the Coefficients

- The coefficient associated with the dummy variable is positive, and the associated *p-value* is statistically significant.
- This indicates that students tend to have higher default probabilities than non-students:

$$\Pr(\text{default} = \text{Yes}|\text{student} = \text{Yes}) = \frac{e^{-3.5041+0.4049\times1}}{1+e^{-3.5041+0.4049\times1}} = 0.0431,$$

$$\Pr(\text{default} = \text{Yes}|\text{student} = \text{No}) = \frac{e^{-3.5041+0.4049\times0}}{1+e^{-3.5041+0.4049\times0}} = 0.0292.$$

- Also, students are approximately 1.5 times more likely to default as compared to those who are not students.

Multiple Logistic Regression

Multiple Logistic Regression

- We now consider the problem of predicting a binary response using multiple predictors.
- By analogy with the extension from simple to multiple linear regression, we can generalize Equation (3) as follows:

$$\ln\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \ldots\ldots\ldots (4)$$

where $X = \left(X_1, X_2, \ldots, X_p\right)$ are p predictors.

Multiple Logistic Regression

- Equation (4) can be rewritten as

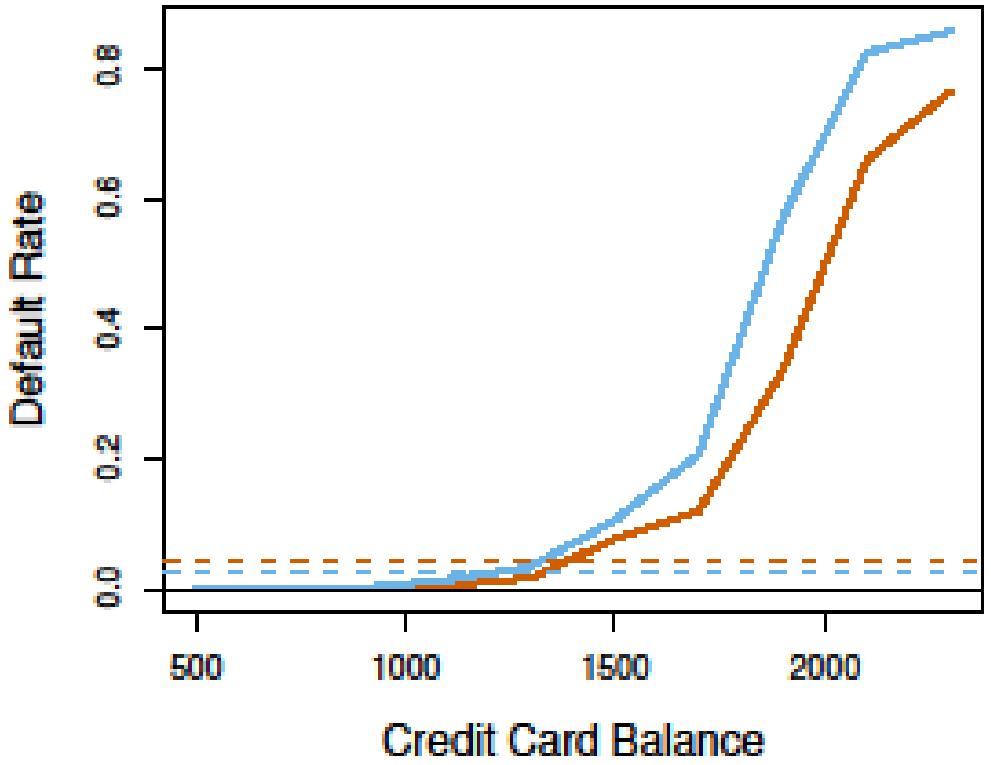
$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \ldots\ldots\ldots (5)$$

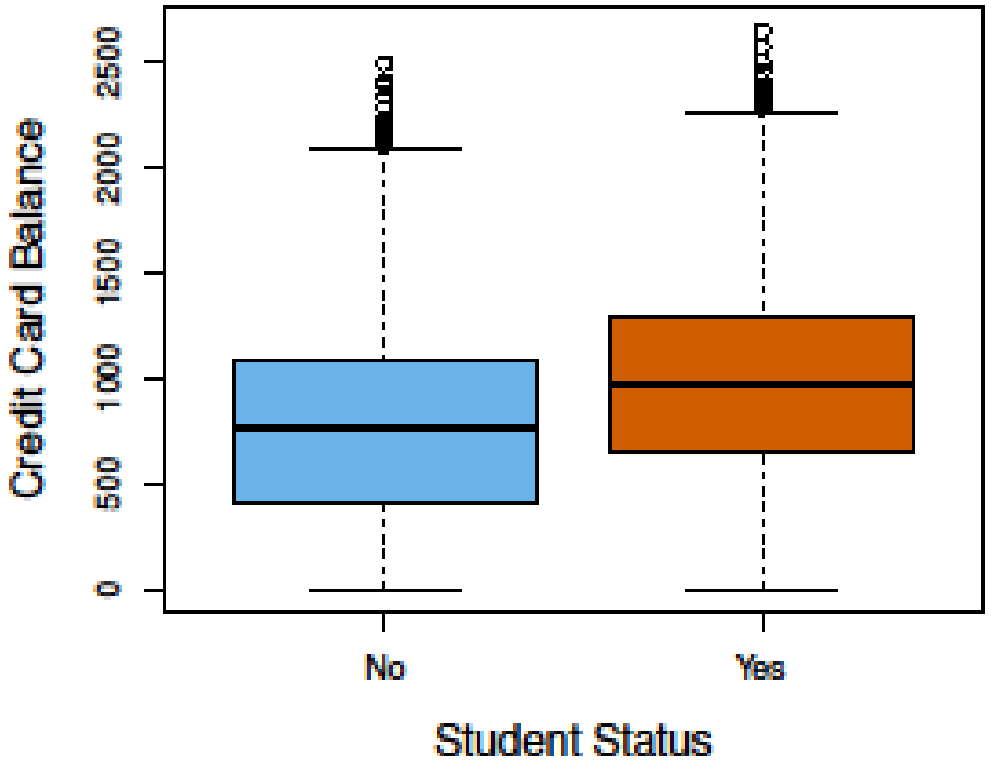
- As before, we use the maximum likelihood method to estimate the parameters $\beta_0, \ldots, \beta_p$.
- The next Table shows the coefficient estimates for a logistic regression model that uses balance, income (in thousands of dollars), and student status to predict probability of default.

Multiple Logistic Regression: Coefficient Table

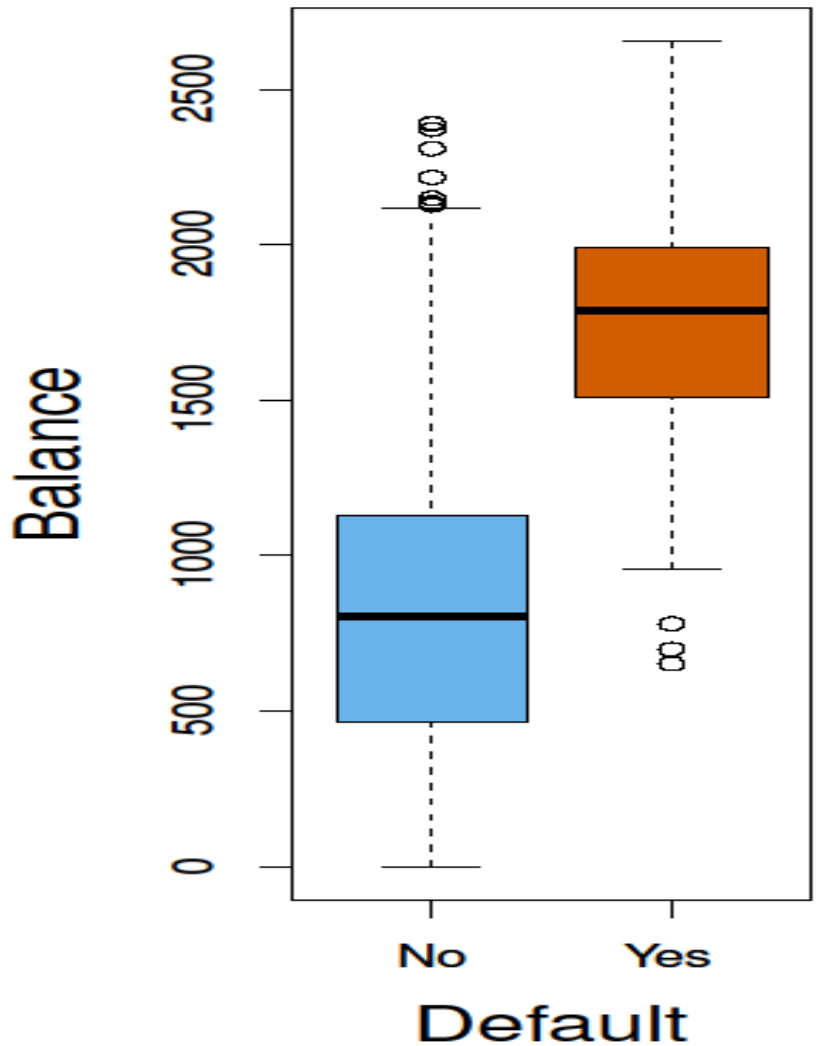
	Coefficient	Std. error	z-statistic	p-value	**Odds Ratio**
Intercept	−10.8690	0.4923	−22.08	<0.0001	
balance	0.0057	0.0002	24.74	<0.0001	1.0002
income	0.0000	0.0000	0.37	0.7115	1.0000
student[Yes]	−0.6468	0.2362	−2.74	0.0062	0.5237

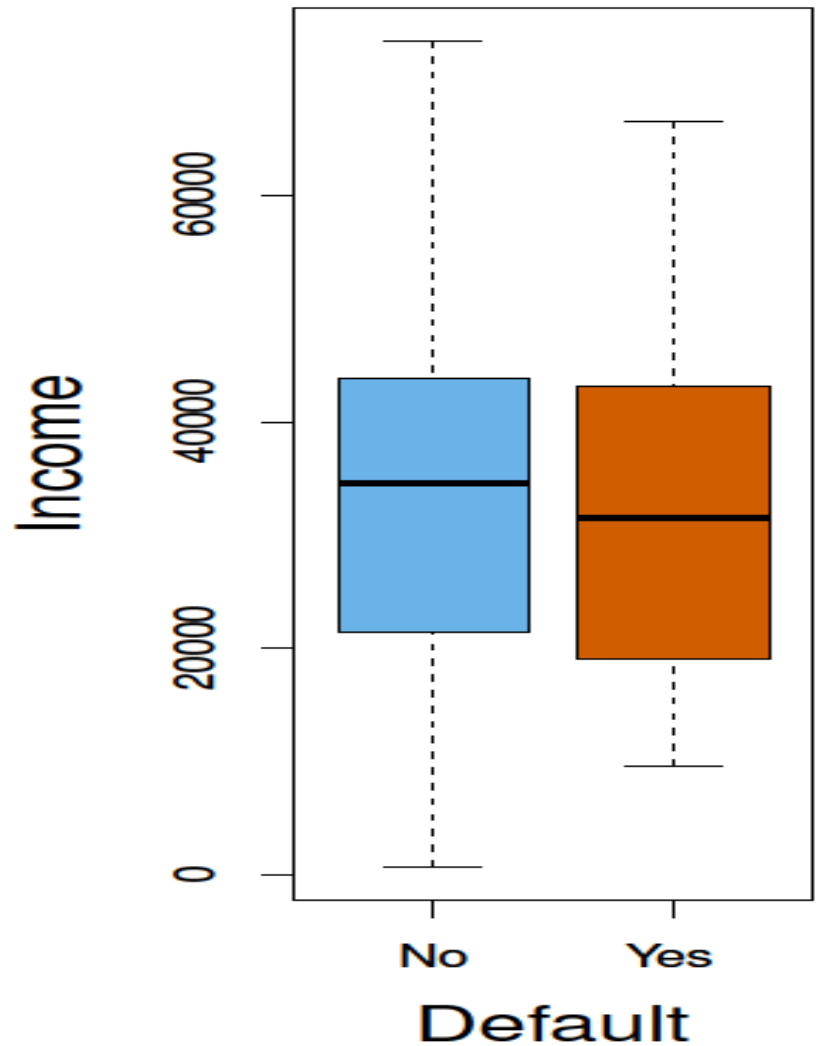
Multiple Logistic Regression: A paradox!!!





Source: http://www-bcf.usc.edu/~gareth/ISL/


Balance
0
500
1000
1500
2000
2500
No
Yes
Default


Income
0
20000
40000
60000
No
Yes
Default

Multiple Logistic Regression: A paradox!!!

- The right-hand panel of the previous Figure provides an explanation for this discrepancy.
- The variables student and balance are correlated.
- Students tend to hold higher levels of debt, which in turn is associated with higher probability of default.
- That is, students are more likely to have large credit card balances, which, from the left-hand panel of the Figure, tend to be associated with high default rates.

Multiple Logistic Regression: A paradox!!!

- This is an important point for a credit card company that is trying to determine to whom they should offer credit.
- A student is riskier than a non-student if no information about the student's credit card balance is available.
- However, that student is less risky than a non-student *with the same credit card balance*!
- In general, the phenomenon seen in the previous Figure is known as *confounding*.

Making Prediction

- By substituting estimates for the regression coefficients from the Table into Equation (5), we can make predictions.
- For example, a student with a credit card balance of $1,500 and an income of $40, 000 has an estimated probability of default of

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times1500+0.003\times40-0.6468\times1}}{1+e^{-10.869+0.00574\times1500+0.003\times40-0.6468\times1}} = 0.058.$$

- A non-student with the same balance and income has an estimated probability of default of

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times1500+0.003\times40-0.6468\times0}}{1+e^{-10.869+0.00574\times1500+0.003\times40-0.6468\times0}} = 0.105.$$

Logistic Regression Classifier

- In this case, we have assigned an observation to the *default* class if

$$\Pr(\text{default} = \text{Yes}|\text{X} = \text{x}) > 0.5 \ldots \ldots \ldots (6)$$

- Thus, the logistic regression here has used a threshold of 50% for the probability of default in order to assign an observation to the *default* class.
- However, if we are concerned about incorrectly predicting the default status for individuals who default, then we may consider lowering this threshold.

Training Error: Points to Remember!!

- First of all, training error rates will usually be lower than test error rates, which are the real quantity of interest.
- In other words, we might expect this classifier to perform worse if we use it to predict whether or not a new set of individuals will default.
- The reason is that we specifically adjust the parameters of our model to do well on the training data.
- The higher the ratio of parameters p to number of samples n, the more we expect this *overfitting* to play a role.

Training Error: Points to Remember!!

- Second, only 3.33% of the individuals in the training sample defaulted.
- A simple but useless classifier that always predicts that each individual will not default, regardless of his or her credit card balance and student status, will result in an error rate of 3.33%.
- In other words, the trivial *null* classifier will achieve an error rate that is only a bit higher than the logistic regression training set error rate.

Confusion Matrix

- In practice, a binary classifier such as logistic regression can make two types of errors.
- It can incorrectly assign an individual who defaults to the *no default* category, or it can incorrectly assign an individual who does not default to the *default* category.
- It is often of interest to determine which of these two types of errors are being made.
- A *confusion matrix*, shown for the Default data in the next Table, is a convenient way to display this information.

Confusion Matrix

	True default status			
Predicted Default Status		No	Yes	Total
	No	9627	228	9855
	Yes	40	105	145
	Total	9667	333	10000

$$Sensitivity = \frac{105}{333} = 31.53\%$$

$$Specificity = \frac{9627}{9667} = 99.59\%$$

$$Total\ Error\ Rate = \frac{268}{10000} = 2.68\%$$

Confusion Matrix

- The table reveals that the logistic regression predicted that a total of 145 people would default.
- Of these people, 105 actually defaulted and 40 did not.
- Hence only 40 out of 9, 667 of the individuals who did not default were incorrectly labelled.
- This looks like a pretty low error rate!

Confusion Matrix

- However, of the 333 individuals who defaulted, 228 (or 68.47%) were missed by Logistic Regression.
- So while the overall error rate is low, the error rate among individuals who defaulted is very high.
- From the perspective of a credit card company that is trying to identify high-risk individuals, an error rate of 228/333 = 68.47% among individuals who default may well be unacceptable.

Sensitivity and Specificity

- Class-specific performance is also important in medicine and biology, where the terms *sensitivity* and *specificity* characterize the performance of a classifier or screening test.
- In this case the sensitivity is the percentage of true defaulters that are identified, a low 31.53% in this case.
- The specificity is the percentage of non-defaulters that are correctly identified, here $(1 - 40/9,667) \times 100 = 99.59\%$.

Improving Logistic Regression Classifier

- The credit card company may be more interested to avoid incorrectly classifying an individual who will default, whereas incorrectly classifying an individual who will not default, though still to be avoided, is possibly less problematic.
- We now to modify logistic regression classifier in order to develop a classifier that better meets the credit card company's needs.

Improving Logistic Regression Classifier

- In this case, we have assigned an observation to the *default* class if

$$\Pr(\text{default} = \text{Yes}|\text{X} = \text{x}) > 0.5 \ldots \ldots \ldots (6)$$

- Thus, the logistic regression here has used a threshold of 50% for the probability of default in order to assign an observation to the *default* class.
- However, if we are concerned about incorrectly predicting the default status for individuals who default, then we may consider lowering this threshold.

Improving Logistic Regression Classifier

- For example, we might label any customer with a probability of default above 20% to the *default* class.
- That is, we may assign an observation to *default* class if

$$\Pr(\text{default} = \text{Yes}|\text{X} = \text{x}) > 0.2 \ldots \ldots \ldots (7)$$

Confusion Matrix

	True default status			
Predicted Default Status		No	Yes	Total
	No	9390	130	9520
	Yes	277	203	480
	Total	9667	333	10000

$$Sensitivity = \frac{203}{333} = 60.96\%$$

$$Specificity = \frac{9390}{9667} = 97.13\%$$

$$Total\ Error\ Rate = \frac{407}{10000} = 4.07\%$$

Improving Logistic Regression Classifier

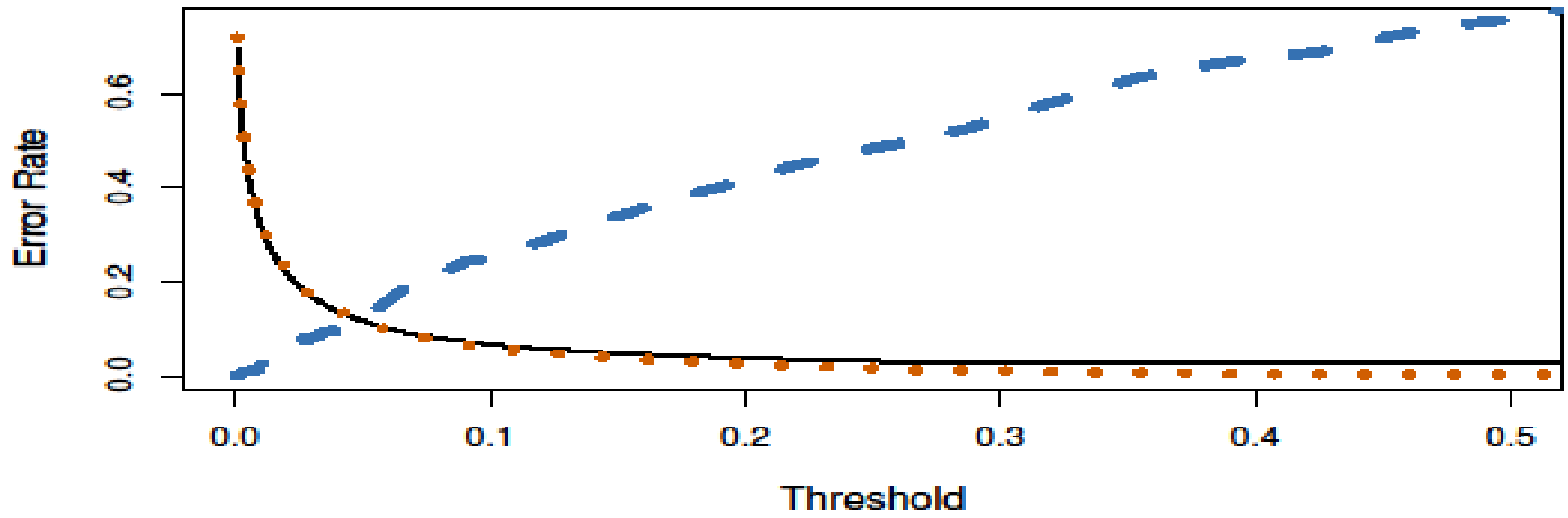
- The error rates that result from taking this approach are shown in the previous Table.
- Now logistic regression predicts that 480 individuals will default.
- Of the 333 individuals who default, logistic regression correctly predicts all but 130, i.e., 39.04%.
- This is a vast improvement over the error rate of 68.47% that resulted from using the threshold of 50%.

Improving Logistic Regression Classifier

- However, this improvement comes at a cost.
- Now 277 individuals who do not default are incorrectly classified.
- As a result, the overall error rate has increased slightly to 4.07 %.
- But this is a small price to be paid by a credit card company for more accurate identification of individuals who do indeed default.

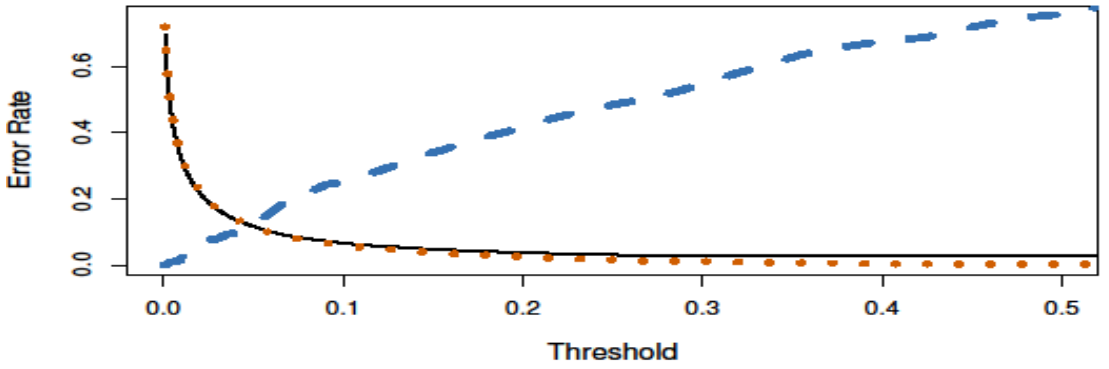
Optimal value of Classifier

Error Rate Plot



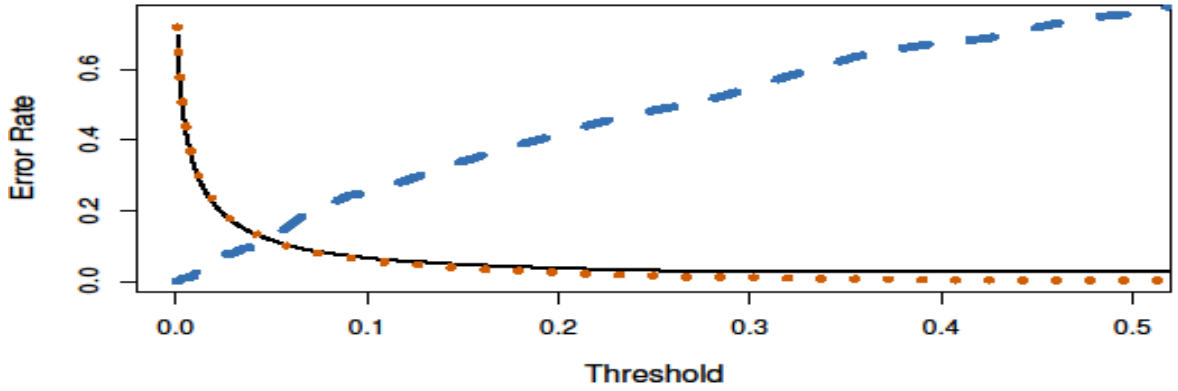
Source: http://www-bcf.usc.edu/~gareth/ISL/

Error Rate Plot



- The previous Figure illustrates various error rates as a function of the threshold value.
- *The black solid line displays the overall error rate.*
- *The blue dashed line represents the fraction of defaulting customers that are incorrectly classified.*
- *The orange dotted line indicates the fraction of errors among the non-defaulting customers.*

Error Rate Plot

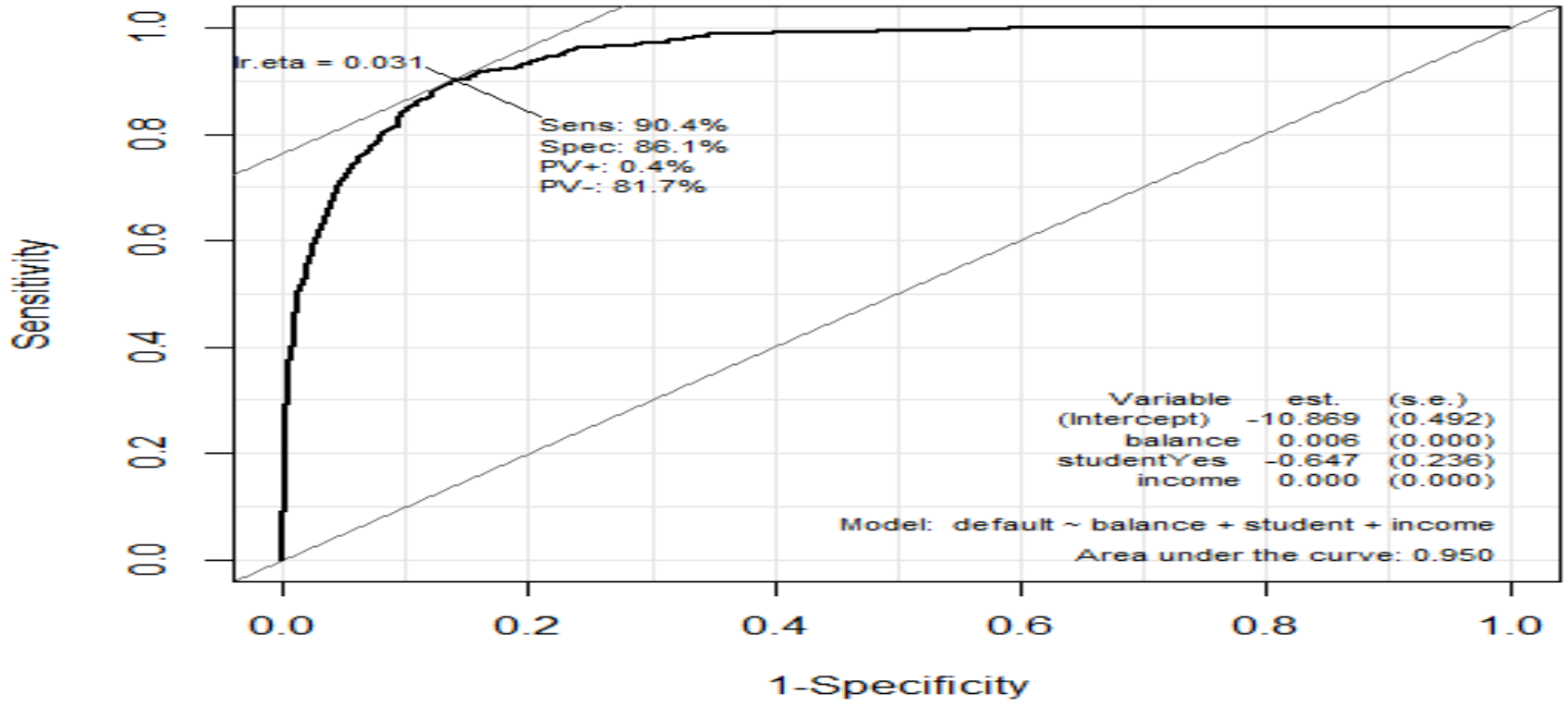


- Using a threshold of 0.5 minimizes the overall error rate (black solid line).
- When a threshold of 0.5 is used, the error rate among the individuals who default is quite high (blue dashed line).
- As the threshold is reduced, the error rate among individuals who default decreases steadily.
- But the error rate among the individuals who do not default increases (orange dotted line).

Deciding the Optimal Threshold

- How can we decide which threshold value is best?
- Such a decision generally depends on *domain knowledge*, such as detailed information about the costs associated with defaulting.
- However, some common approaches involve maximizing the Sensitivity or Specificity.
- Another approach available is to maximize the $Youden\ Index\ (Sens.+Spec.-1)$ should be closer to 1. (see http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2749250/).
- The R library "Epi" determines the threshold by maximizing the sum of specificity and sensitivity.

Deciding the Optimal Threshold



Confusion Matrix using Optimal Threshold

	True default status			
Predicted Default Status		No	Yes	Total
	No	8318	32	8350
	Yes	1349	301	1650
	Total	9667	333	10000

$$Sensitivity = \frac{301}{333} = 90.39\%$$

$$Specificity = \frac{8318}{9667} = 86.05\%$$

$$Total\ Error\ Rate = \frac{1381}{10000} = 13.81\%$$

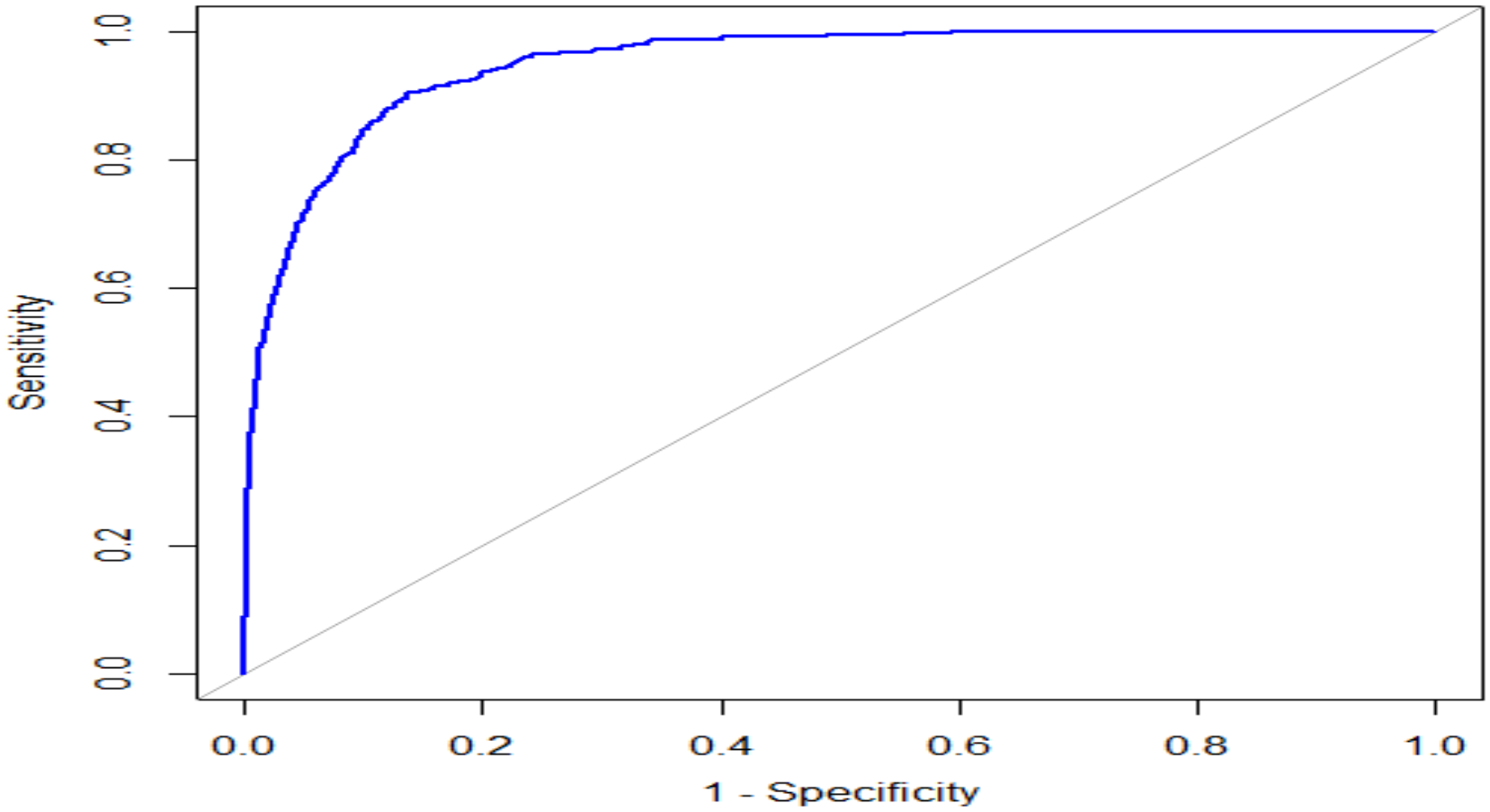
ROC Curve

- The *ROC curve* is a used for simultaneously displaying two types of errors for all possible thresholds.
- The name “ROC” comes from communications theory. It is an acronym for *receiver operating characteristics*.
- The overall performance of a classifier, summarized over all possible thresholds, is given by the *area under the (ROC) curve* (AUC).
- An ideal ROC curve should touch the top left corner, so the larger the AUC the better the classifier.

ROC Curve

Threshold Point	Sensitivity	Specificity	1 – Specificity
0.0	1.000	0.000	1.000
0.1	0.745	0.942	0.056
0.2	0.610	0.971	0.029
0.3	0.508	0.986	0.014
0.4	0.402	0.992	0.008
0.5	0.315	0.996	0.004
0.6	0.243	0.998	0.002
0.7	0.171	0.999	0.001
0.8	0.090	1.000	0.000
0.9	0.030	1.000	0.000
1.0	0.000	1.000	0.000

ROC Curve



General Rule

AUC	Decision
$AUC = 0.5$	No Discrimination
$0.7 \leq AUC < 0.8$	Acceptable Discrimination
$0.8 \leq AUC < 0.9$	Excellent Discrimination
$AUC \geq 0.9$	Outstanding Discrimination

ROC Curve

- For this data the AUC is 0.95, which is close to the maximum of one so would be considered very good.
- We expect a classifier that performs no better than chance to have an AUC of 0.5.
- ROC curves are extremely useful for comparing different classifiers, since they take into account all possible thresholds.

True Positive and False Positive Rate

- As we have seen above, varying the classifier threshold changes its true positive and false positive rate.
- These are also called the *sensitivity* and one minus the *specificity* of our classifier.
- To make the connection with the epidemiology literature, we may think of “+” as the “disease” that we are trying to detect, and “–” as the “non-disease” state.
- To make the connection to the classical hypothesis testing literature, we think of “–” as the null hypothesis and “+” as the alternative (non-null) hypothesis.
- In the context of the Default data, “+” indicates an individual who defaults, and “–” indicates one who does not.

True Positive and False Positive Rate

	Predicted Class			
True Class		- Or Null	+ or Non-null	Total
	- Or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

True Positive and False Positive Rate

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, sensitivity
Pos. Pred. value	TP/P*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N*	

- TP=True Positive: cases with the disease correctly classified as positive
- FN= False Negative: cases with the disease incorrectly classified as negative
- TN= True Negative: cases without the disease correctly classified as negative
- FP= False Positive: cases without the disease incorrectly classified as positive

	Disease				
Test	**Present**	**n**	**Absent**	**n**	**Total**
Positive	True Positive (TP)	*a*	False Positive (FP)	*c*	*a + c*
Negative	False Negative (FN)	*b*	True Negative (TN)	*d*	*b + d*
Total		*a + b*		*c + d*	

Sensitivity	$\frac{a}{a+b}$	**Specificity**	$\frac{d}{c+d}$
Positive Likelihood Ratio	$\frac{\text{Sensitivity}}{1 - \text{Specificity}}$	**Negative Likelihood Ratio**	$\frac{1 - \text{Sensitivity}}{\text{Specificity}}$
Positive Predictive Value	$\frac{a}{a+c}$	**Negative Predictive Value**	$\frac{d}{b+d}$

- Sensitivity: probability that a test result will be positive when the disease is present (true positive rate, expressed as a percentage).
  = a / (a+b)
- Specificity: probability that a test result will be negative when the disease is not present (true negative rate, expressed as a percentage).
  = d / (c+d)
- Positive likelihood ratio: ratio between the probability of a positive test result given the presence of the disease and the probability of a positive test result given the absence of the disease, i.e.= True positive rate / False positive rate = Sensitivity / (1-Specificity)
- Negative likelihood ratio: ratio between the probability of a negative test result given the presence of the disease and the probability of a negative test result given the absence of the disease, i.e.= False negative rate / True negative rate = (1-Sensitivity) / Specificity
- Positive predictive value: probability that the disease is present when the test is positive (expressed as a percentage).
  = a / (a+c)
- Negative predictive value: probability that the disease is not present when the test is negative (expressed as a percentage).
  = d / (b+d)

	Disease				
Test	Present	n	Absent	n	Total
Positive	True Positive (TP)	a	False Positive (FP)	c	$a+c$
Negative	False Negative (FN)	b	True Negative (TN)	d	$b+d$
Total		$a+b$		$c+d$	

Sensitivity	$\frac{a}{a+b}$	Specificity	$\frac{d}{c+d}$
Positive Likelihood Ratio	$\frac{\text{Sensitivity}}{1-\text{Specificity}}$	Negative Likelihood Ratio	$\frac{1-\text{Sensitivity}}{\text{Specificity}}$
Positive Predictive Value	$\frac{a}{a+c}$	Negative Predictive Value	$\frac{d}{b+d}$

Training Error Rate and Test Error Rate

- The misclassification error rate calculated earlier with the optimal threshold was 13.81%.
- However, we have used the same data to train and test our model.
- In reality, this error rate is in fact the *training error* rate.
- In order to assess the accuracy of the model, we should first fit a model using a part of the data and then should examine the performance on the “hold-out” data.
- This error rate is called the *test error* rate.
- Next we have used 80% of the observations to fit the model and 20% of observations are kept aside for validating the model.

Confusion Matrix for the Training Data using Optimal Threshold

	True default status			
Predicted Default Status		No	Yes	Total
	No	6684	27	6711
	Yes	1050	239	1289
	Total	7734	266	8000

$$Sensitivity = \frac{239}{266} = 89.85\%$$

$$Specificity = \frac{6684}{7734} = 86.42\%$$

$$Total\ Error\ Rate = \frac{1077}{8000} = 13.46\%$$

Confusion Matrix for the Test Data using Optimal Threshold

	True default status			
Predicted Default Status		No	Yes	Total
	No	1651	8	1659
	Yes	282	59	341
	Total	1933	67	2000

$$Sensitivity = \frac{59}{67} = 88.06\%$$

$$Specificity = \frac{1651}{1933} = 85.41\%$$

$$Total\ Error\ Rate = \frac{290}{2000} = 14.5\%$$