Inspire…Educate…Transform.

Statistics and Probability in Decision Modeling

Logistic Regression, ROC and AUC, Gains and Lift Charts, Naïve Bayes Classifier, Performance Measures

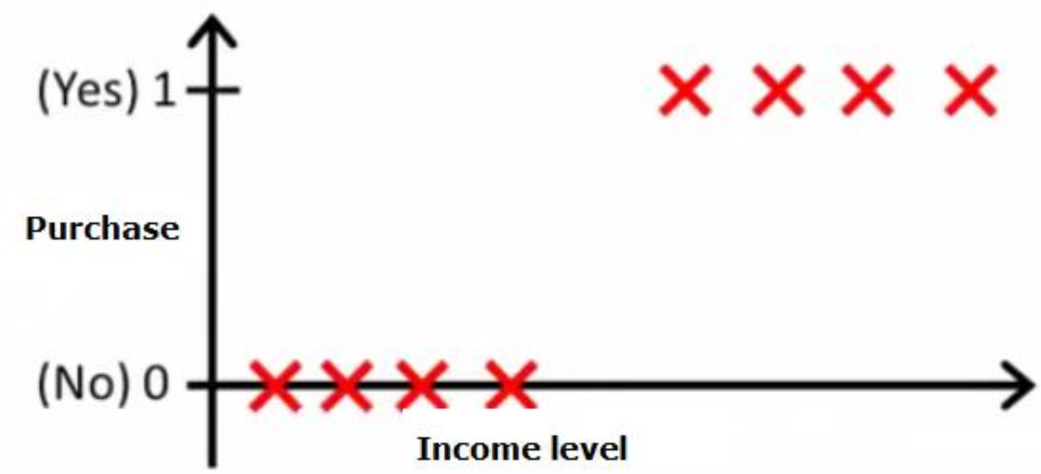
Dr. Sridhar Pappu

Executive VP – Academics, INSOFE

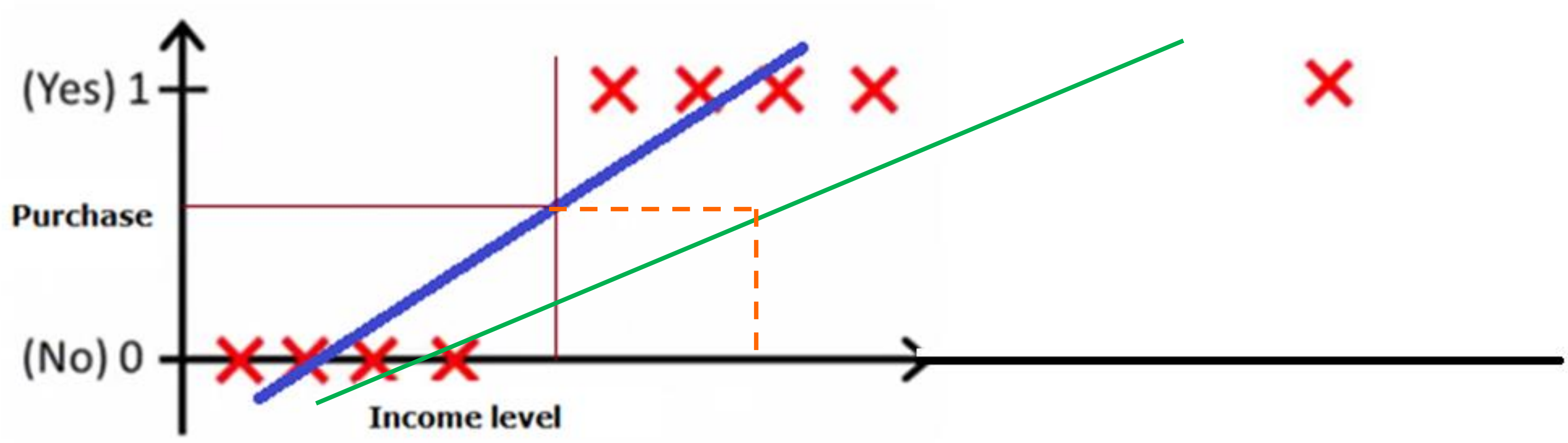
January 07, 2018

LOGISTIC REGRESSION

Classification Tasks: Regression



It could fail



In addition, linear regression hypothesis can be much larger than 1 or much smaller than zero and hence thresholding becomes difficult.

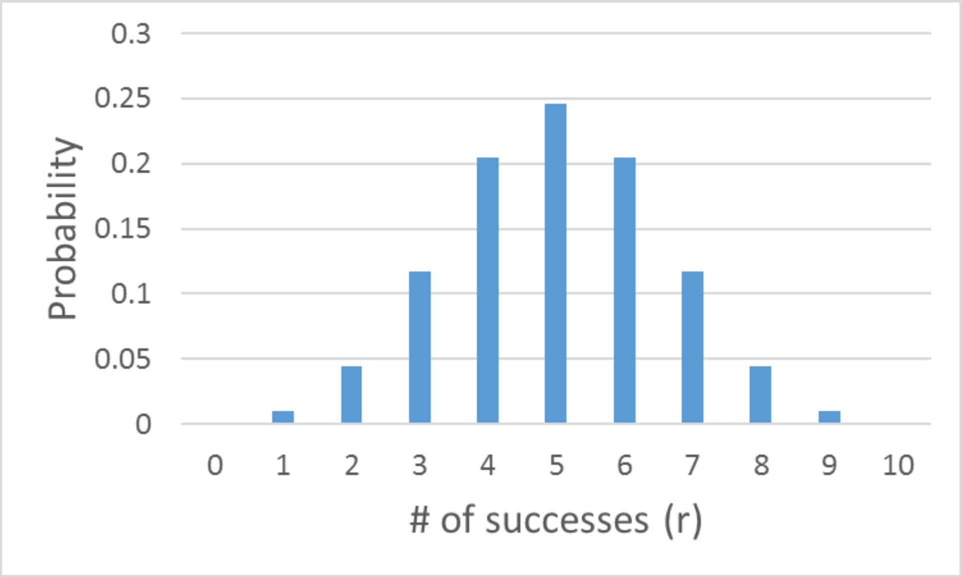
No Assumptions

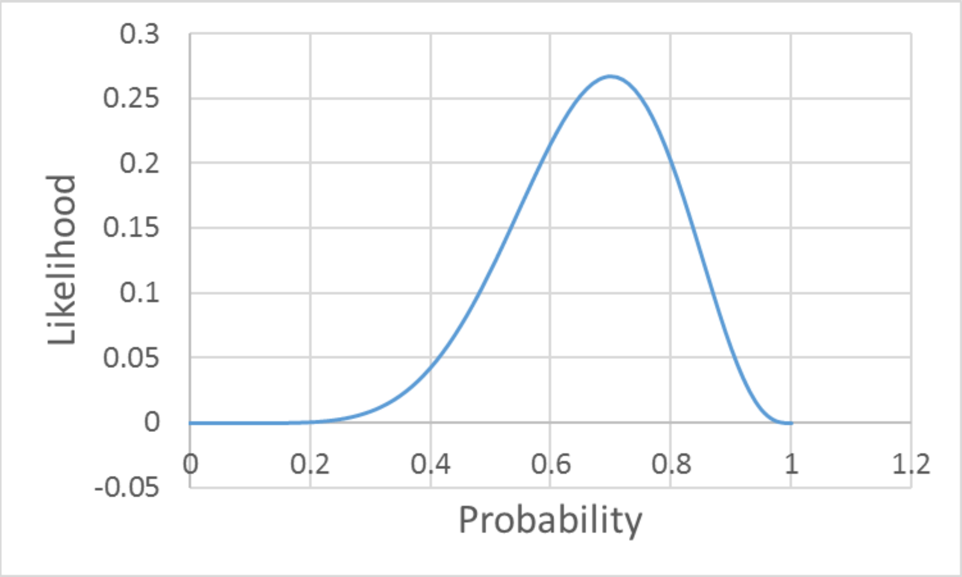
Ordinary Least Squares (OLS) is inappropriate.
Maximum Likelihood Estimation (MLE) is used instead.

Hence avoids assumptions regarding normality and homoscedasticity of errors, and linearity between dependent and independent variables.

Probability vs Likelihood - Excel

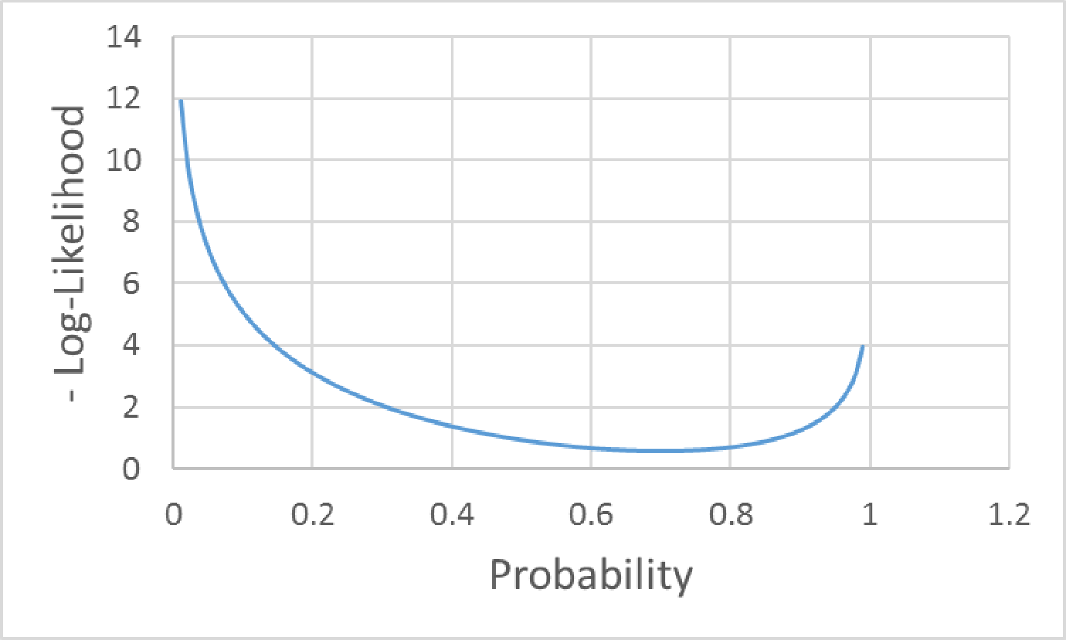
- Likelihood is also known as reverse probability.
- In Probability, we **predict data** based on **known parameters**. *(Recall B(n,p), Geo(p), Po(λ), N(μ, σ^2), etc.)*
- In Likelihood, we **predict parameters** based on **known data**.

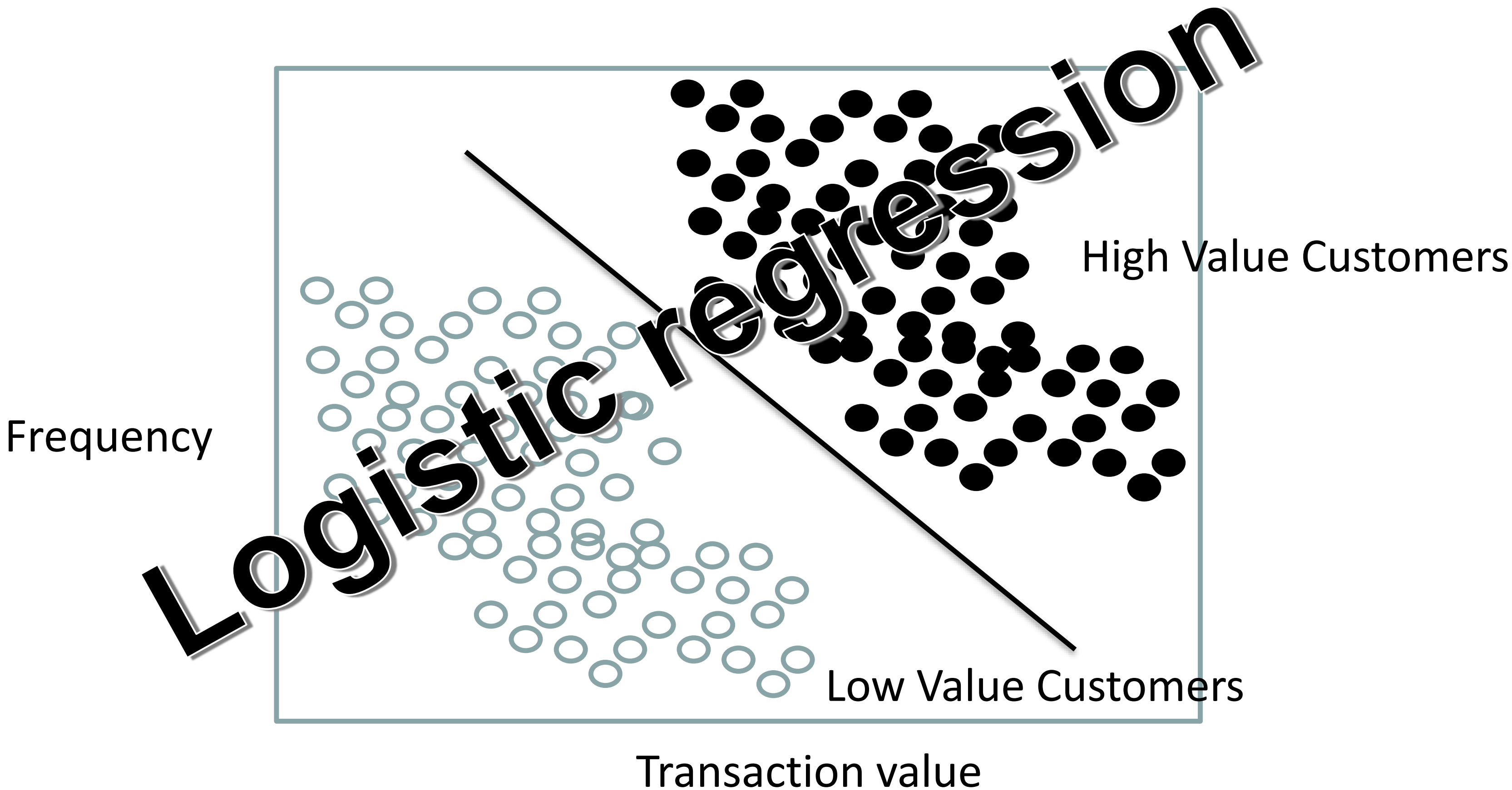




MLE

- Goal is to maximize likelihood.
- In most Data Science optimizations, the goal is to find minima using calculus (minimize sum of squared errors in linear regression, and so on) or numerical techniques like Gradient Descent (minimize deviance in logistic regression, and so on).
- Maximum Likelihood => Minimum of Negative Log-Likelihood.



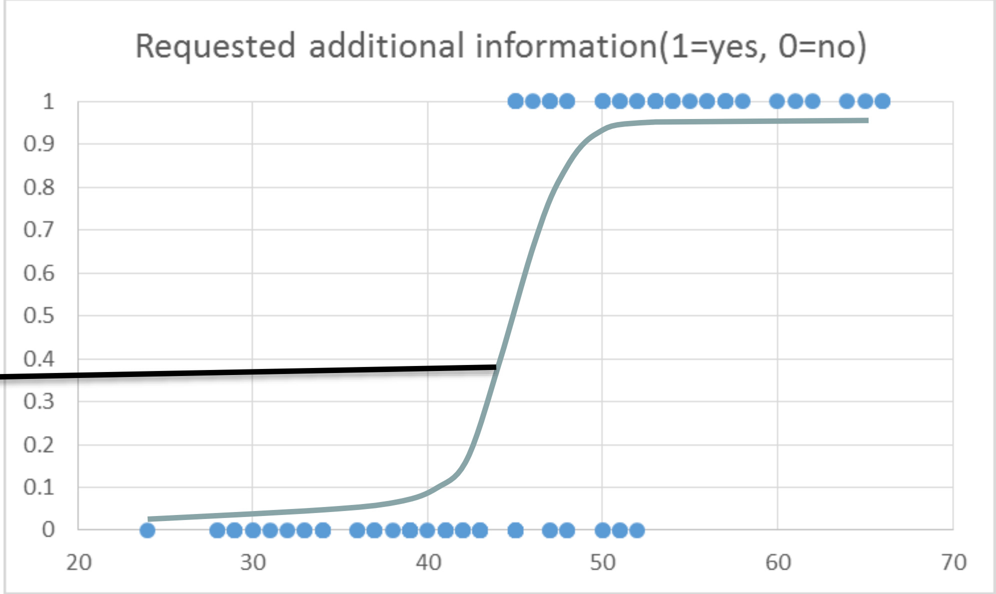

Logistic regression
High Value Customers
Frequency
Low Value Customers
Transaction value

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



$$f(x) = p = \frac{1}{1 + e^{-\mu}} = \frac{e^{\mu}}{1 + e^{\mu}}$$

where $\mu = \beta_0 + \beta_1 x_1$ (also known as the systematic or the structural component or linear predictor).

This is a logistic model. The function is also known as the inverse link function, which links the response with the systematic component.

p is the probability that a club member fits into group 1 (returns the form; success; P(Y=1|X)).

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:

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

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 13:2, what is the probability of winning?	13/15
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

Logistic model

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

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

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

CSE 7302c

Logistic model

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

R 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?

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

CSE 7302c

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 cutoff between predicting a 0 or a 1, this member would be classified as a 1.

Interpreting Output - Deviances

Deviance or **Residual Deviance** is *similar to SSE* in the sense it measures how much remains unexplained by the model built with predictors included.

$$D = -2LL,$$

where LL is the log-likelihood.

```
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
```

Null Deviance shows how well the model predicts the response with only the intercept as a parameter. The intercept is the logarithm of the ratio of cases with *y=1* to the number of cases with *y=0*. This is *similar to SST*, which gives total variation when all coefficients are zero (null hypothesis).

Interpreting Output – Testing the Overall Model

The *z*-values and the associated *p*-values provide significance of individual predictor variables.

R outputs AIC (Akaike's Information Criterion) and you need to pick the model with the lowest AIC.

```
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
```

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. Recall in Linear Regression, it is calculated as ***AIC = nln(RSS/n) + 2k***.
- AIC is *similar to Adjusted* R^2 in the sense it penalizes for adding more parameters to the model.
- 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.

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

Diagnostic Hints

- Coefficients that tend to infinity could be a sign that an input is perfectly correlated with a subset of your responses. Or put another way, it could be a sign that this input is only really useful on a subset of your data, so perhaps it is time to segment the data.

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. $GVIF = VIF^{\left(\frac{1}{2*df}\right)}$

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

CSE 7302c

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

CSE 7302c

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
```

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
	False	9	2170

Testing Set

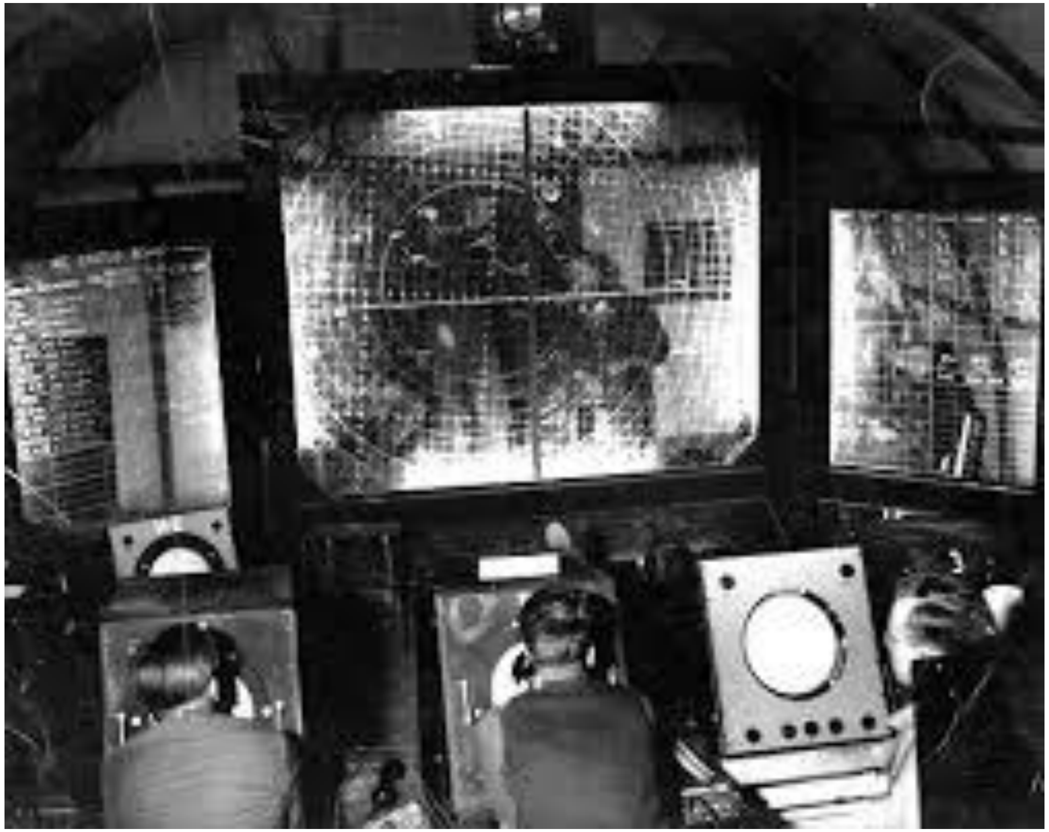
10-year CHD risk		**Predicted**	
Actual		True	False
	True	12	158
	False	7	915

Some More Performance Measures for Regression and Classification Models

CSE 7302c

ROC Curves and AUC

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



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

- Actual Counts
  - Without CHD: 922
  - With CHD: 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	Sensitivity True Positive Rate	False Positive Rate	Specificity True Negative Rate	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

- Actual Counts
  - Without CHD: 922
  - With CHD: 170

ROC Curve

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

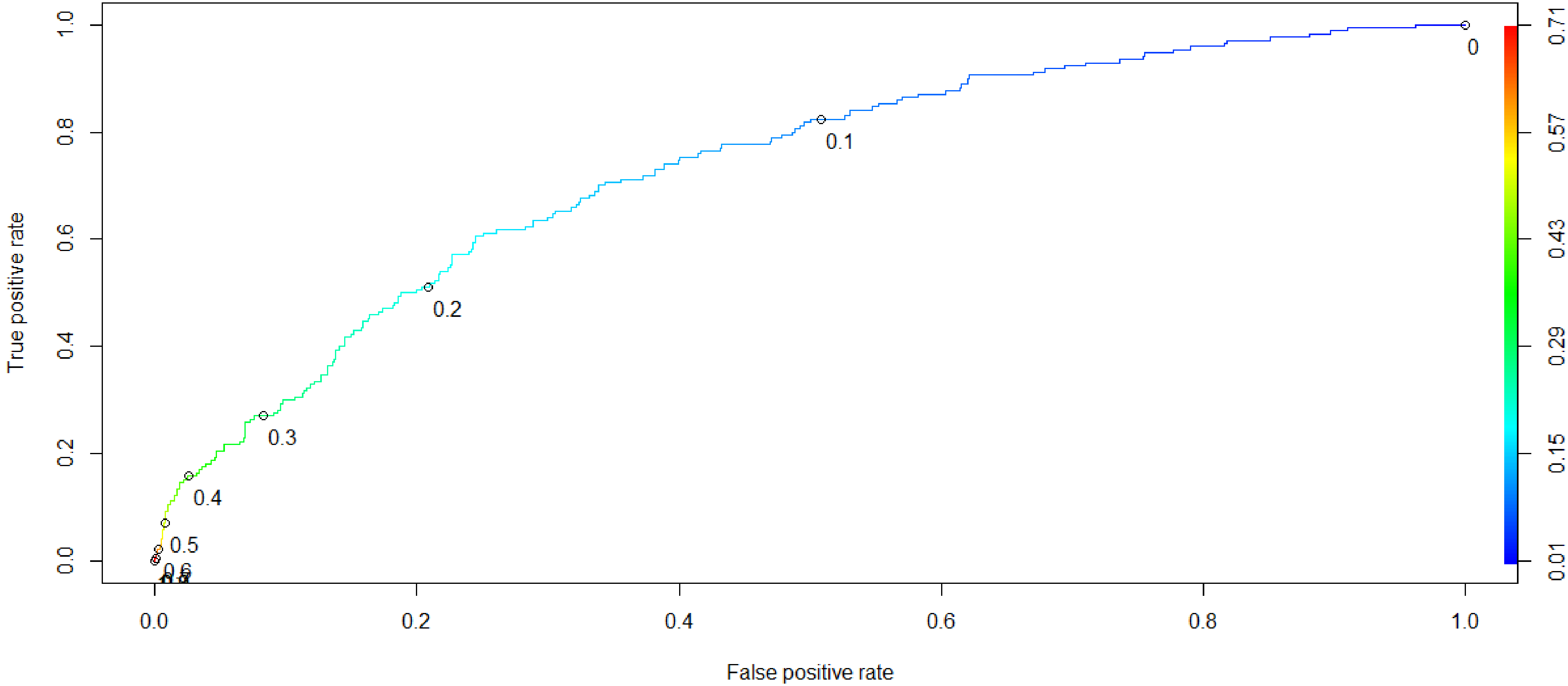
ROC Curve

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

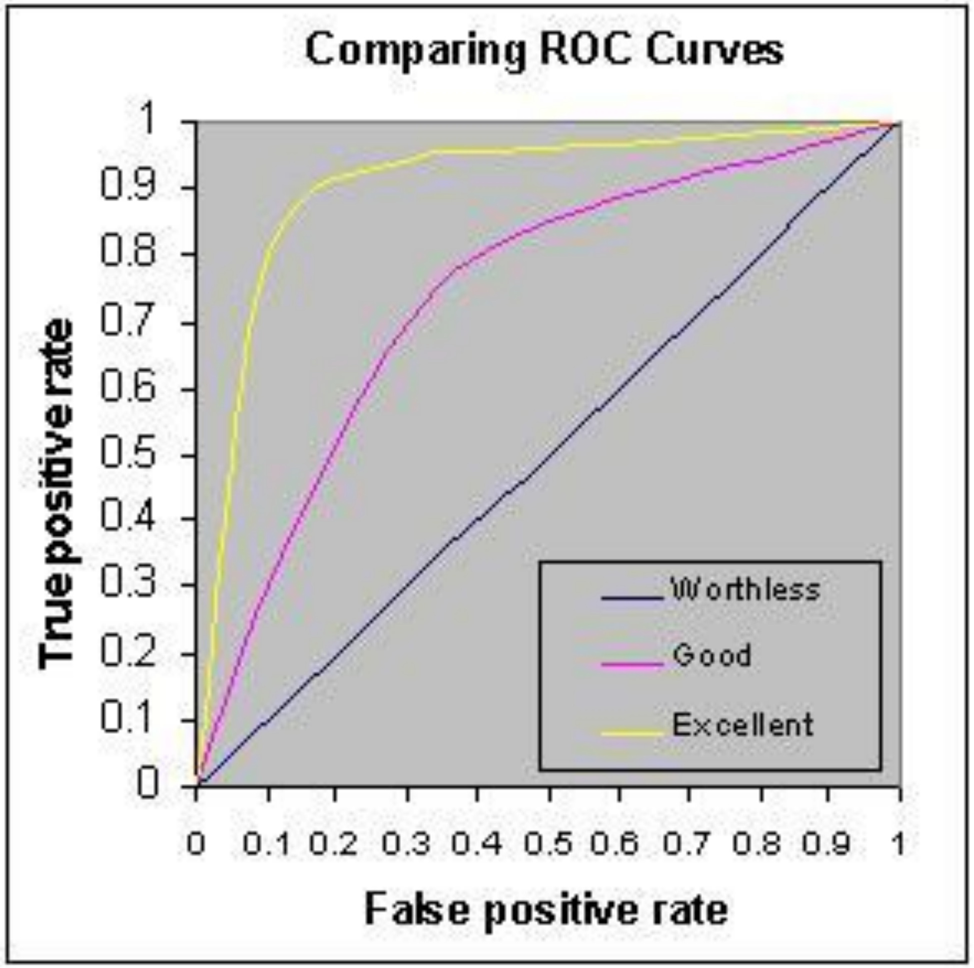


CSE 7302c

ROC Curves and AUC

- AUC – Measures discrimination, i.e., ability to correctly classify those with and without CHD.
- If you randomly pick <u>one</u> person who HAS CHD and <u>one</u> 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



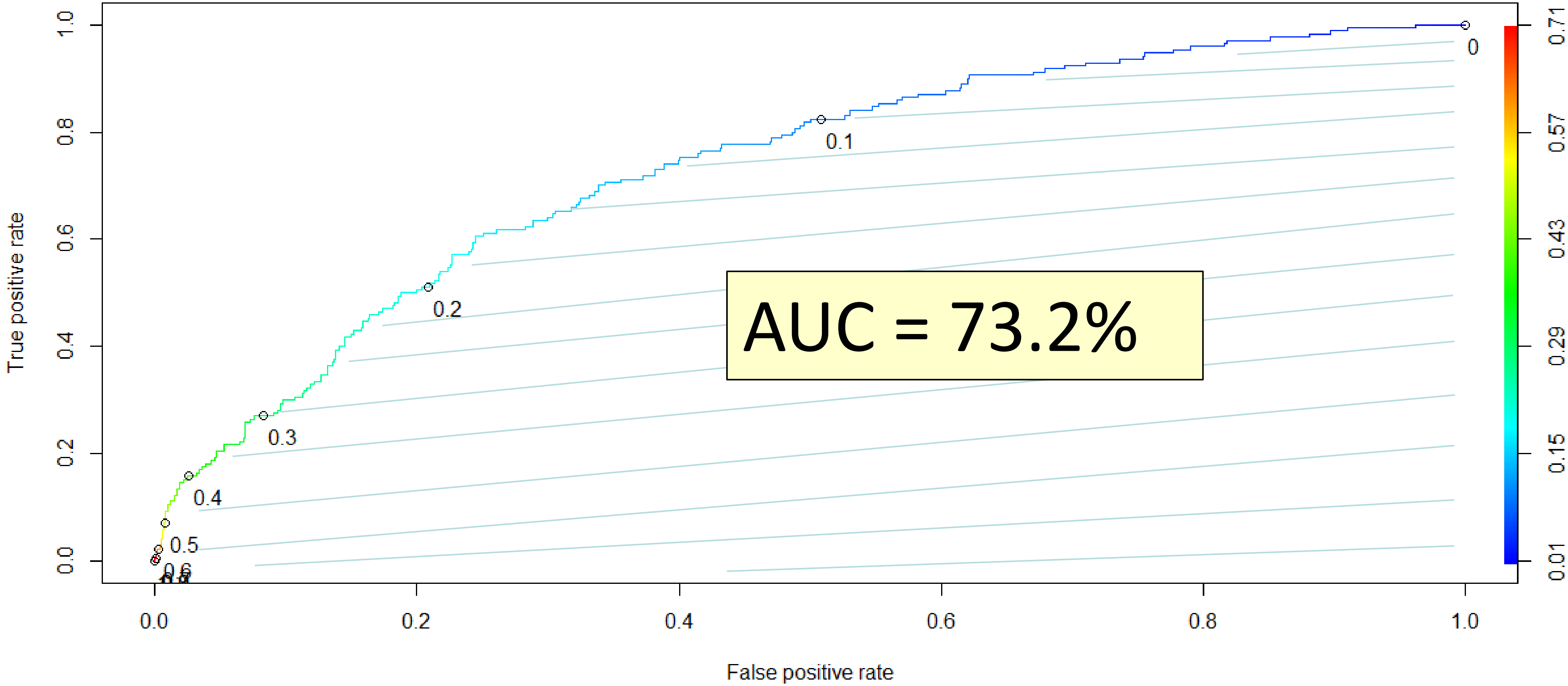
Rough rule of thumb:

- 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.



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.

CSE 7302c

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

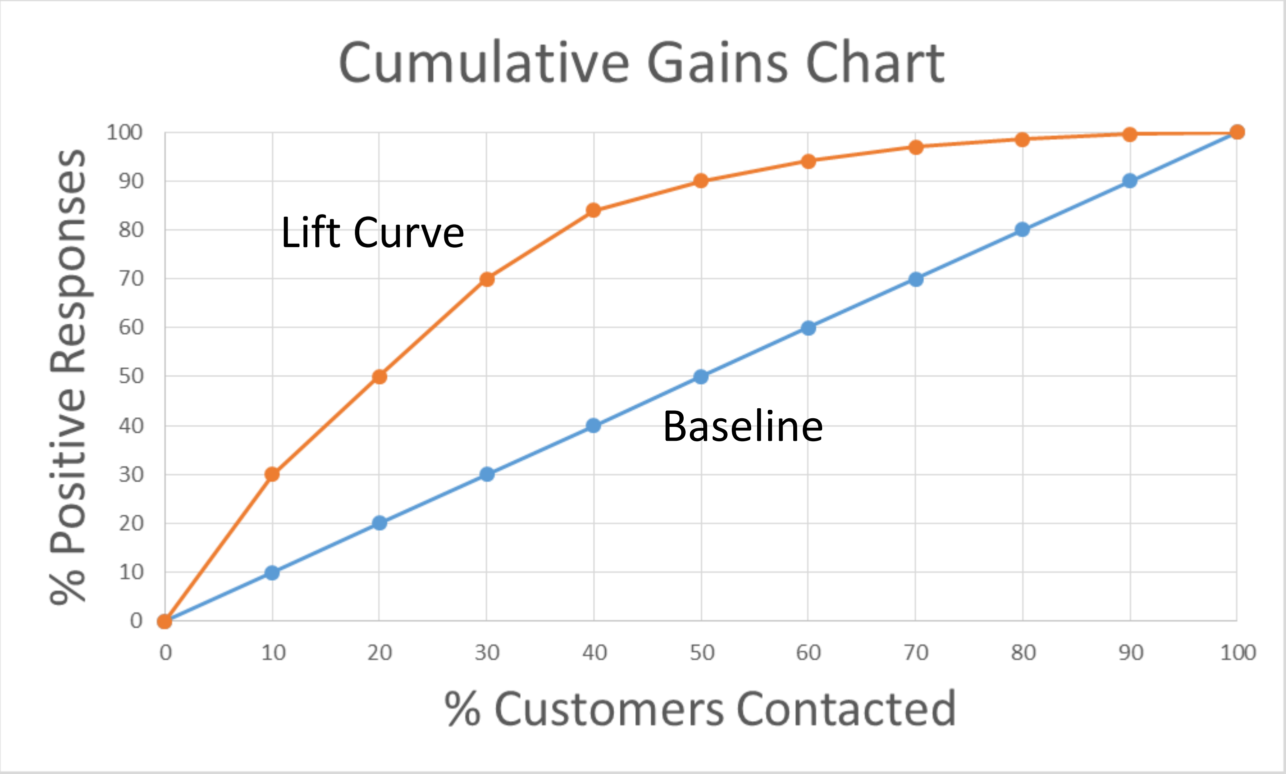
- With a predictive model, where the model assigns a probability to each customer, the customers are ordered and divided into deciles (or any other quantiles). They are then called in decreasing order of probability to buy.

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

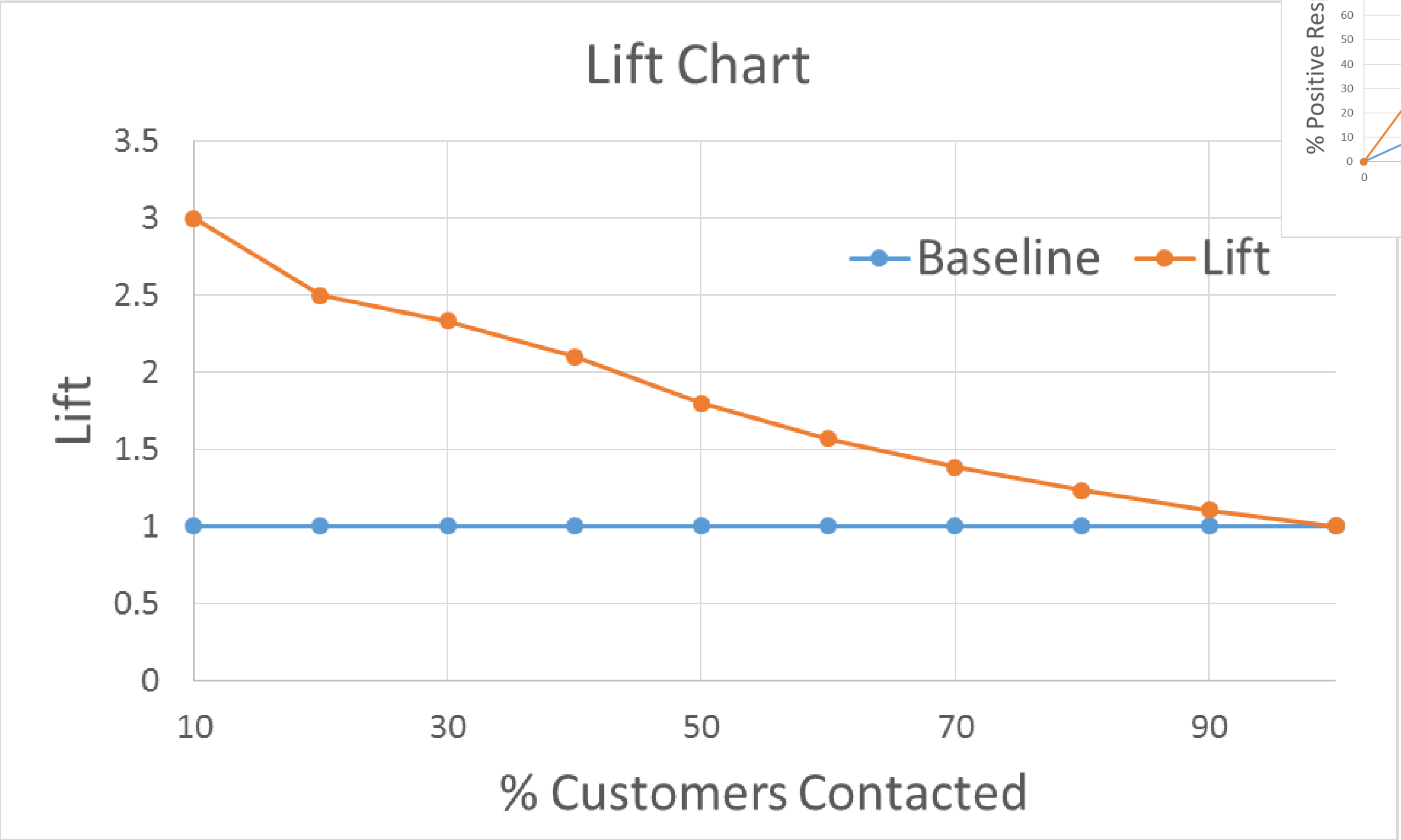
% Called	Called at Random	Called According to Model Score
0	0	0
10	10	30
20	20	50
30	30	70
40	40	84
50	50	90
60	60	94
70	70	97
80	80	98.5
90	90	99.5
100	100	100

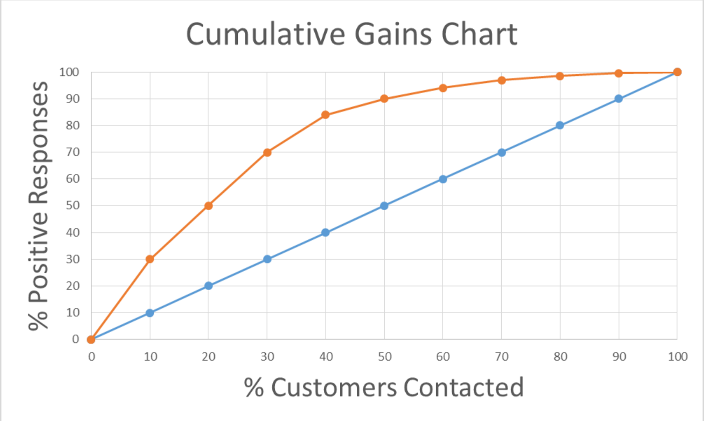
Cost ($)	Decile contacted	Cumulative responses
10000	10 (top decile)	1500
20000	9	2500
30000	8	3500
40000	7	4200
50000	6	4500
60000	5	4700
70000	4	4850
80000	3	4925
90000	2	4975
100000	1	5000



CSE 7302c

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.

Classification

NAÏVE BAYES ALGORITHM

Classification Problems with Multiple Classes

- Given an article, predict which section of the newspaper (Current News, International, Arts, Sports, Fashion, etc.) it is supposed to go to

- Given a photo of a car number plate, identify which state it belongs to

- Given an audio clip of a song, identify the genre

- Given an email, predict whether it is spam or not spam (a 2-class problem)

Classification Problems

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

Naïve Bayes Algorithm

- Naïve Bayes computes $P(Y_i \mid X)$ by using Bayes theorem (computes the joint probability - inverse conditional probability $P(X \mid Y_i)$ times the prior)

- These type of methods are called Generative Learning Models

- A simple classifier that performs surprisingly well on a large class of problems

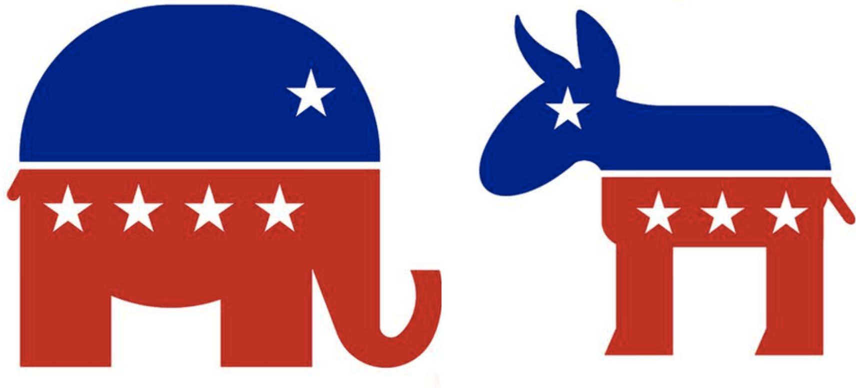
US House of Congress Voting Patterns

Class	V1	V2	V3	V4	V5	V6	V7
republican	n	y	n	y	y	y	n
republican	n	y	n	y	y	y	n
democrat	NA	y	y	NA	y	y	n
democrat	n	y	y	n	NA	y	n
democrat	y	y	y	n	y	y	n
democrat	n	y	y	n	y	y	n
democrat	n	y	n	y	y	y	n
republican	n	y	n	y	y	y	n
republican	n	y	n	y	y	y	n
democrat	y	y	y	n	n	n	y
republican	n	y	n	y	y	n	n
republican	n	y	n	y	y	y	n
democrat	n	y	y	n	n	n	y
democrat	y	y	y	n	n	y	y
republican	n	y	n	y	y	y	n
republican	n	y	n	y	y	y	n
democrat	y	n	y	n	n	y	n
democrat	y	NA	y	n	n	n	y
republican	n	y	n	y	y	y	n
democrat	y	y	y	n	n	n	y

1	**Class Name: 2 (democrat, republican)**
2	handicapped-infants: 2 (y,n)
3	water-project-cost-sharing: 2 (y,n)
4	adoption-of-the-budget-resolution: 2 (y,n)
5	physician-fee-freeze: 2 (y,n)
6	el-salvador-aid: 2 (y,n)
7	religious-groups-in-schools: 2 (y,n)
8	anti-satellite-test-ban: 2 (y,n)
9	aid-to-nicaraguan-contras: 2 (y,n)
10	mx-missile: 2 (y,n)
11	immigration: 2 (y,n)
12	synfuels-corporation-cutback: 2 (y,n)
13	education-spending: 2 (y,n)
14	superfund-right-to-sue: 2 (y,n)
15	crime: 2 (y,n)
16	duty-free-exports: 2 (y,n)
17	export-administration-act-south-africa: 2 (y,n)

House Votes 1984 Dataset: Voting patterns of Members of Congress.
A data frame with 435 observations on 17 variables. 168 Republicans, 267 Democrats

Republican or Democrat?



Republican – R – Red

Democrat – D - Donkey

Given a Congressman's voting pattern (v1 = *y*, v2 = *n*), what is the probability that this person is a Democrat?

$$P(D \mid v1 = y, v2 = n) = ?$$

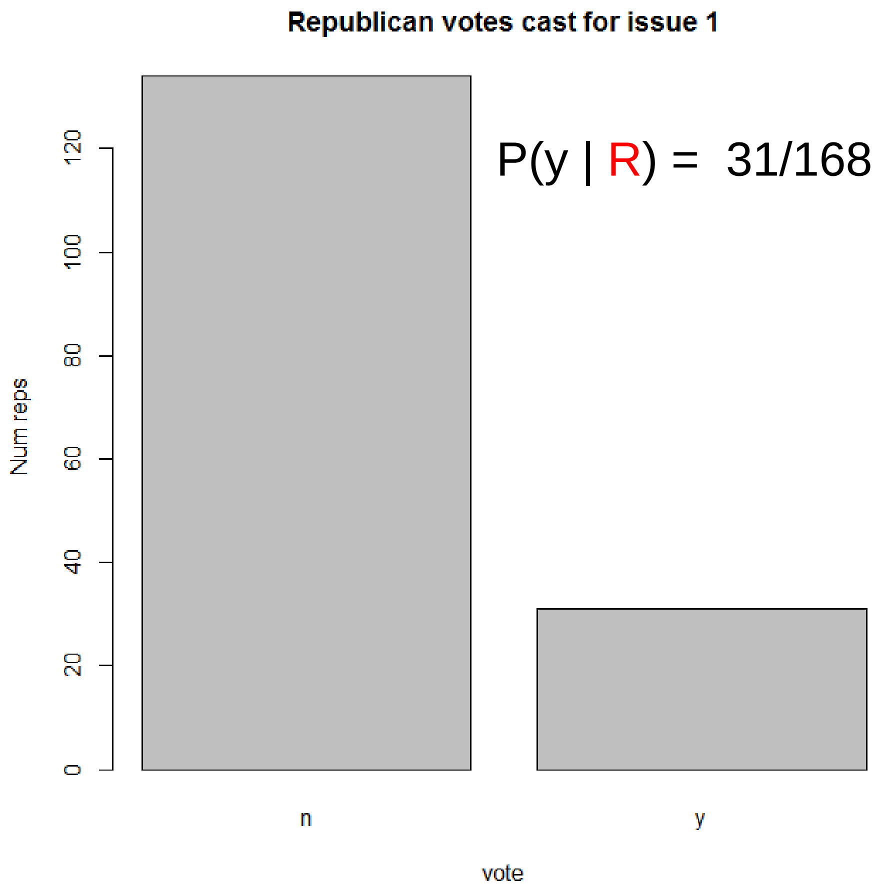
Prior Belief - Simplest Solution

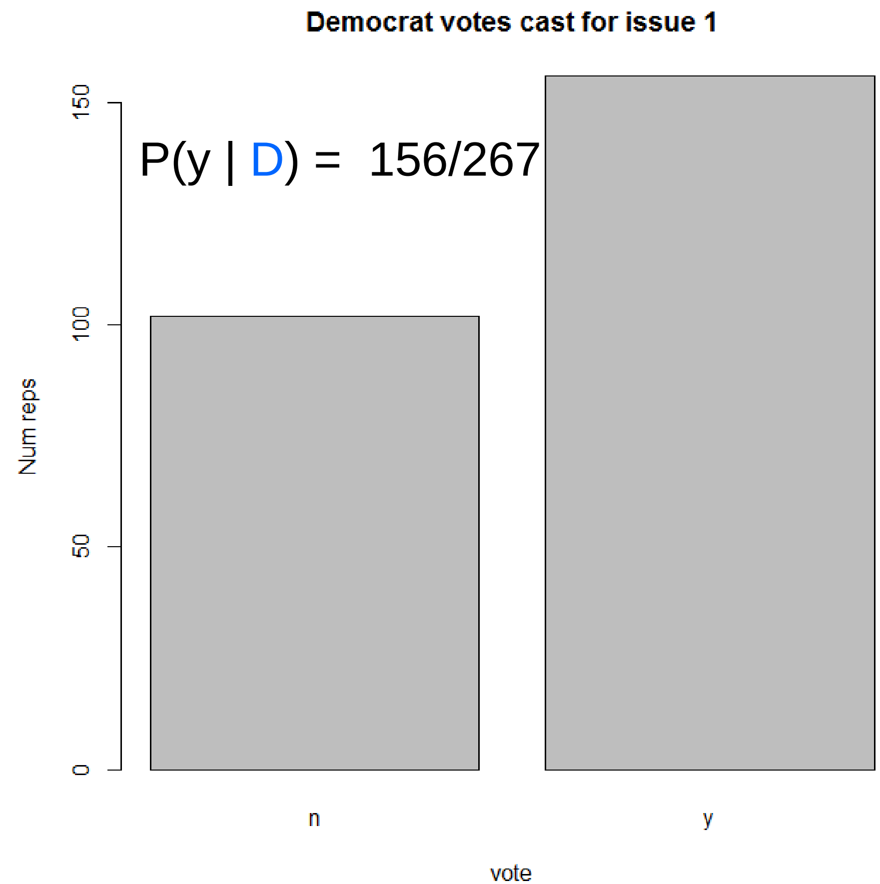
- The house has a majority of Democrats
  - 168 Republicans, 267 Democrats
- Probability of a random person being Democrat is
  - P(D) = 267/435 = 0.61
- Can we do better by incorporating the evidence of their voting patterns?

Voting Patterns for V1

Handicapped Infants. The vote failed to pass: 236 to 187

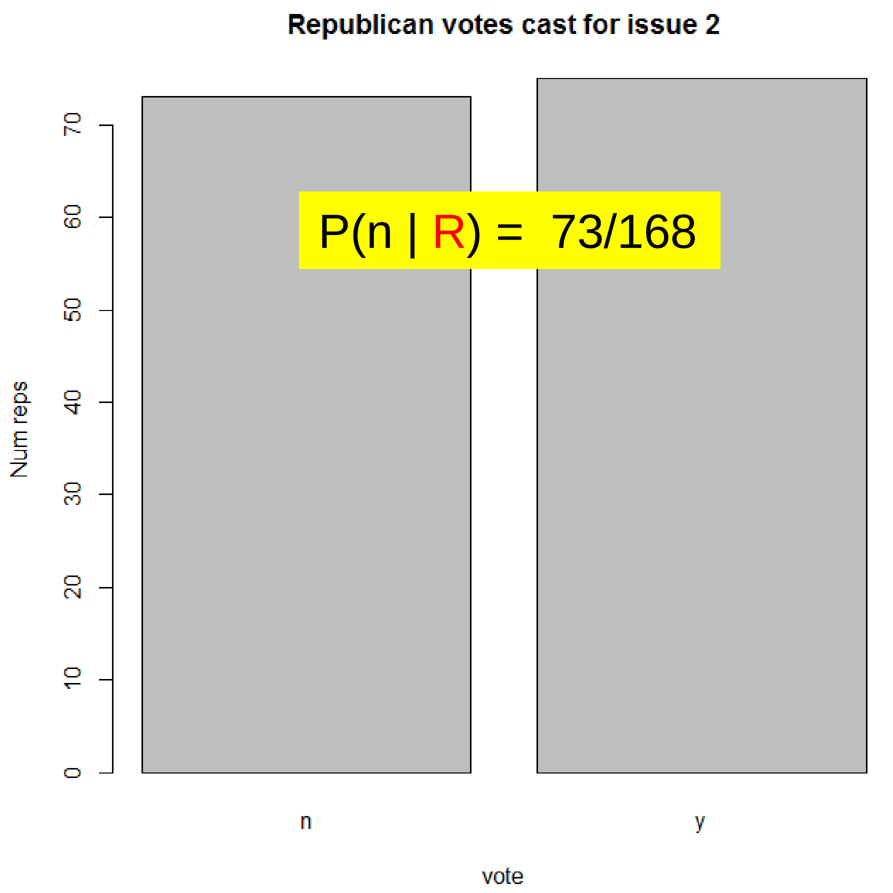
```
> Repub <-  HouseVotes84$Class=="republican"
> Democrat <- HouseVotes84$Class=="democrat"
> plot(as.factor(HouseVotes84[Repub,2]))
> title(main="Republican votes cast for issue 1",
xlab="vote", ylab="Num reps")
> plot(as.factor(HouseVotes84[Democrat,2]))
> title(main="Democrat votes cast for issue 1", x
lab="vote", ylab="Num reps")
```

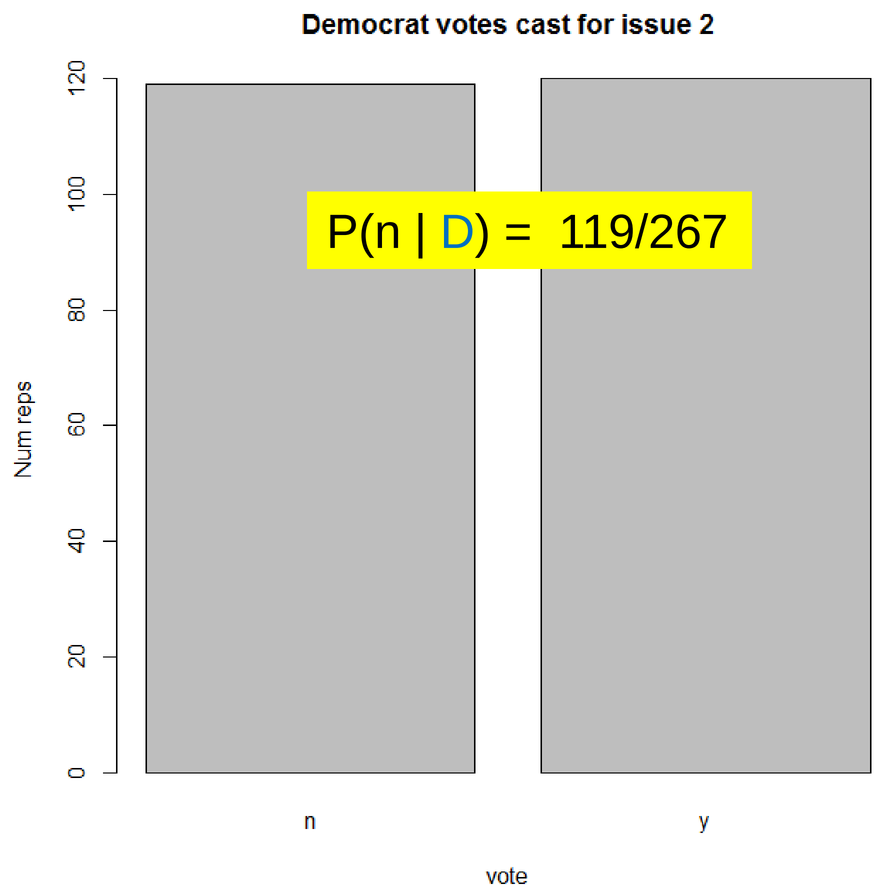




Voting Patterns for V2

Water-project-cost-sharing. The vote passed: 195 to 192





CSE 7302c

Bayes Theorem

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

$$P(D|\text{v1} = y, \text{v2} = n) = ?$$

$$P(D|\text{v1} = y, \text{v2} = n) = \frac{P(D) * P(\text{v1} = y, \text{v2} = n|D)}{P(\text{v1} = y, \text{v2} = n)}$$

CSE 7302c

Naïve Bayes

Naïve Assumption: *Conditional probability of each feature given the class, is independent of all other features*

$$P(\text{v1} = y, \text{v2}=n \mid D) \quad = \quad P(\text{v1} = y \mid D) * P(\text{v2} = n \mid D)$$

$$P(D \mid \text{v1} = y, \text{v2}=n) = \frac{P(D) * P(\text{v1} = y \mid D) * P(\text{v2} = n \mid D)}{P(\text{v1} = y, \text{v2}=n)}$$

Naïve Bayes

We are trying to decide, given the voting pattern, if that person is a Democrat or a Republican.

$$P(D|\text{v1} = y, \text{v2}=n) = \frac{P(D) * P(\text{v1} = y|D) * P(\text{v2} = n|D)}{P(\text{v1} = y, \text{v2}=n)}$$

$$P(R|\text{v1} = y, \text{v2}=n) = \frac{P(R) * P(\text{v1} = y|R) * P(\text{v2} = n|R)}{P(\text{v1} = y, \text{v2}=n)}$$

Whichever probability is higher, we would classify the person into that party.

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

Naïve Bayes

$$P(D|\text{v1} = y, \text{v2}=n) \propto P(D) * P(\text{v1} = y|D) * P(\text{v2} = n|D)$$

P(D) = 267/435 (267 Democrats among 435 Congressmen)

$$P(D|\text{v1} = y, \text{v2}=n) \propto \frac{267}{435} * \frac{156}{267} * \frac{119}{267} = 0.15$$

From voting pattern [slide]

$$P(R|\text{v1} = y, \text{v2}=n) \propto \frac{168}{435} * \frac{31}{168} * \frac{73}{168} = 0.03$$

Since the conditional probability for being Democrat is higher, he is likely to be Democrat.

Naïve Bayes: Voting patterns

```r
library(e1071)
nb_model <- naiveBayes(Class~.,data = trainHouseVotes84)
```

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
  democrat republican
 0.6111111  0.3888889

Conditional probabilities:
            V1
Y                    n         y
  democrat   0.4066986 0.5933014
  republican 0.8195489 0.1804511

            V2
Y                    n         y
  democrat   0.5119617 0.4880383
  republican 0.4586466 0.5413534
```

CSE 7302c

Naïve Bayes Assumption

- The key assumption of independence of features, is almost never true

- Still Naïve Bayes does surprisingly well in a lot of situations

- It works best when all the predictor variables are categorical variables

- Very frequently used in text mining, character image analysis problems

Evaluating Model Accuracy

BIAS-VARIANCE TRADEOFF

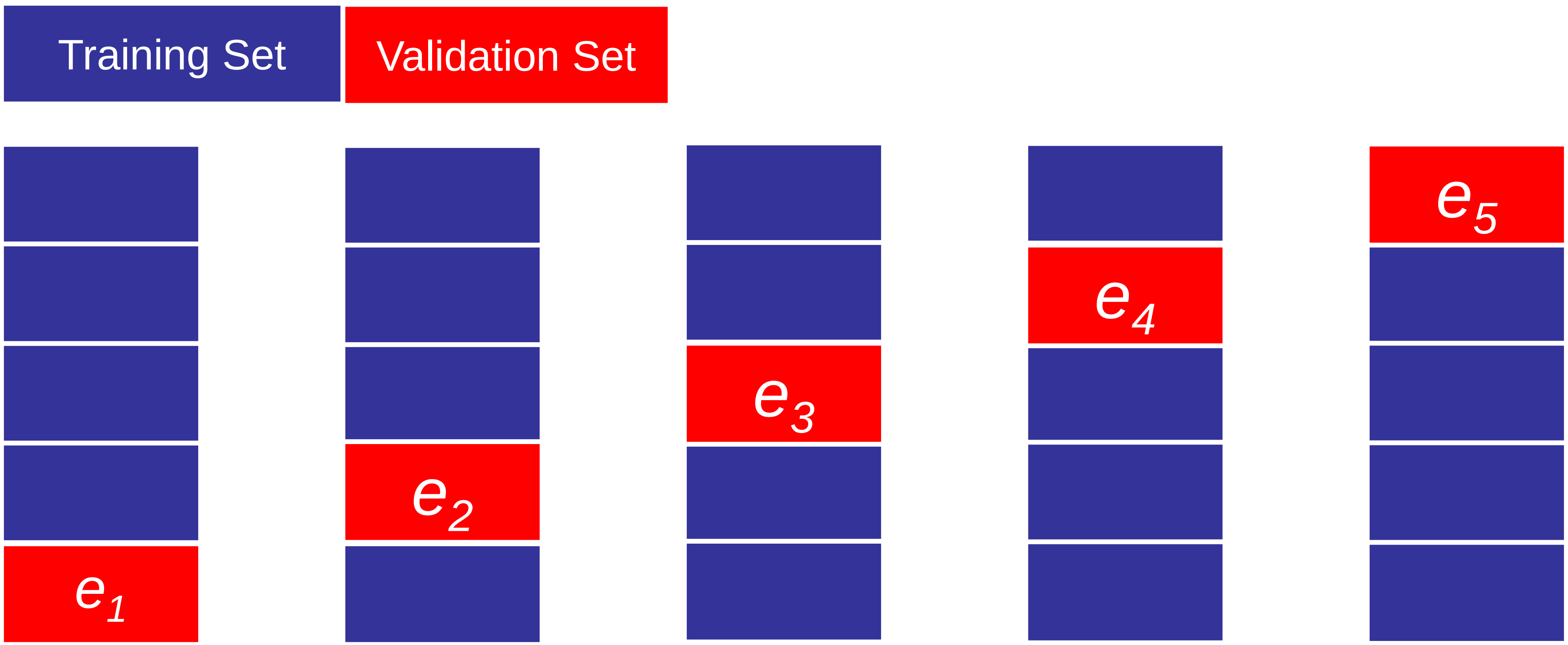
CSE 7302c

The Ultimate Test of Model Accuracy

- Holdout set: Split data into train, validation and test sets (in 70:20:10 or 60:20:20, etc. ratios), and **ensure model performance is similar**.
  - Training Set: For fitting a model
  - Validation Set: For selecting a model based on estimated prediction errors
  - Test Set: For assessing selected model's performance on "new" data

k-fold Cross-Validation

Common values of *k* are 5 to 10.



k-fold Cross-Validation

A good model will have

- a small mean of the errors (low bias, i.e., the model accurately captures the behaviour of the data), and
- a small standard deviation of the errors (low variance, i.e., error does not vary much based on the choice of the dataset)

Appropriate Error Measures for Evaluating Model Accuracy

- Use accurate measures of prediction error, experiment with different models and use the model with minimum error.
- Some measures for comparing models within the same technique (e.g., Linear Regression):
  - R^2
  - AIC

Appropriate Error Measures for Evaluating Model Accuracy

Some measures for comparing models across techniques:

- MAE (Mean Absolute Error): Mean of the absolute value of the difference between the predicted and actual values.
- MAPE (Mean Absolute Percentage Error): Same as above but converted into percentages to allow for comparison across different scales (e.g., comparing accuracies of forecasts on BSE vs NSE).
- RMSE (Root Mean Square Error): Accounts for infrequent large errors, whose impact may be understated by the mean-based error measures.

Bias-Variance Tradeoff

- Total error is composed of Bias, Variance and a Random irreducible error. Bias and Variance can be managed.

- If the model performance on training and testing data sets is inconsistent, it indicates a problem either with Bias or Variance.

Bias-Variance Tradeoff

- Bias arises when you make assumptions preventing you from finding relevant relationships between inputs (independent variables) and outputs (dependent variable). This causes the model to ***underfit*** the data. For example, assuming linearity when there is non-linearity in the data.

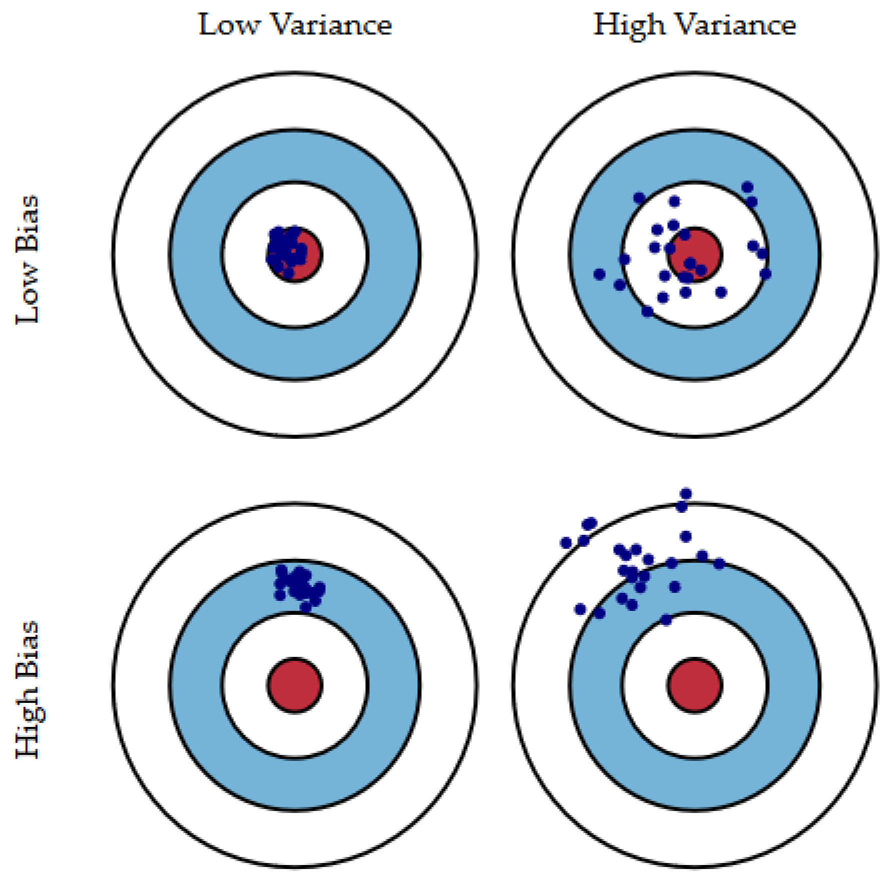
- Variance arises due to the model being overly sensitive to small fluctuations in the training data. Such a model ***overfits*** the data, including the random noise rather than just the actual behaviour.

Bias-Variance Tradeoff

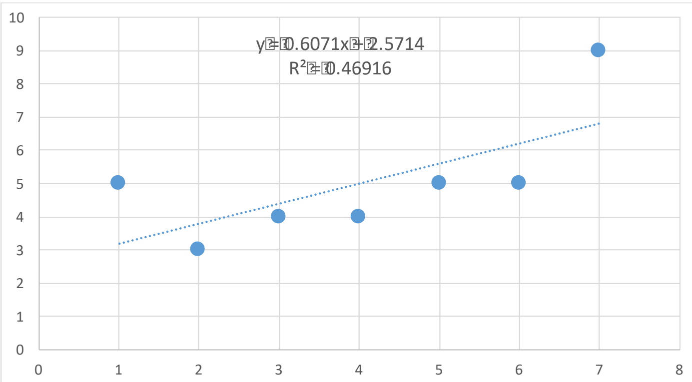
- An ideal model will **both** capture the patterns in the training data **and** generalize well enough to the unseen (testing) data.
- Unfortunately, it is generally impossible to do both and hence the tradeoff.
- All supervised models (classification, regression, etc.) are affected by this.

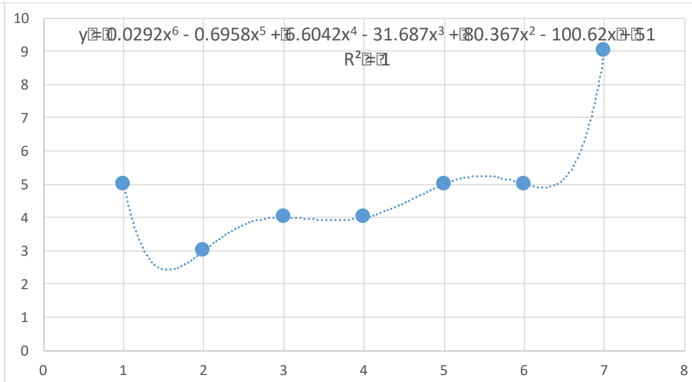
Bias-Variance Tradeoff

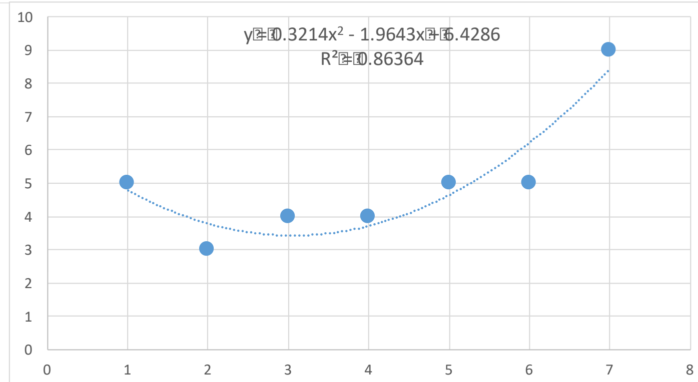
- Bulls-eye is a model that correctly predicts the real values.
- Each hit is a model based on chance variability in training datasets.



Bias-Variance Tradeoff and Underfitting vs Overfitting Excel

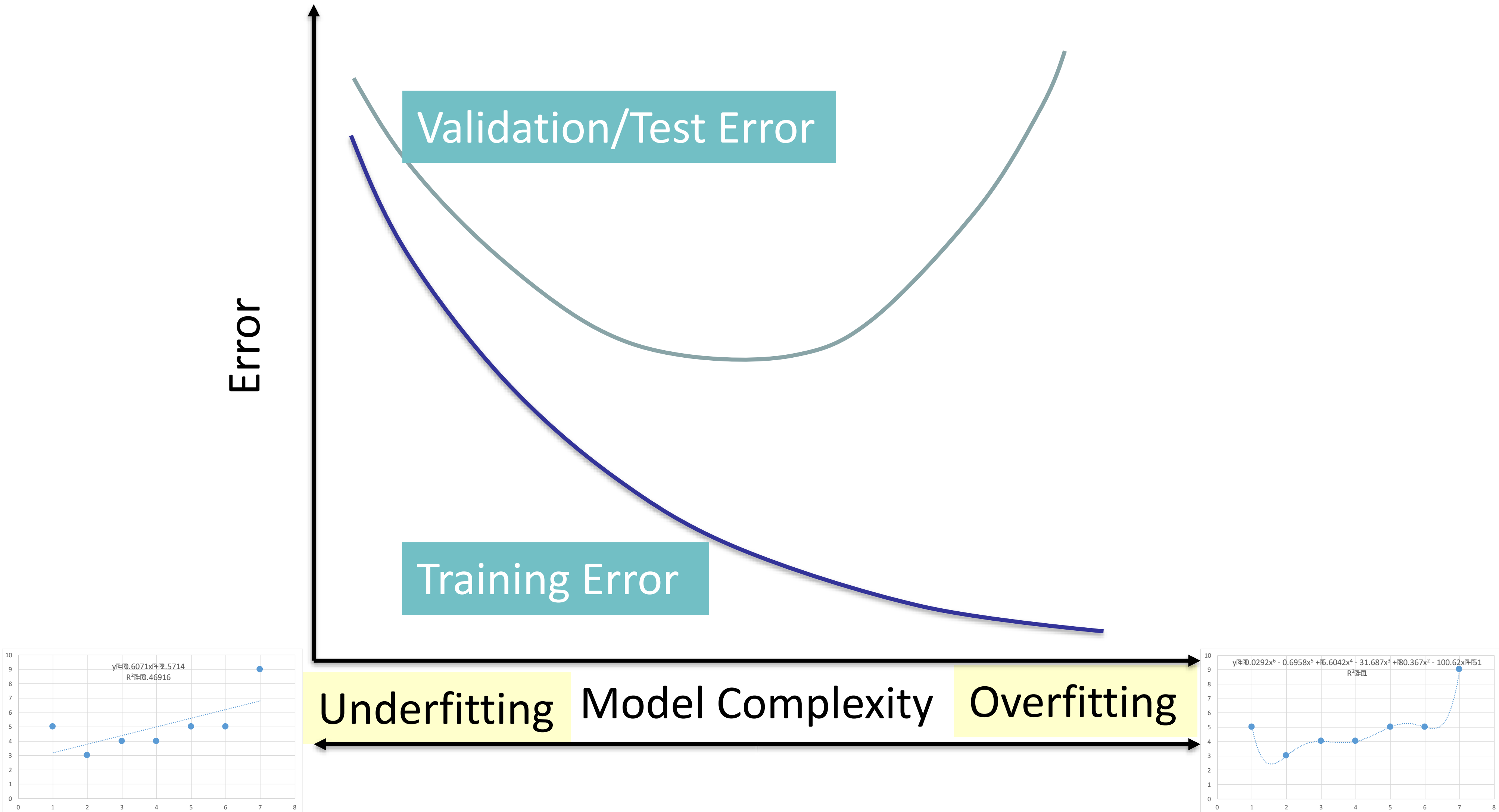




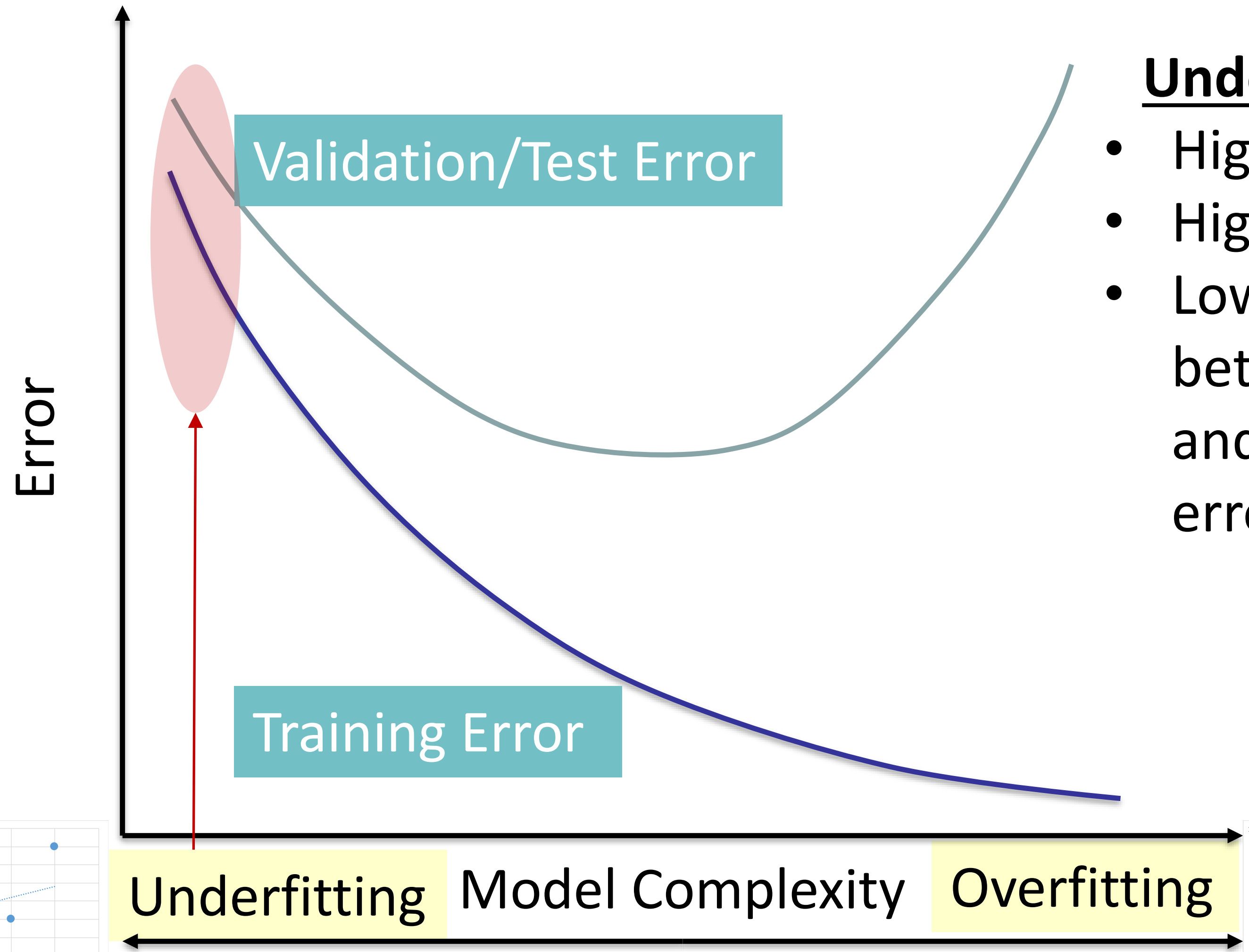


Too Simple a Model Underfit	Too Complex a Model Overfit	Right Model Reasonable fit

Bias-Variance Tradeoff and Underfitting vs Overfitting

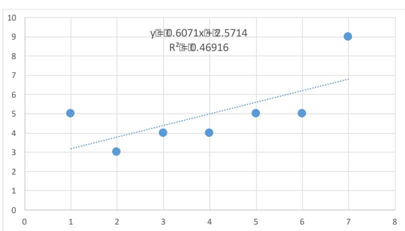


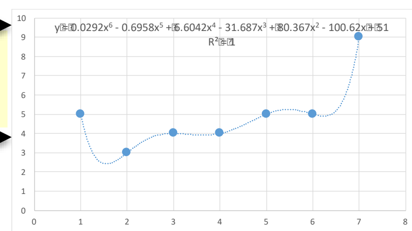
Diagnosing Bias and Variance



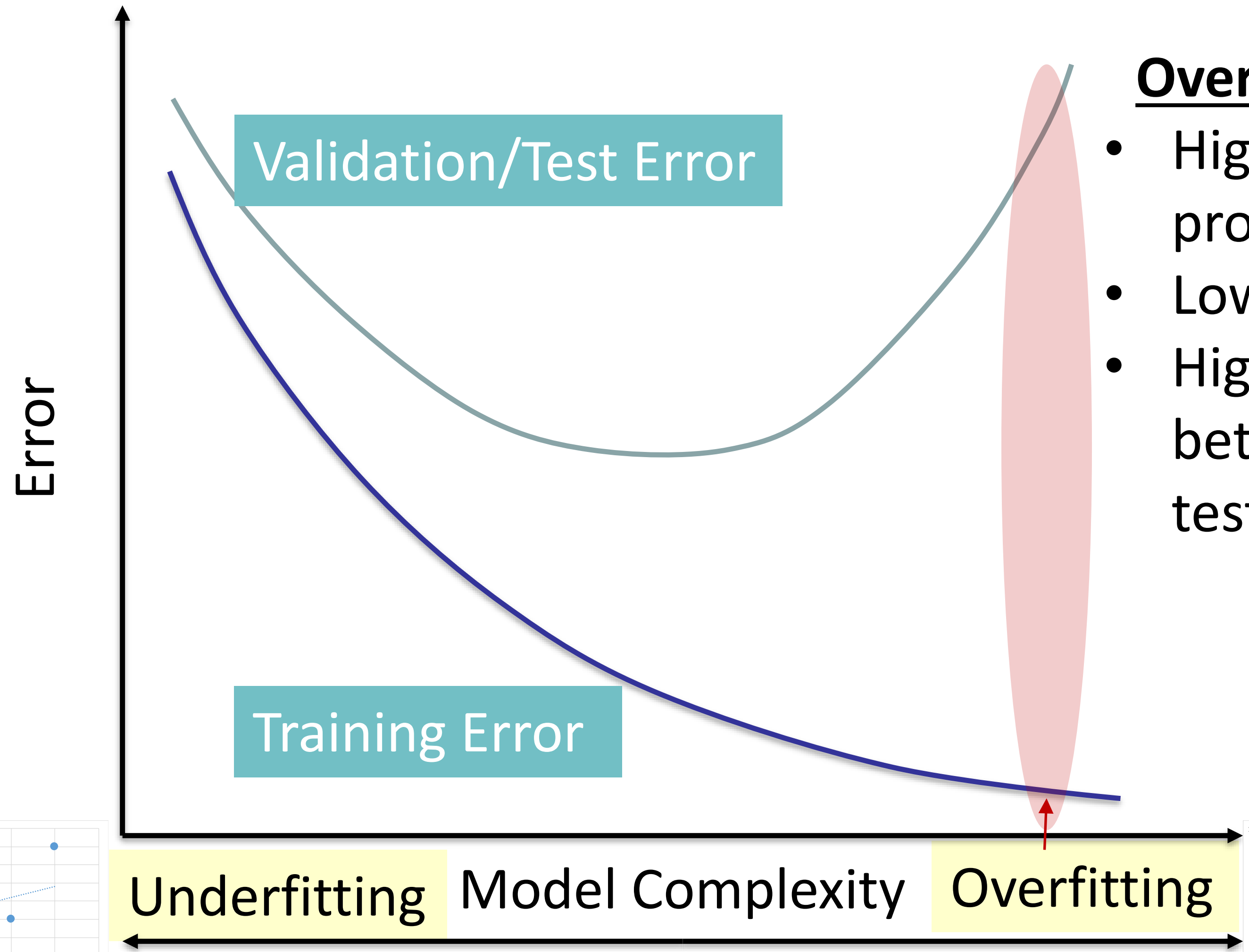
Underfitting (Bias)

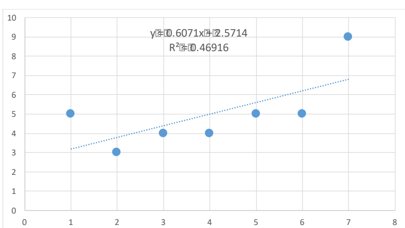
- High Bias problem
- High training error
- Low difference between training and test/validation errors

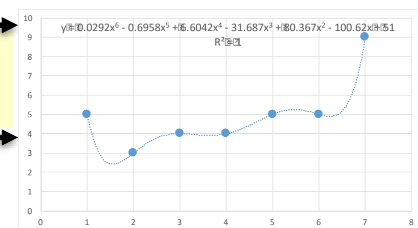




Diagnosing Bias and Variance







<u>Overfitting (Variance)</u>

- High Variance problem
- Low training error
- High difference between training and test/validation errors

Bias-Variance Tradeoff

Ways of detecting and minimizing Bias and Variance

- Outliers and Influential Observations can cause statistical bias. Can be identified using various methods like Box plots, points outside $\pm 2\ or\ \pm 3$ standard deviations/errors, residual plots, etc.

- Bias cannot be corrected by increasing training sample size.

- Adding features (independent variables or predictors) tends to decrease bias.

- Variance or standard error can be minimized by increasing training sample size.

- Dimensionality reduction and feature selection methods decrease variance.

Bias-Variance Tradeoff

Ways of detecting and minimizing Bias and Variance

Parameters can be tuned in supervised models to control bias and variance:

- Regularization decreases variance at the cost of increasing bias. It can be applied to a variety of techniques (not just linear models).
- In Artificial Neural Networks, bias decreases at the cost of increasing variance with addition of hidden units.
- In kNN, increasing k lowers variance at the cost of increasing bias.
- In Decision Trees, increasing the length of the tree increases variance. Pruning is used to control variance.

Ref: https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff
Last accessed: July 08, 2017

Bias-Variance Tradeoff

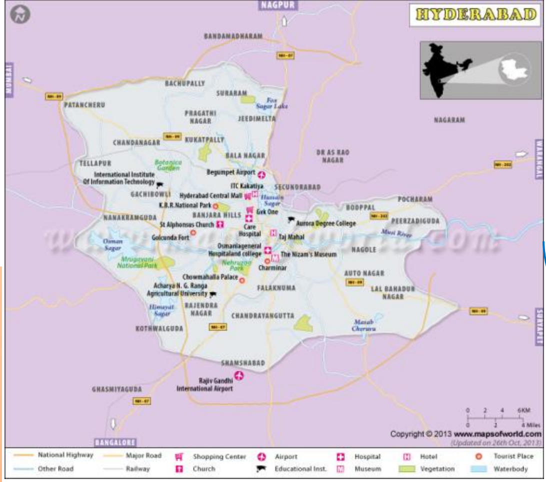
Ways of detecting and minimizing Bias and Variance

Ensemble models help resolve the tradeoff *(taught later in the program)* .

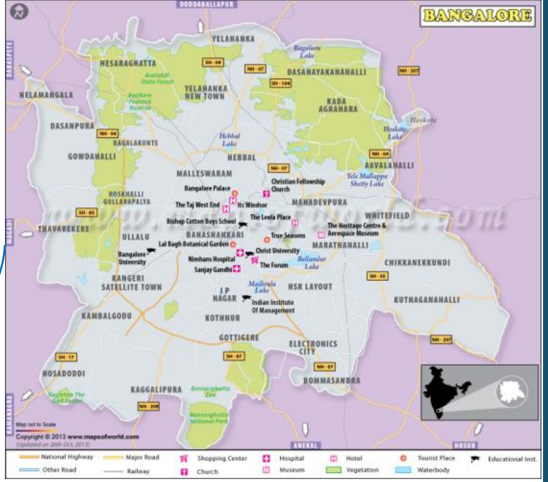
- Boosting methods combine many “weak” (high bias) models in an ensemble that lowers bias compared to individual models.

- Bagging (bootstrap aggregating) techniques combine “strong” models to minimize variance.

Ref: https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff
Last accessed: July 08, 2017



Inspire...Educate...Transform.



HYDERABAD

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

BENGALURU

Floors 1-3, L77, 15th Cross Road, 3A Main Road, Sector 6, HSR Layout, Bengaluru – 560 102
+91-9502334561 (Individuals)
+91-9502799088 (Corporates)

Social Media

Web:	http://www.insofe.edu.in
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	http://www.youtube.com/InsofeVideos
SlideShare:	http://www.slideshare.net/INSOFE
LinkedIn:	http://www.linkedin.com/company/international-school-of-engineering

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.