Inspire…Educate…Transform.

Statistics and Probability in Decision Modeling

Logistic Regression and Naïve Bayes

Dr. Anand Narasimhamurthy

Acknowledgements : A number of slides are due to Dr.Sridhar Pappu

Outline

- Motivation for and basics of logistic regression with simple examples
- Understanding the nuts and bolts of logistic regression
  - Link function, log odds
  - Parameter estimation : Maximum likelihood
- Short review of classification performance evaluation concepts
  - Confusion matrix, Sensitivity, Specificity, ROC
- Hands on exercise
- Naive Bayes
  - Review of Bayes Theorem
  - Understanding the “Naive” part of Naive Bayes

Applications

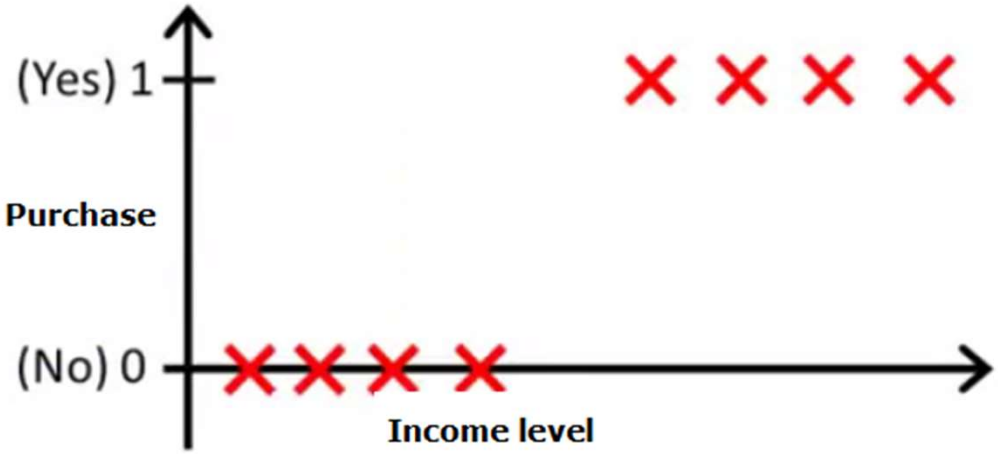
- Predicting stock price movement (up/down)
- Predict whether a patient has diabetes or not
- Predict whether a customer will buy or not
- Predict the likelihood of loan default

Examples of classification problems with multiple classes

- Given an article – predict which section of the newspaper (Current News, International, Arts, Sports,Fashion etc) it supposed to go
- Given a photo of a car number plate, identify which state it belongs to
- Audio clip of a song, identify the genre

LOGISTIC REGRESSION

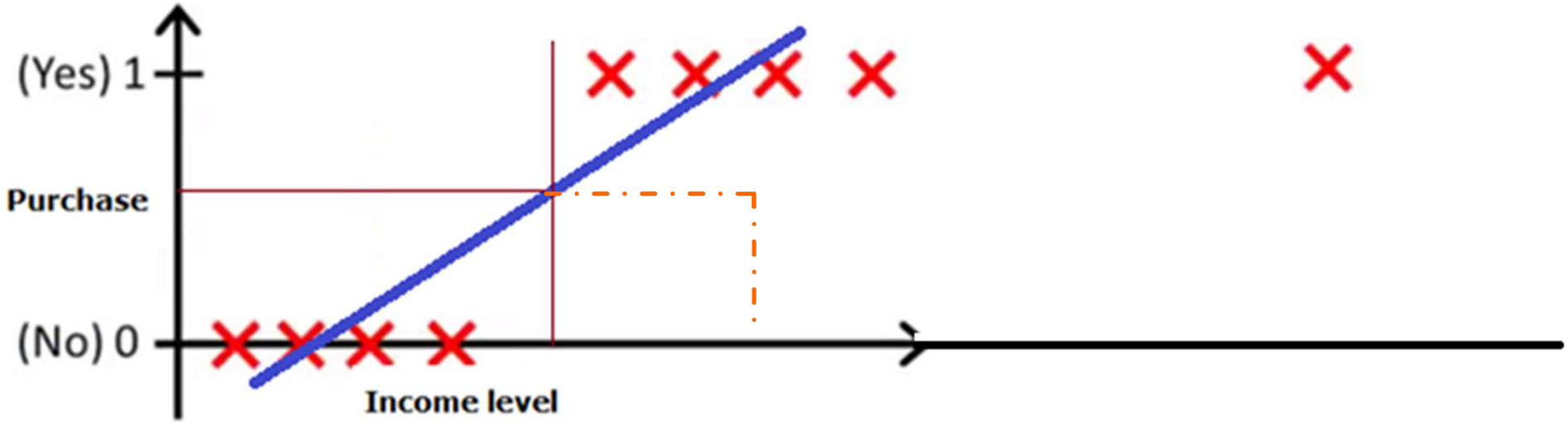
Classification Tasks: Can regression be used?



Consider a two class problem, where the class labels are coded as 0 and 1

How about using linear regression for classification with target values as the class labels 0 and 1?

Linear regression for classification has major drawbacks



Unsuitability of Linear regression

- The output of linear regression is not naturally constrained to lie within a range.

- Linear regression slopes can be much larger than 1 or much smaller than zero and hence thresholding becomes difficult.

Other reasons why Linear regression is not suitable

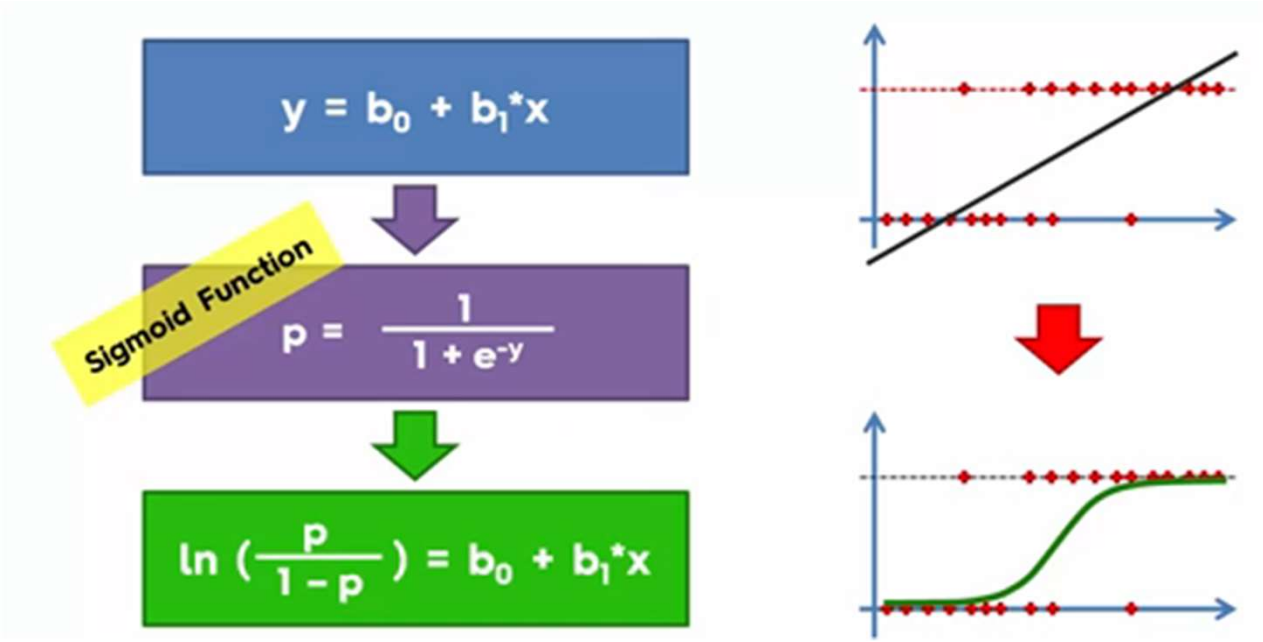
- Basic assumptions of linear regression are clearly violated
  - Error terms do not follow normal distribution.
  - Error terms are not independent.
  - Error variances are heteroscedastic.
- Hence, Linear Regression via Least Squares is inappropriate.

What is actually required?

Mapping from a continuous numeric variable (say x) which possibly lies in a large range to an output value constrained to lie between 0 and 1.

A possible solution : Use of the **sigmoid function**

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



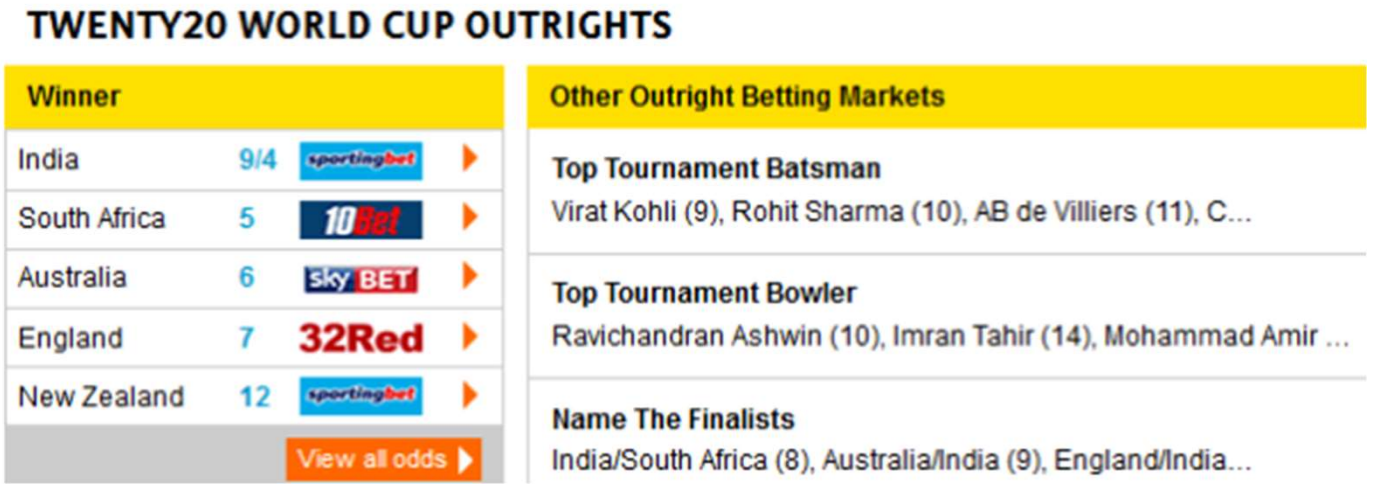
where $z = b_0 + b_1 x$

This can be interpreted as the **odds** i.e. p/(1-p) where p is the probability of success

Slide credit : A youtube video on Logistic regression

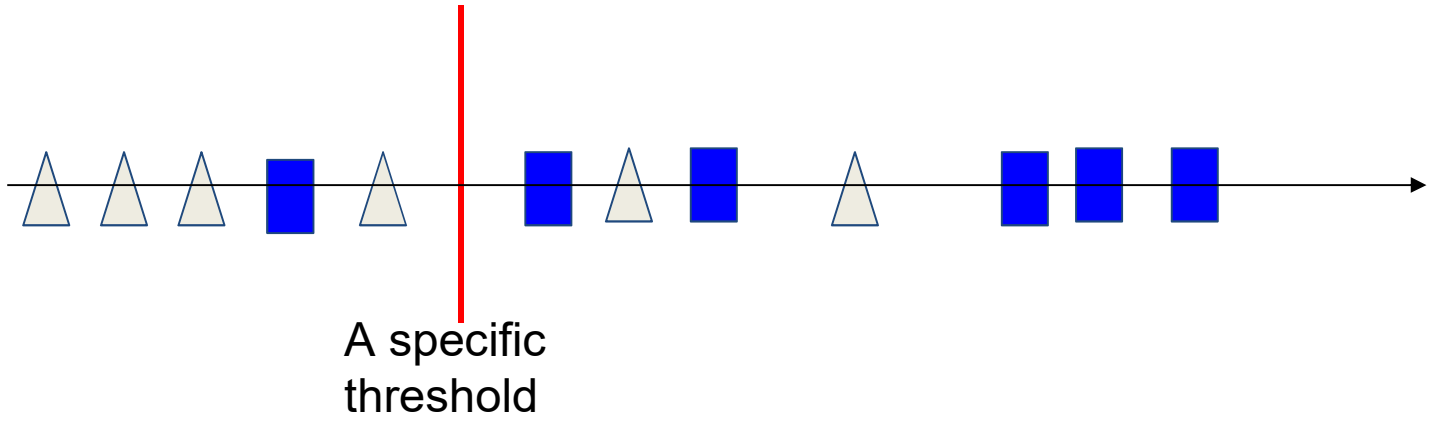
Attention Check – Probability and Odds

If the probability of winning is 6/12, what are the odds of winning?	1:1 (Note, the probability of losing also is 6/12)
If the odds of winning are 19:2, what is the probability of winning?	19/21
If the odds of winning are 3:8, what is the probability of losing?	8/11
If the probability of losing is 6/8, what are the odds of winning?	2:6 or 1:3



Are there easier alternatives for a classification with one predictor variable?

Yes. For example a simple scheme where a threshold is varied in steps and the optimal threshold is determined using the ROC curve.



However we are using the one predictor example to understand the logistic regression technique which can be applied much more generally.

An example : Predict approval based on credit score

MODEL DATA

creditScore	approved
655	0
692	0
681	0
663	1
688	1
693	1
699	0
699	1
683	1
698	0
655	1
703	0
704	1
745	1
702	1

$n = 1000$

creditScore is the applicant's credit score

approved is coded "1" for approved and "0" for not approved; it is a binary, mutually exclusive variable.

* Only 15 of 1000 observations shown

Source : Brandon Foltz,Logistic Regression youtube

Typical questions related to credit score

Source : Brandon Foltz,Logistic Regression youtube

Generalized linear model : Description

- A **generalized linear model** is a broad class of models that includes

many models as special cases.

- In a GLM, there is a **link function g( )** between η (=$\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots \boldsymbol{\beta}_n x_n$) and the mean of the response variable (i.e. E(Y/X=x))

  The general form for n predictor variables is :
  - $g(\eta) = E(Y/X=x)$ (where $\eta = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots \boldsymbol{\beta}_n x_n$)

- **Special cases :**
  - If $g(\eta) = \eta$ (identity function), this reduces to multiple linear regression.
  - If $g(\eta) = 1/(1 + e^{-\eta})$, (sigmoid function) the model is logistic regression

Generalized linear model : More flexible class of models

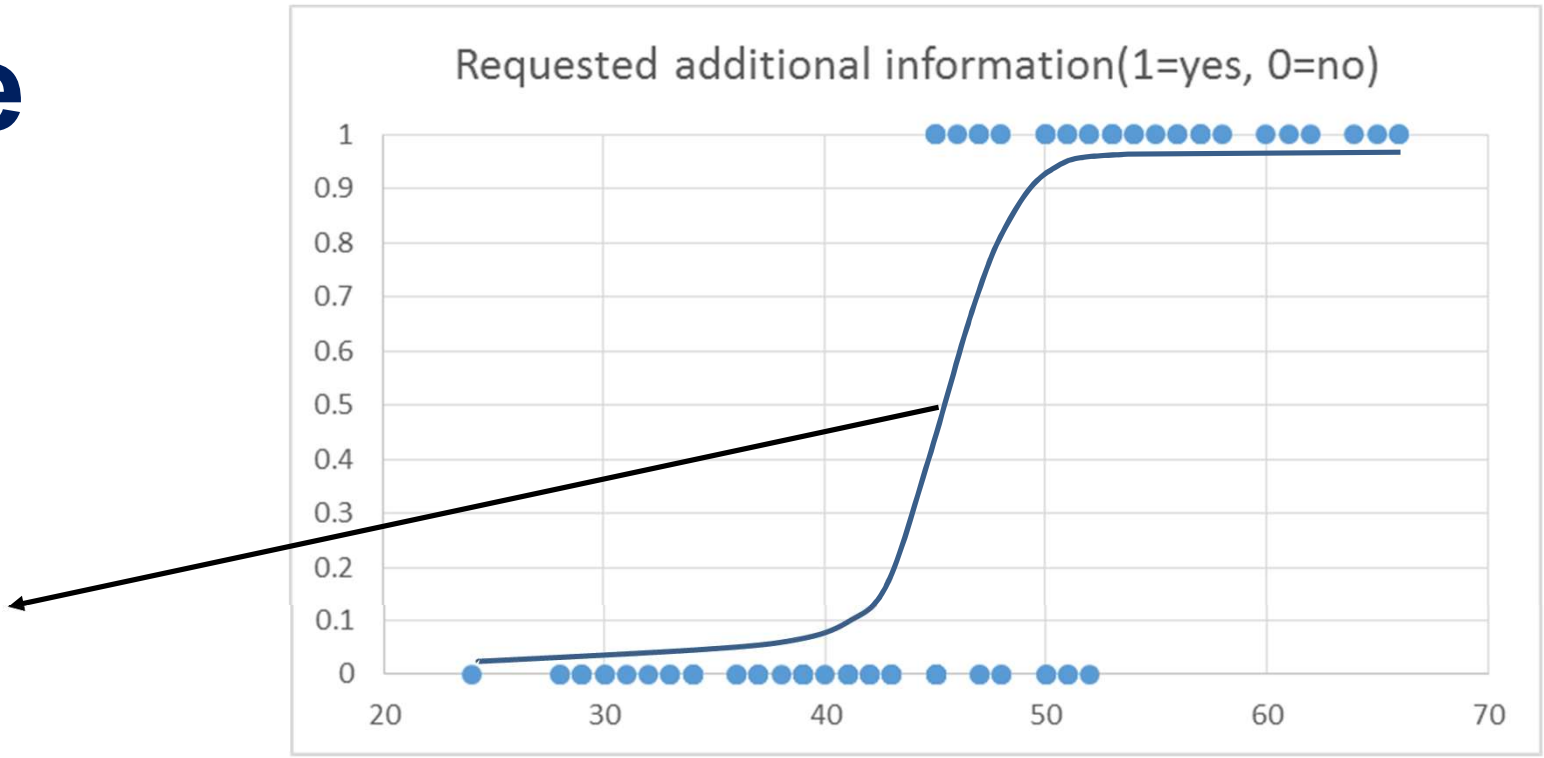
- GLM does not assume a linear relationship between dependent and independent variables. However, it assumes a linear relationship between link function and independent variables in logit model.
- The dependent variable need not to be normally distributed.
- Errors need to be independent but not normally distributed.
- It does not uses OLS (Ordinary Least Square) for parameter estimation. Instead, it uses maximum likelihood estimation (MLE).

Example

An auto club mails a flier to its members offering to send more information regarding a supplemental health insurance plan if the member returns a brief enclosed form.

Can a model be built to predict if a member will return the form or not?

Example


Requested additional information(1=yes, 0=no)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
20
30
40
50
60
70

Logistic model

$$f(x) = p = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}$$

Odds Ratio is obtained by the probability of an event occurring divided by the probability that it will not occur.

Logistic model can be transformed into an odds ratio:

$$S = Odds\ ratio = \frac{p}{1 - p}$$

Logistic model

$$S = Odds\ ratio = \frac{p}{1-p}$$

$$S = \frac{\frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}}{1 - \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}}$$

$$\therefore, S = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}$$

$$\ln(S) = \ln\left(e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Logistic model

The log of the odds ratio is called logit, and the transformed model is linear in βs.

and Interpreting the output

```
Call:
glm(formula = Response ~ Age, family = "binomial", data = flierresponse)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.95015  -0.32016  -0.05335   0.26538   1.72940

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -20.40782    4.52332  -4.512 6.43e-06 ***
Age           0.42592    0.09482   4.492 7.05e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 123.156  on 91  degrees of freedom
Residual deviance:  49.937  on 90  degrees of freedom
AIC: 53.937

Number of Fisher Scoring iterations: 7
```

What is the logit equation?

Determining Logistic Regression Model

Suppose we want a probability that a 50-year old club member will return the form.

$$\ln(S) = -20.40782 + 0.42592 * 50 = 0.89$$
$$S = e^{0.89} = 2.435$$

The odds that a 50-year old returns the form are 2.435 to 1.

Determining Logistic Regression Model

$$\hat{p} = \frac{S}{S+1} = \frac{2.435}{2.435+1} = 0.709$$

Using a probability of 0.50 as a cut-off between predicting a 0 or a 1, this member would be classified as a 1.

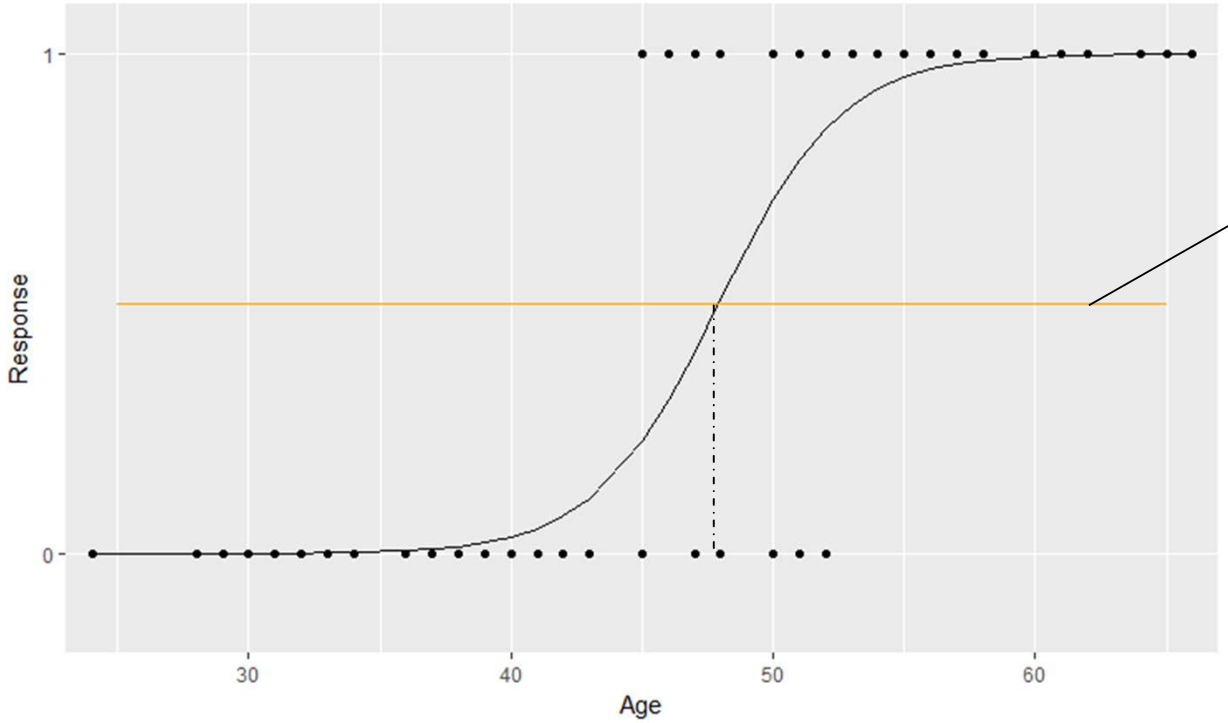
The output of the logistic regression forecast is a probability value. One needs to decide on a threshold value before a class is assigned.

Computing using R

What is the probability that a 50 year-old will return the form?

```
> flierresponseglm <- glm(Response~Age, data = flierresponse, family = "binomial")
> nd <- data.frame(Age=50)  #To predict the probability for Age=50, put that info in a data-frame
> predict(flierresponseglm,newdata=nd)  # This gives the log-Odds
        1
0.8879707
> predict(flierresponseglm,newdata=nd,type="response")  # Compute the probability
        1
0.7084712
```

Visualizing the fit



The threshold of p=0.5, corresponds to the point where Ln(S) = 0.

We can obtain the age at which
the model switches from class 0 to class 1, by setting Ln(S) to be zero in the logistic equation.

$$\ln(S) = -20.40782 + 0.42592\,Age$$

Setting ln(S) = 0 , we get the Age at which probability = 0.5

$Age_c = 20.40782/0.42592 = 47.9$

Interpreting Output – Testing the Overall Model

- AIC provides a means for model selection.
- ***AIC = D + 2k***, where k is the # of parameters in the model including the intercept.
- AIC is *similar to Adjusted* R^2 in the sense it penalizes for adding more parameters to the model.
- It offers a relative estimate of the information lost when a model is used to represent the process that generated the data.
- It does not test a model in the sense of null hypothesis and hence doesn't tell anything about the quality of the model. It is only a relative measure between multiple models.
- AIC = n Log(SSE/n) + 2k for Ordinary Least Squares

Diagnostic Hints

- Overly large coefficient magnitudes, overly large error bars on the coefficient estimates, and the wrong sign on a coefficient could be indications of correlated inputs.

- VIF can be used to check for multicollinearity. R outputs a Generalized Variance Inflation Factor, which is obtained by correcting VIF to the degrees of freedom for categorical predictors. $GVIF = VIF^{\left(\frac{1}{2*df}\right)}$

Parameter estimation :
Maximum Likelihood Estimation (MLE)

How to estimate parameters given the data?

A simple illustrative example (problem statement only)
Suppose we have test scores of 800 students. We believe scores follow a normal distribution (or at least that a normal distribution is a reasonable model for the distribution of scores).

Parameter estimation problem : What are the parameters of the normal distribution (namely μ and σ) that "best fit" the given observations?

Common estimation methods

- Maximum Likelihood Estimation (MLE) :
- Maximum Aposteriori Probability (MAP) : More common in classification settings or Bayesian models.

Likelihood function

Let X be a random variable following an absolutely continuous probability distribution with density function f depending on a parameter Θ.

Then the function $\mathcal{L}(\theta|x) = f_\theta(x)$

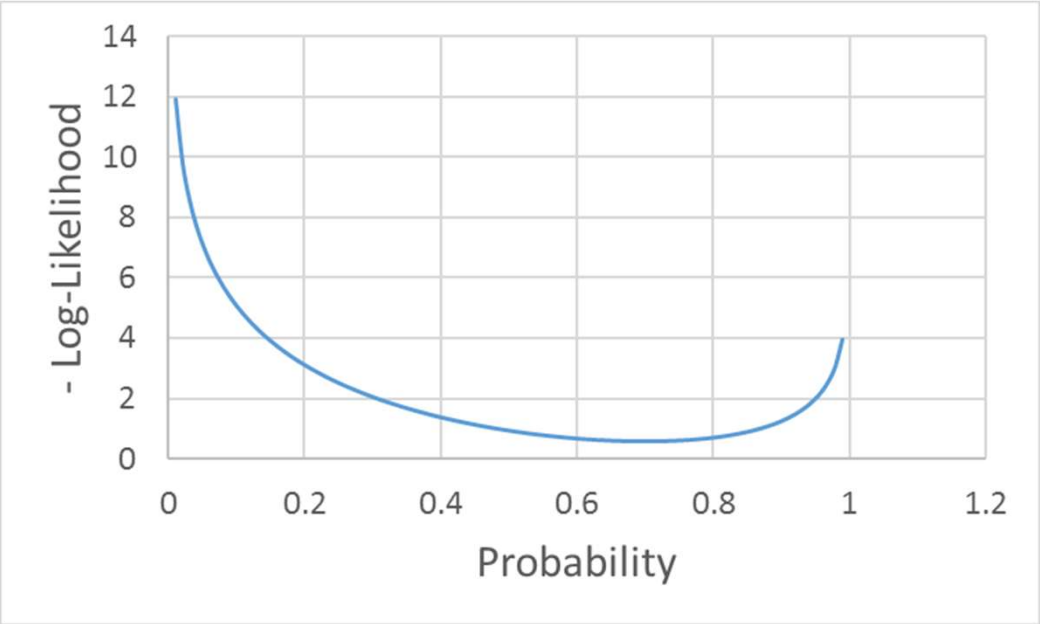
considered as a function of Θ, is the likelihood function (of Θ, given the outcome x of X).

Example : If the density function is the univariate Normal distribution, then the set of parameters (to be estimated) is given by

$$\Theta = [\mu, \sigma]^T \text{ and } f_\theta(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Maximum Likelihood Estimation

- MLE : Goal is to **maximize likelihood**.
- In many cases it is mathematically more convenient to deal with log of the likelihood, and often negative of log likelihood.
- Maximum Likelihood => Minimum of Negative Log-Likelihood.



Graph of an arbitrary likelihood function

MLE specific examples

For the Normal distribution, it can be analytically shown that the **sample mean** and **sample standard deviation** are Maximum Likelihood Estimates (MLE) of the parameters μ and σ of the normal distribution (for an appropriately chosen likelihood function).

For logistic regression, computing a closed form solution analytically is very cumbersome. Usually, an iterative optimization method is used for solving for the parameters $[\beta_0, \beta_1, \ldots \beta_n]$

Performance Measures for Regression and Classification Models

A short review of performance metrics for classification

Recall concepts covered in previous sessions

- Accuracy alone can be misleading, especially if there is a class imbalance.

- Hence, depending on the application better to report additional metrics
  Eg. For a two class problem with a target class, report either
  - Sensitivity and specificity or
  - Precision and Recall

Note : Recall is same as sensitivity but Precision is not same as specificity

- Most of the above measures can be derived from the **Confusion Matrix**.
- The **ROC** curve provides a good visual aid to compare performance of different predictors.

Kappa Metric

- Accuracy can often be a misleading metric, when one category occurs more often than other in the given data-set
  - For eg: Occurrence of cancer in general population is 0.4%
  - If a prediction system blindly marks everyone as “No cancer”, it will 99.6% accurate

Kappa Metric

- Kappa metric quantifies how accurate the prediction algorithm is when compared to a random prediction

$$kappa = \frac{totalAccuracy - randomAccuracy}{1 - randomAccuracy}$$

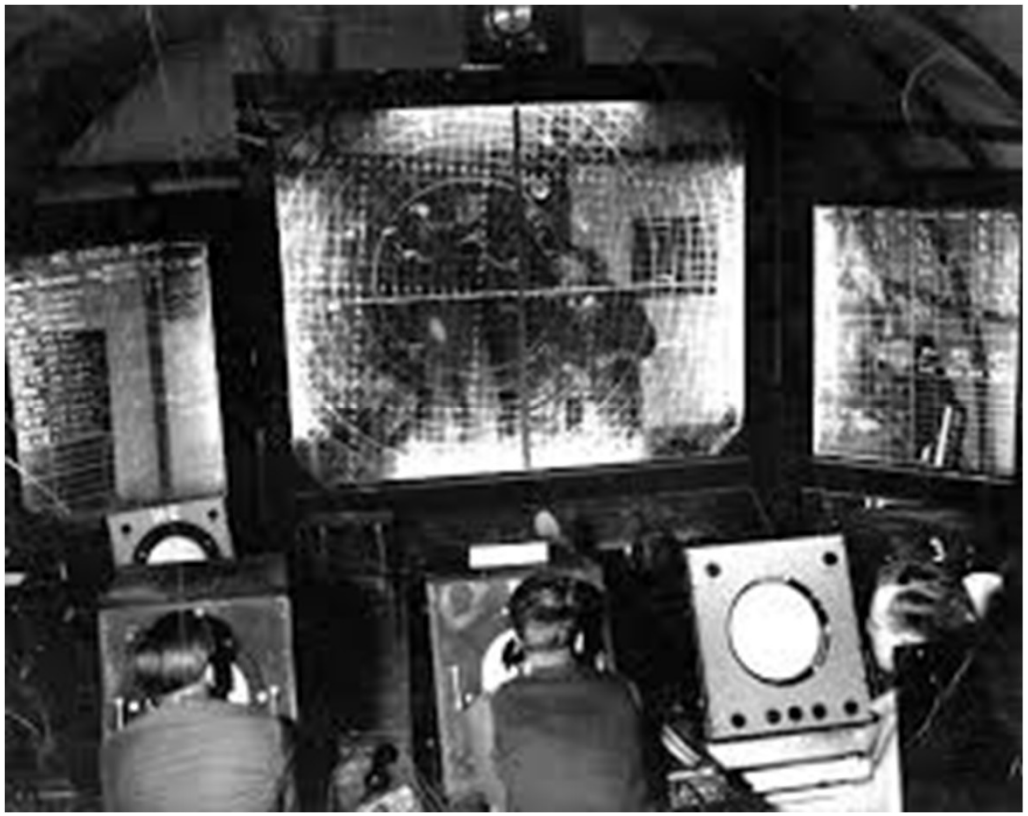
$$\text{totalAccuracy} = \frac{CorrectPredictions}{Total}$$

$$\text{randomAccuracy} = \frac{ActualFalse}{Total} * \frac{PredictedFalse}{Total} + \frac{ActualTrue}{Total} * \frac{PredictedTrue}{Total}$$

Kappa Value	
<0	No agreement
0-0.2	Slight
0.21 to 0.4	Fair
0.4 to 0.6	Moderate
0.6 to 0.8	Substantial
0.8 to 1	Almost Perfect

ROC Curves and AUC

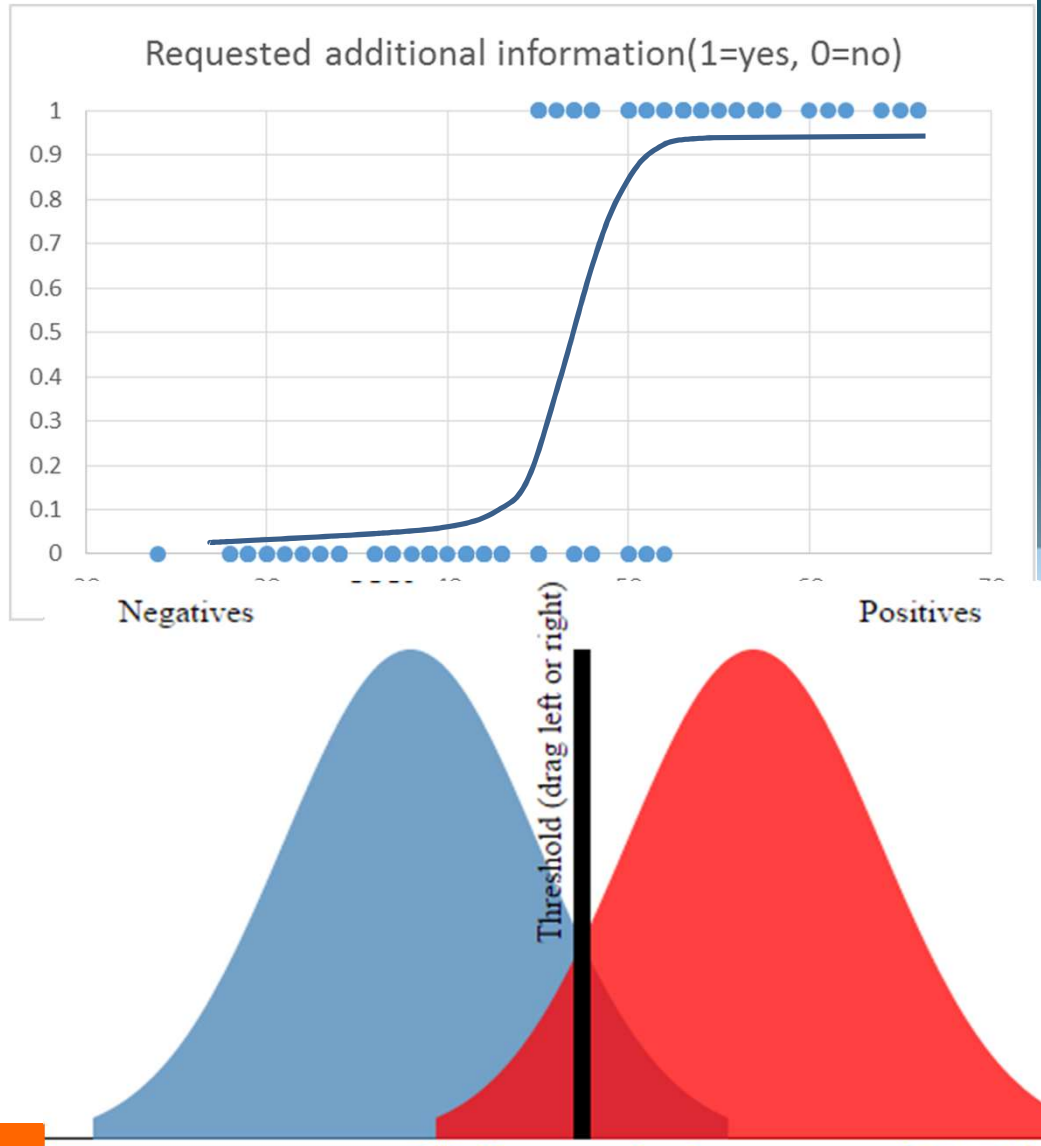
- ROC – Receiver Operating Characteristics
- AUC – Area Under the ROC Curve



Logistic regression gives Probability forecasts for the given data point to be in a given bucket.

.

- A threshold needs to be chosen to finally translate this probability to a bucket allocation



- At a given threshold, we can evaluate the classification accuracy (accuracy, sensitivity, recall, kappa etc)
- ROC curve tries to evaluate how well the regression has achieved the separation between the classes at all threshold values

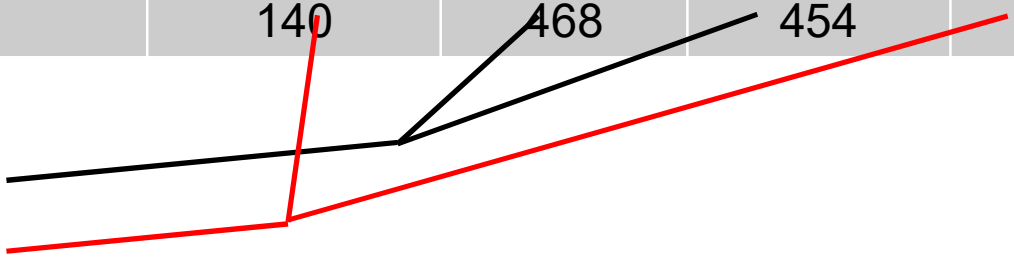
ROC Curve Demo

- http://www.navan.name/roc/
- See: https://youtu.be/OAl6eAyP-yo

ROC Curves and AUC

- ROC – Plot of True Positive Rate vs False Positive Rate, i.e., Sensitivity vs 1-Specificity

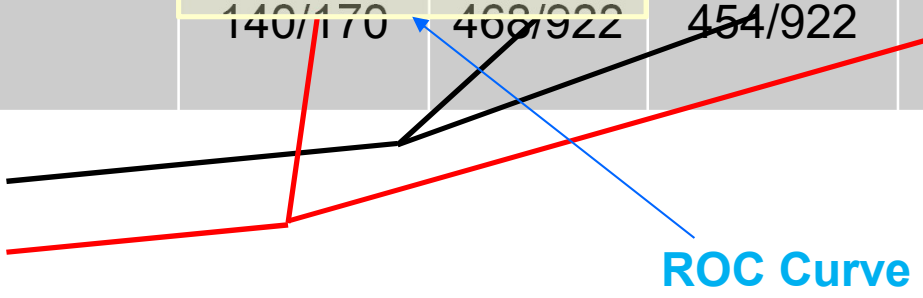
Probability Threshold for Discriminating Between High Risk and Low Risk of Having Ten Year CHD	True Positives	False Positives	True Negatives	False Negatives
0.9	0	0	922	170
0.7	1	1	921	169
0.5	12	7	915	158
0.3	46	76	846	124
0.1	140	468	454	30



ROC Curves and AUC

- ROC – Plot of True Positive Rate vs False Positive Rate, i.e., Sensitivity vs 1-Specificity

Probability Threshold for Discriminating Between High Risk and Low Risk of Having Ten Year CHD	Sensitivity: True Positive Rate	Sensitivity: False Positive Rate	Specificity: True Negative Rate	Specificity: False Negative Rate
0.9	0/170	0/922	922/922	170/170
0.7	1/170	1/922	921/922	169/170
0.5	12/170	7/922	915/922	158/170
0.3	46/170	76/922	846/922	124/170
0.1	140/170	468/922	454/922	30/170



ROC Curves and AUC

- ROC – Plot of True Positive Rate vs False Positive Rate, i.e., Sensitivity vs 1-Specificity

Probability Threshold for Discriminating Between High Risk and Low Risk of Having Ten Year CHD	True Positive Rate (Sensitivity)	False Positive Rate
0.9	0/170	0/922
0.7	1/170	1/922
0.5	12/170	7/922
0.3	46/170	76/922
0.1	140/170	468/922

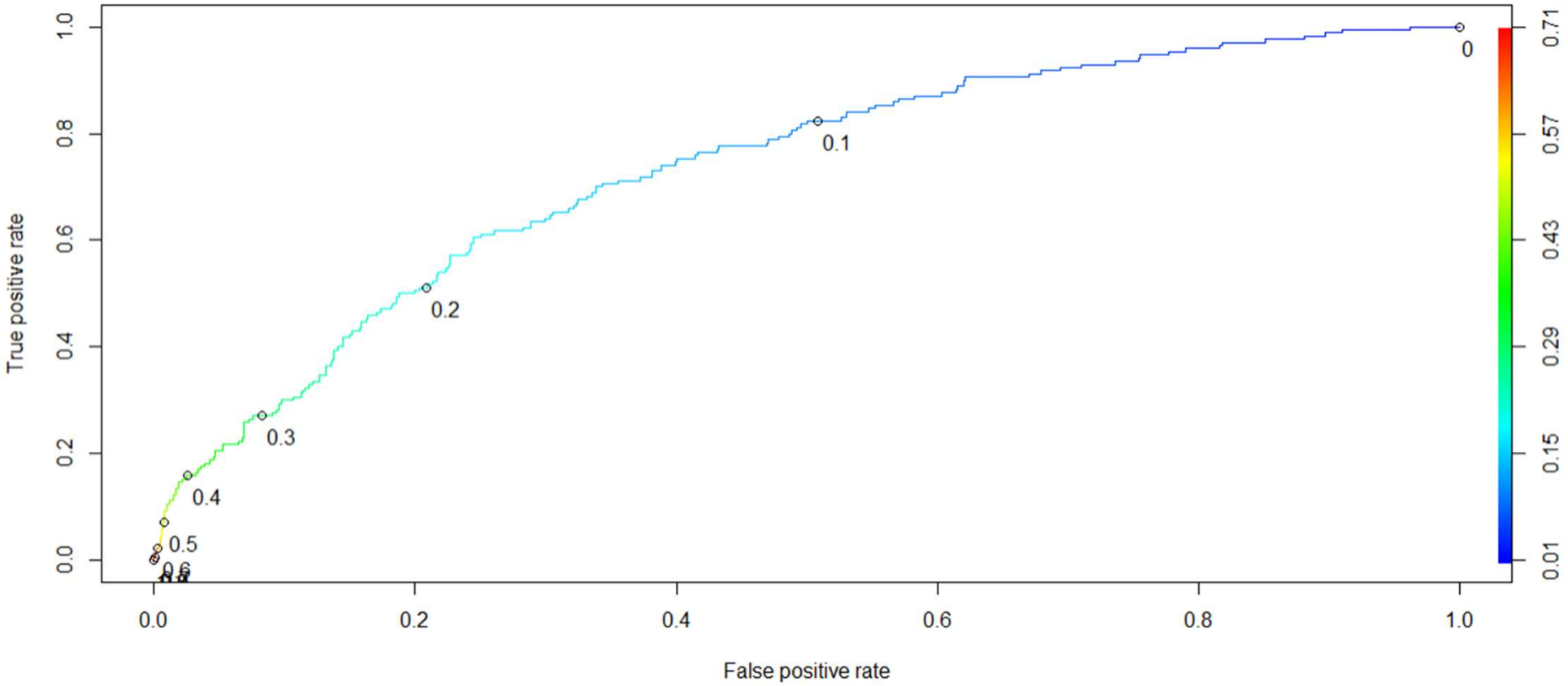
ROC Curve

P(Predicting CHD | Have CHD)

P(Predicting CHD | Do Not Have CHD)

ROC Curves and AUC

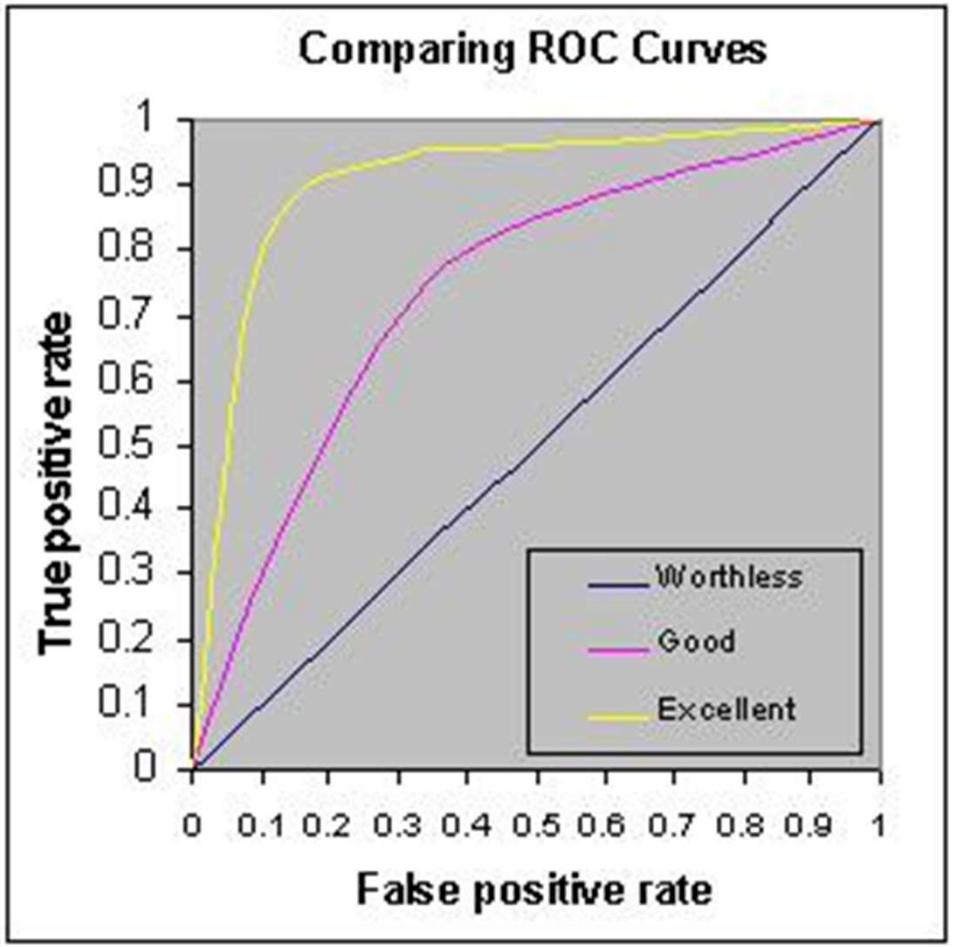
- ROC – Plot of True Positive Rate vs False Positive Rate, i.e., Sensitivity vs 1-Specificity



ROC Curves and AUC

- AUC – Measures discrimination, i.e., ability to correctly classify those with and without CHD.
- If you randomly pick one person who HAS CHD and one who DOESN'T and run the model, the one with the higher probability should be from the high risk group.
- AUC is the percentage of randomly drawn such pairs for which the classification is done correctly.

ROC Curves and AUC



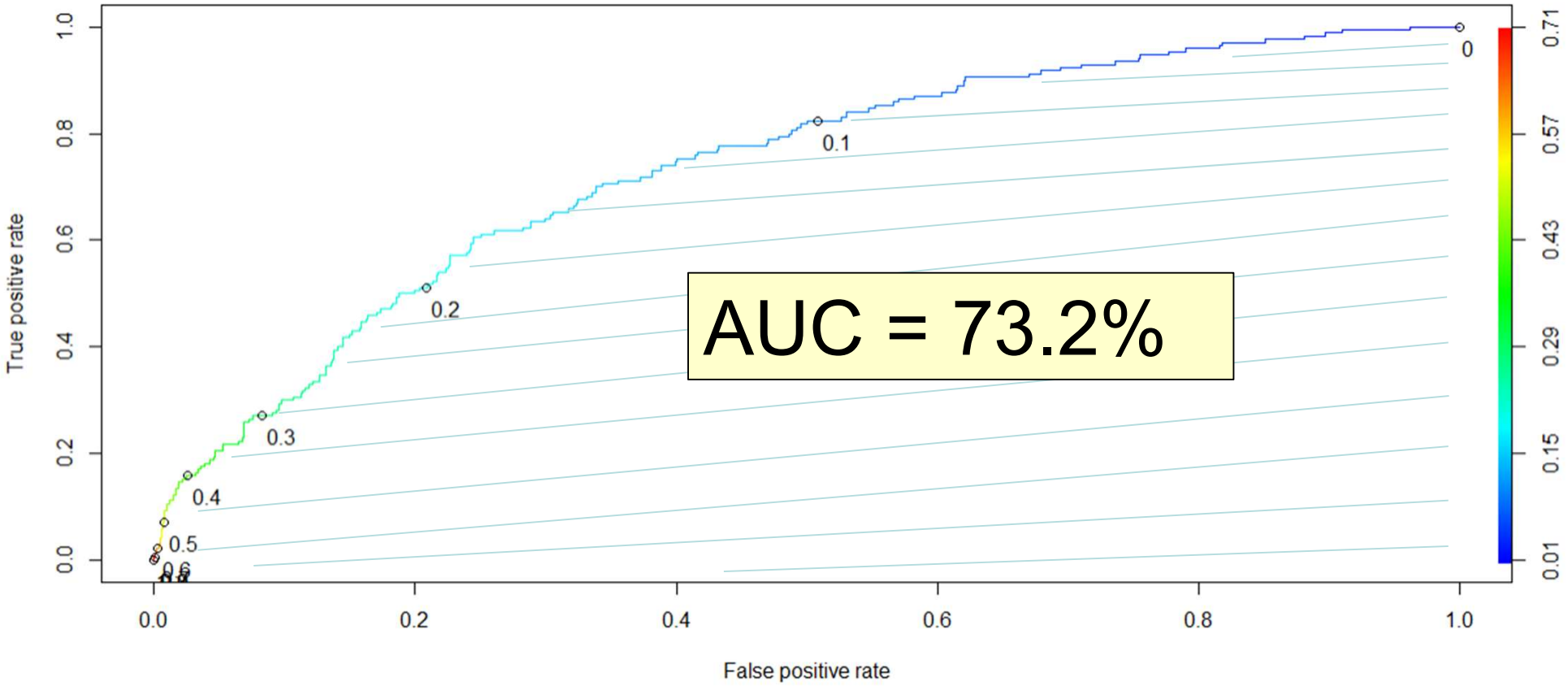
Rough rule of thumb:

- 0.90 -1.0 = Excellent
- 0.80 – 0.90 = Good
- 0.70 – 0.80 = Fair
- 0.60 – 0.70 = Poor
- 0.50 – 0.60 = Fail

- <0.50 – You are better off doing a coin toss than working hard to build a model 😊

ROC Curves and AUC

- The model does a fair job of discrimination between high risk and low risk people.
- Useful for comparing different models.



Hands on exercise

Example – Will the client subscribe a term deposit or not?

A Portuguese banking institution conducted a direct marketing campaign based on phone calls. Often, more than one contact to the same client was required, in order to assess if the product (bank term deposit) would be subscribed ('yes') or not ('no').

Citation: [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014
Data source: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Example – Will the client subscribe a term deposit or not?

Data description an

 # bank client data

- *age* (numeric)
- *job*: type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
- *marital*: marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)

Example – Will the client subscribe a term deposit or not?

 # bank client data

- *education* (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
- *default*: has credit in default? (categorical: 'no','yes','unknown')
- *balance*: money in account at the end of the year (numeric)
- *housing*: has housing loan? (categorical: 'no','yes','unknown')
- *loan*: has personal loan? (categorical: 'no','yes','unknown')

Example – Will the client subscribe a term deposit or not?

 # related with the last contact of the current campaign

- *contact*: contact communication type (categorical: 'cellular','telephone')
- *month*: last contact month of year (categorical: 'jan', 'feb',..., 'nov', 'dec')
- day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
- *duration*: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call, *y* is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Example – Will the client subscribe a term deposit or not?

 # other attributes

- *campaign*: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- *pdays*: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
- *previous*: number of contacts performed before this campaign and for this client (numeric)
- *poutcome*: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

Example – Will the client subscribe a term deposit or not?

Case – Framingham Heart Study

Framingham Heart Study
A Project of the National Heart, Lung, and Blood Institute and Boston University

- Committed to identifying common factors contributing to cardiovascular disease (CVD).
- Setup in the town of Framingham, MA in 1948.
- Random sample consisting of 2/3rds of adult population in the town.

AGE-SEX DISTRIBUTION AT ENTRY (1948)

Age	29-39	40-49	50-62	Totals
Men	835	779	722	2,336
Women	1,042	962	869	2,873
Totals	1,877	1,741	1,591	5,209

Case Study – Data (framinghamheartstudy.org and MITx)

- 5209 men and women participated.
- Age range: 30-62
- People who had not yet developed overt symptoms of CVD or suffered a heart attack or stroke.
- Careful monitoring of Framingham Study population has led to identification of major CVD risk factors.
- Led to development of Framingham Risk Score, a gender specific algorithm used to estimate the 10-year cardiovascular risk of an individual: http://cvdrisk.nhlbi.nih.gov/

Case Study – Predicting Coronary Heart Disease (CHD)

Data description

4240 observations; 15 predictor and 1 predicted variables

- *TenYearCHD* – To be predicted. Risk of having a heart attack or stroke in the next 10 years.

Predictors

- Demographic Risk Factors
  - *male*: Gender of subject – Yes or No
  - *age*: Age of subject at first examination
  - *education*: some high school (1), high school (2), some college/vocational college (3), college (4)

Case Study – Predicting Coronary Heart Disease (CHD)

- Behavioural Risk Factors
  - *currentSmoker*: Yes or No
  - *cigsPerDay*: No. of cigarettes smoked per day if smoker
- Medical History Risk Factors
  - *BPmeds*: On BP medication at the time of first examination – Yes or No
  - *prevalentStroke*: Did the subject have a previous stroke – Yes or No
  - *prevalentHyp*: Is the subject currently hypertensive – Yes or No
  - *diabetes*: Does the subject currently have diabetes – Yes or No

Case Study – Predicting Coronary Heart Disease (CHD)

- Risk Factors from First Examination
  - *totChol*: Total cholesterol (mg/dL)
  - *sysBP*: Systolic blood pressure (the higher number in BP result)
  - *diaBP*: Diastolic blood pressure (the lower number in BP result)
  - *BMI*: Body Mass Index (kg/m^2)
  - *heartRate*: # of beats per minute
  - *glucose*: Blood glucose level (mg/dL)

Case Study – Predicting Coronary Heart Disease (CHD)

Approach

- "Randomly" split data into training and test in 70:30 ratio.
- Measure prediction accuracies on training and test data

- Although , the split is random, we need to make sure the frequency of the categories are roughly the same in both training and test set.

Test/Train split

```
> # Randomly split the data into training and testing sets
> set.seed(1000)
> split = sample.split(framingham$TenYearCHD, SplitRatio = 0.70)
>
> # Split up the data using subset
> train = subset(framingham, split==TRUE)
> test = subset(framingham, split==FALSE)
> #Check the frequency of CHD in both sets
> cat(sum(train$TenYearCHD)/nrow(train),sum(test$TenYearCHD)/nrow(test))
0.1519542 0.1517296
```

Case Study – Predicting Coronary Heart Disease (CHD)

Results

- Significant variables that cannot be controlled
  - Gender
  - Age
  - Medical history
- Significant variables that can be controlled
  - Smoking habits
  - Cholesterol
  - Systolic BP
  - Blood glucose

```
Call:
glm(formula = TenYearCHD ~ ., family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9392  -0.5998  -0.4211  -0.2771   2.8632

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -8.360272   0.864696  -9.668  < 2e-16 ***
male             0.524080   0.130836   4.006 6.19e-05 ***
age              0.065429   0.008049   8.129 4.34e-16 ***
education       -0.041105   0.059185  -0.695 0.487366
currentSmoker    0.120498   0.187629   0.642 0.520735
cigsPerDay       0.016471   0.007488   2.200 0.027825 *
BPMeds           0.169118   0.282140   0.599 0.548898
prevalentStroke  1.156666   0.560179   2.065 0.038940 *
prevalentHyp     0.307077   0.166034   1.849 0.064389 .
diabetes        -0.319937   0.392574  -0.815 0.415087
totChol          0.003799   0.001330   2.856 0.004290 **
sysBP            0.011144   0.004446   2.507 0.012188 *
diaBP           -0.001861   0.007760  -0.240 0.810517
BMI              0.008812   0.015662   0.563 0.573702
heartRate       -0.007273   0.005131  -1.418 0.156296
glucose          0.009227   0.002752   3.353 0.000798 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2176.6  on 2565  degrees of freedom
Residual deviance: 1919.9  on 2550  degrees of freedom
  (402 observations deleted due to missingness)
AIC: 1951.9
```

Missing Values

There are several ways of dealing with missing values.
If large percentage of data for a given variable is missing, then we don't use that variable for building the model.

If the percentage of missing values is small (5 to 10%)

- Naïve method: Replace the missing values with either mean, median or mode
- Intelligent method: Impute the missing values from the relationship between the variables.

See for eg: https://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/

Case Study – Predicting Coronary Heart Disease (CHD)

Results

- Accuracy in training set = 2200/2566 = 85.7%
- Accuracy in testing set = 927/1092 = 84.9%

- Accuracy is affected by imbalance between positives and negatives.
- There is a trade-off between sensitivity and specificity.

Training Set

10-year CHD risk		Predicted	
Actual		True	False
	True	30	357

Testing Set

10-year CHD risk		Predicted	
Actual		True	False
	True	12	158

CSE 7202c

Gains and Lift Charts

- In some business problems, it is not good enough to just classify. For example, in direct mail or phone marketing campaigns, where it costs money to send a mail to each prospect, it is better to be able to rank the prospective buyers by their probability to buy. That way, you can order them and start calling or mailing them in their decreasing order of propensity to buy.
- **Lift** is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model (random selection).

Gains and Lift Charts

- A Lift Chart describes how well a model ranks samples in a particular class.
- The greater the area between the lift curve and the baseline (random selection), the better the model.

Gains and Lift Charts

- A company sends mail catalogs to prospective buyers. It costs the company $1 to print and mail one catalog.
- From past data, they know the response rate is 5%, i.e., if 100,000 prospective customers are contacted, 5000 buy.
- This means that if there is no model and the company randomly contacts the prospects, they will have the following result.

No. of customers contacted	No. of responses
10000	500
20000	1000
30000	1500
.	.
.	.
100000	5000

Gains and Lift Charts

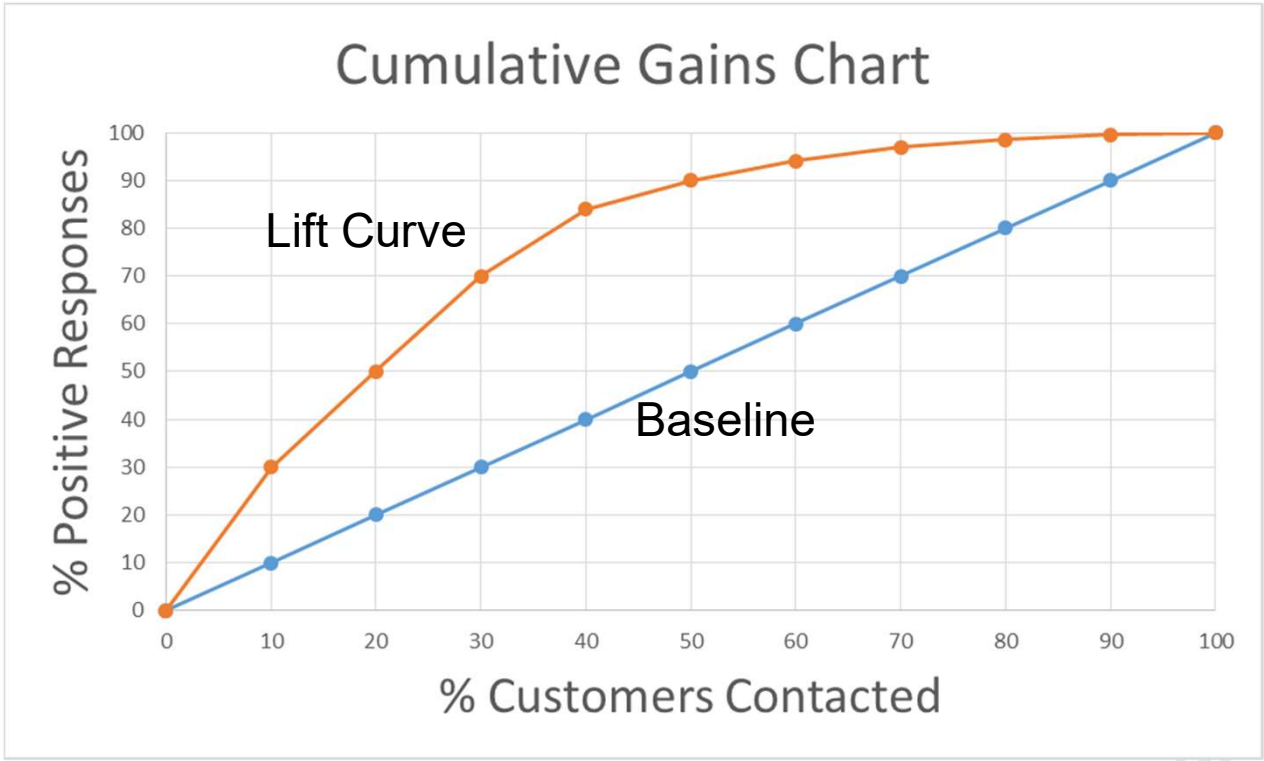
- With a predictive model, where the model assigns a probability to each customer, the customers are ordered and divided into deciles (or any other quantiles). They are then called in decreasing order of probability to buy.

Cost ($)	Decile contacted	Cumulative responses
10000	10 (top decile)	1500
20000	9	2500
30000	8	3500
40000	7	4200
50000	6	4500
60000	5	4700
70000	4	4850
80000	3	4925
90000	2	4975
100000	1	5000

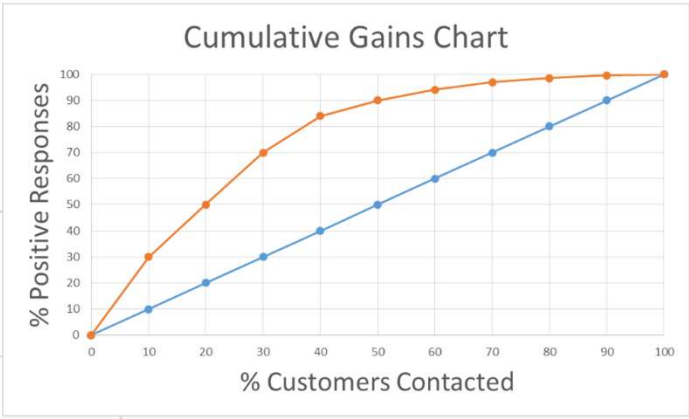
Gains and Lift Charts

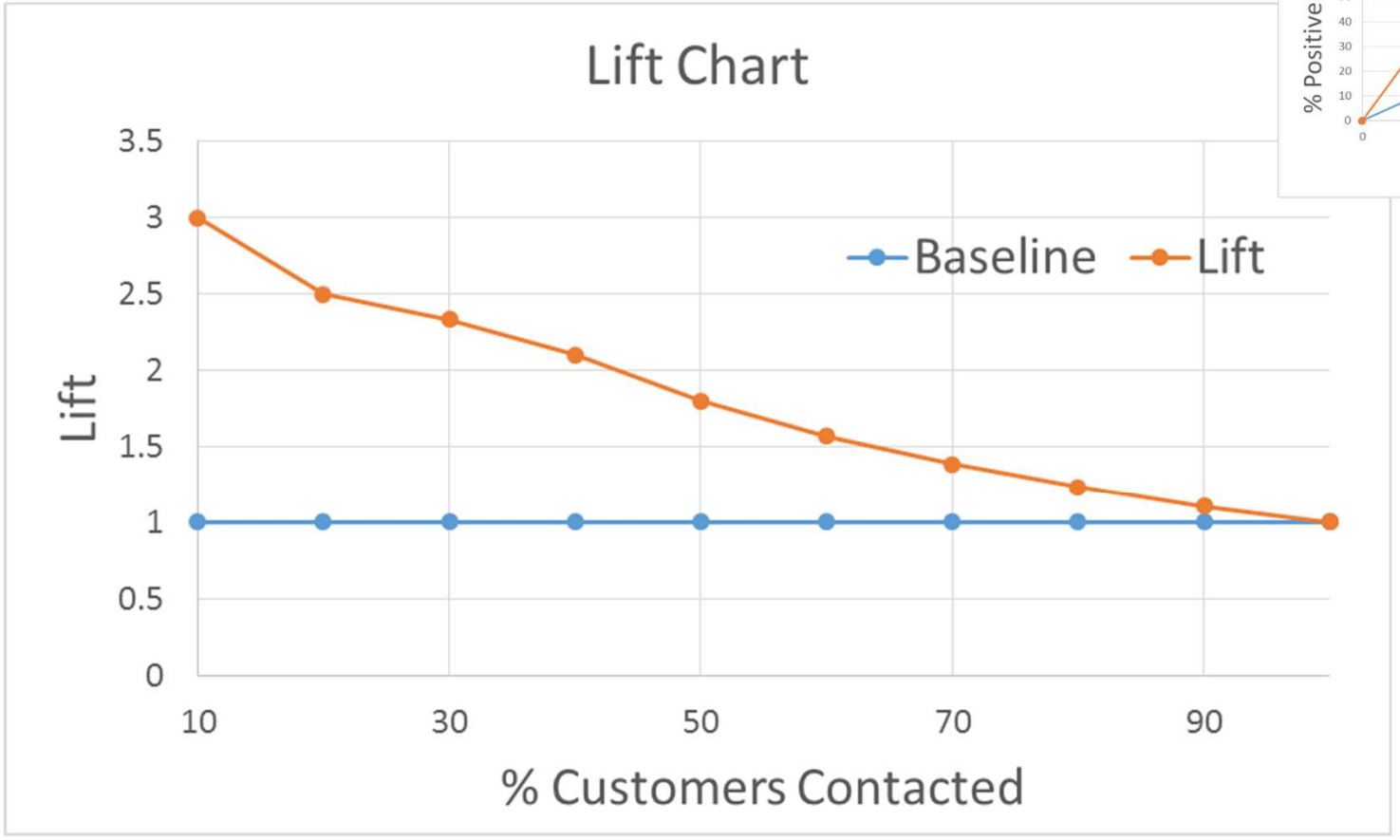
% Called	Called at Random	Called According to Model Score
0	0	0
10	10	30
20	20	50
30	30	70
40	40	84
50	50	90
60	60	94
70	70	97
80	80	98.5
90	90	99.5
100	100	100

Cost ($)	Decile contacted	Cumulative responses
10000	10 (top decile)	1500
20000	9	2500
30000	8	3500
40000	7	4200
50000	6	4500
60000	5	4700
70000	4	4850
80000	3	4925
90000	2	4975
100000	1	5000



Gains and Lift Charts





- Max lift of 3 at the top decile.
- Model advantage diminishes as more customers are contacted, especially in lower deciles.
- Useful to compare different models.

NAÏVE BAYES ALGORITHM

Classification problems

- All classification problems essentially equivalent to evaluating conditional probability
- $P(Y_i \mid X)$ *i.e.* Given certain evidence X, what is the probability that this if from class Y_i
- Logistic Regression solves this problem by modelling the probabilistic relationship between X and Y (sigmoid function, linear in X etc)
- Such models are called Discriminative models

Naïve Bayes Algorithm

- A simple classifier that performs surprisingly well on a large class of problems
- It belongs to a class of methods called Generative Learning Models
- It works best when all the predictor variables are categorical variables.
- Very frequently used in text mining, character image analysis problems.

Review of Bayes Theorem

A review problem :
Suppose there are only two factories A and B that produce a particular machine component. Suppose that it is known from historical data that Factory A on average produces 3.5 defective pieces per 1000 and factory B produces 2 defective pieces per thousand.
B accounts for 60% of total production and A for the remaining.

a) Compute the probability of a machine part being defective.
Hint : Use total probability formula.

(b) Suppose a particular piece was chosen at random and found to be defective. What is the probability that it was manufactured in factory A?
Hint : Use Bayes theorem and express posteriori probabilities in terms of prior probabilities and likelihood.

Recall

$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$

Classification according to Maximum Aposteriori Probability (MAP) rule

MAP rule : Assign the class label which corresponds to the **maximum aposteriori probability**
i.e.

Maximum Aposteriori Probability (MAP) rule
Given an observation x, assign the class which yields highest value for $P(y_j|x)$
i.e.

$$k^* = argmax_j P(y_j|x)$$

If there are K classes y_1, y_2, y_K, compute $P(y_1|x), \ldots, P(y_K|x)$ and assign to x the class that yields the highest value among these.

$$P(y_j|x) = \frac{P(x|y_j)P(y_j)}{P(x)}$$

Classification according to Maximum Aposteriori Probability (MAP) rule and Bayes Theorem

Question. But how to compute P(class=y_j/ **x**) **j=1,…,K?**
Answer. Use Bayes Theorem and related results.

P(class=y_j/ **x**) **= P(x/**class = y_j) x P(class = y_j) / P(**x**)

Class conditional probability also called **Likelihood**

Prior probability

Probability of observing x

Note that the denominator (P(**x**) is the same for all classes and is positive.
We need to focus only on numerator, if interested in just finding out which y_j yields the highest P(class=y_j/ **x**)

An example : Lab activity problem

Goal: given the weather condition, calculate whether tennis can be played?

Assign the class label corresponding to the Maximum Aposteriori Probability (MAP)

P(class = p/o=rain,t=cool,h=normal,w=true)
= P(o=rain,t=cool,h=normal,w=true/ class =p) x P(class = p)
/ P(o=rain,t=cool,h=normal,w=true/)

P(class = n/o=rain,t=cool,h=normal,w=true)
= P(o=rain,t=cool,h=normal,w=true/ class = n) x P(class = n)
/ P(o=rain,t=cool,h=normal,w=true/)

Whichever probability is higher, we would classify the result to that class.

Note that the denominator is the same for both. So we need to focus only on numerator.

An example : Lab activity problem

Goal: given the weather condition, calculate whether tennis can be played?

Assign the class label corresponding to the Maximum Aposteriori Probability (MAP)

Whichever probability is higher, we would classify the result to that class.

Note that the denominator is the same for both. So we need to focus only on numerator.

Naïve Bayes

$$P(class = p|o = rain, t = cool, h = normal, w = true) =$$

$$\frac{P(o=rain,t=cool,h=normal,w=true|class=p)P(clas}{P(o=rain,t=cool,h=normal,w=true)}$$

How to compute joint probabilities such as $P(o = rain, t = cool, h = normal, w = true|class = p)$?

Imposing naive conditional independence assumption, we get :

$$P(o = rain, t = cool, h = normal, w = true|class = p) =$$

$$P(o = rain|class = p)\times$$
$$P(t = cool|class = p)\times$$
$$P(h = normal|class = p)\times$$
$$P(w = true|class = p)$$

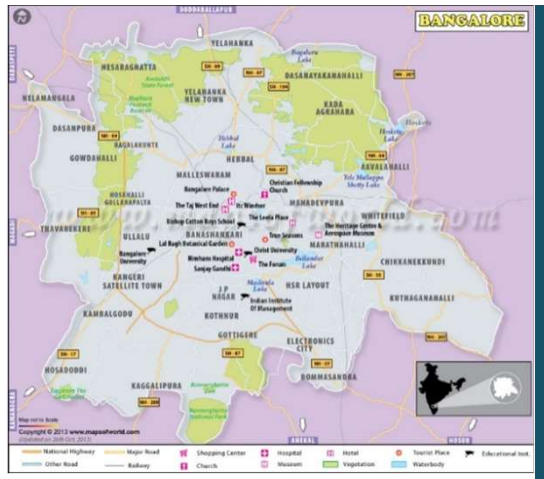
Naïve Bayes

- No parametric fit needed to compute the class
- Prior probabilities can be computed from data
- Individual conditional probabilities were evaluated, and using Bayes relationship the final class probability was evaluated

Naïve Bayes Assumption

- The key assumption of independence of features, is almost never true (and often demonstrably false)
- Still Naïve Bayes does surprisingly well in a lot of situations

Inspire...Educate...Transform.



HYDERABAD

2nd Floor, Jyothi Imperial, Vamsiram Builders, Old Mumbai Highway, Gachibowli, Hyderabad - 500 032
+91-9701685511 (Individuals)
+91-9618483483 (Corporates)

BENGALURU

L77, 15th Cross Road, 3rd Main Road, Sector 6, HSR Layout, Bengaluru – 560 102
+91-9502334561 (Individuals)
+91-9502799088 (Corporates)

Social Media

Web:	
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	
SlideShare:	
LinkedIn:	

This presentation may contain references to findings of various reports available in the public domain. INSOFE makes no representation as to their accuracy or that the organization subscribes to those findings.