




Inspire…Educate…Transform.

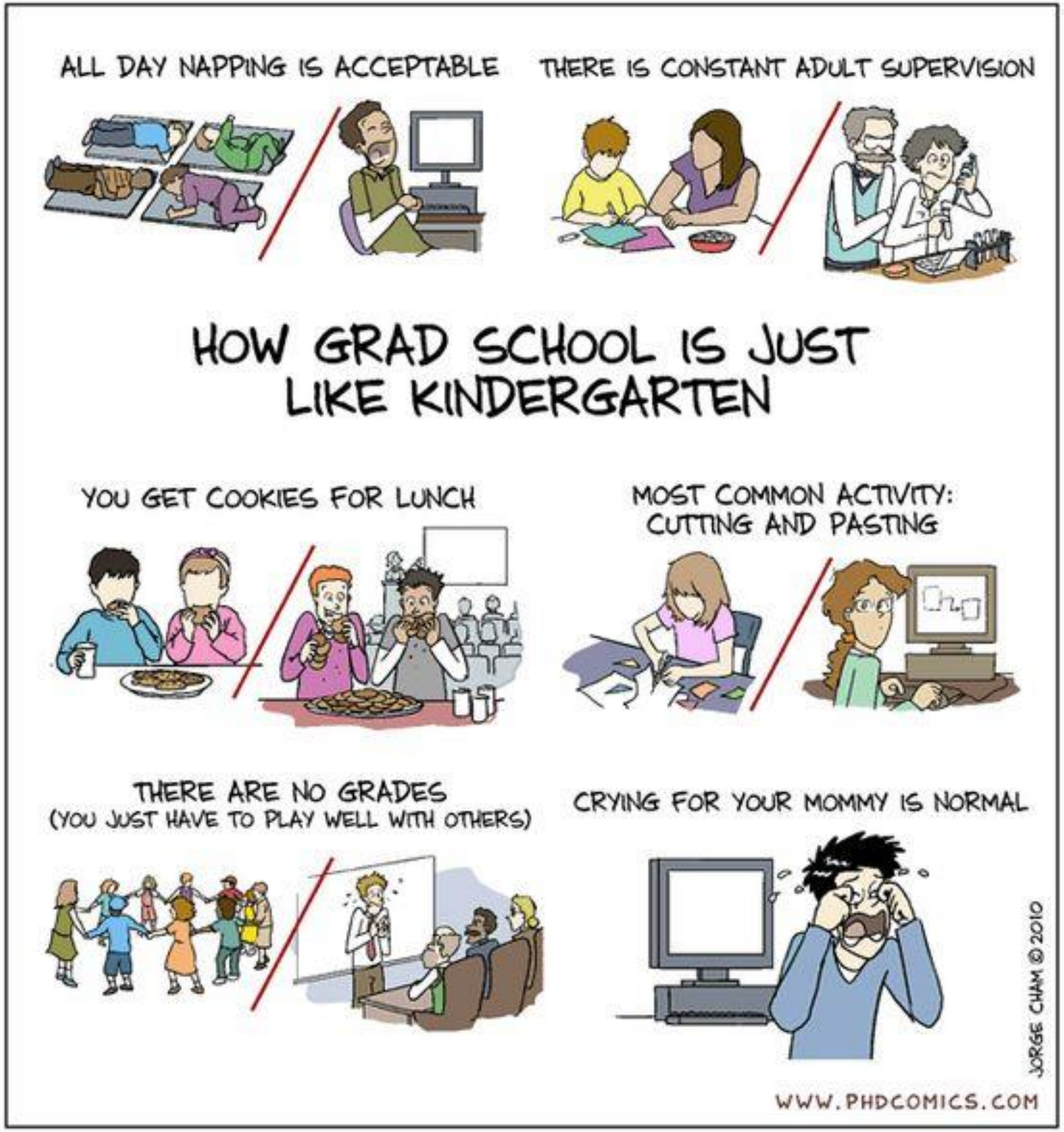
Supervised models

Logistic Regression

Dr. Anand Jayaraman
anand.jayaraman@insofe.edu.in

Apr 30, 2017

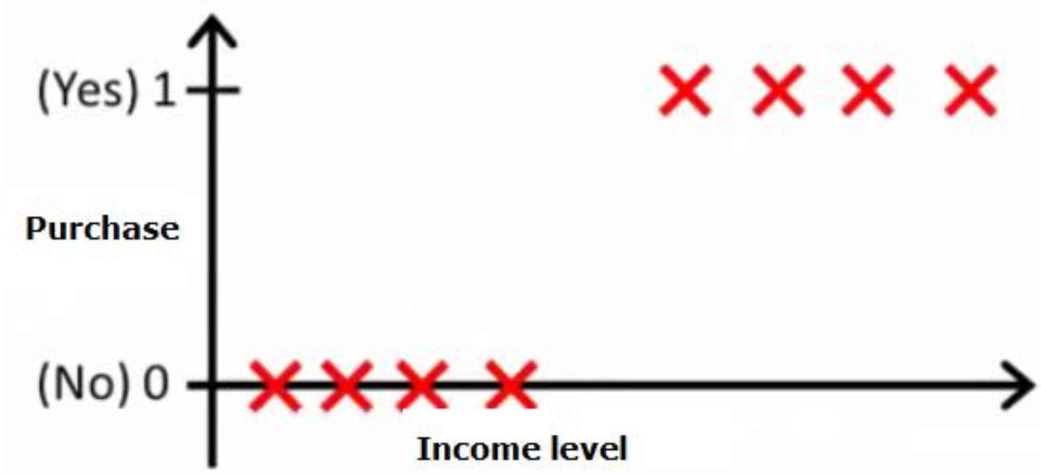
Thanks to Dr.Sridhar Pappu for the material


ALL DAY NAPPING IS ACCEPTABLE
THERE IS CONSTANT ADULT SUPERVISION
HOW GRAD SCHOOL IS JUST LIKE KINDERGARTEN
YOU GET COOKIES FOR LUNCH
MOST COMMON ACTIVITY: CUTTING AND PASTING
THERE ARE NO GRADES
(YOU JUST HAVE TO PLAY WELL WITH OTHERS)
CRYING FOR YOUR MOMMY IS NORMAL
JORGE CHAM © 2010
WWW.PHDCOMICS.COM

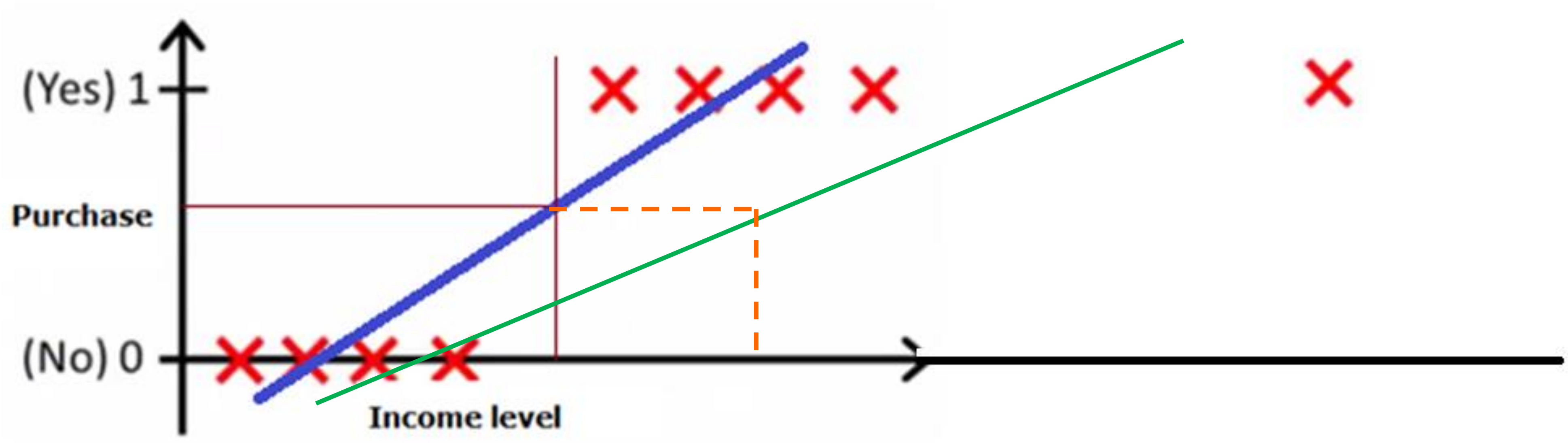
CSE 7202c

LOGISTIC REGRESSION

Classification Tasks: Regression



It could fail



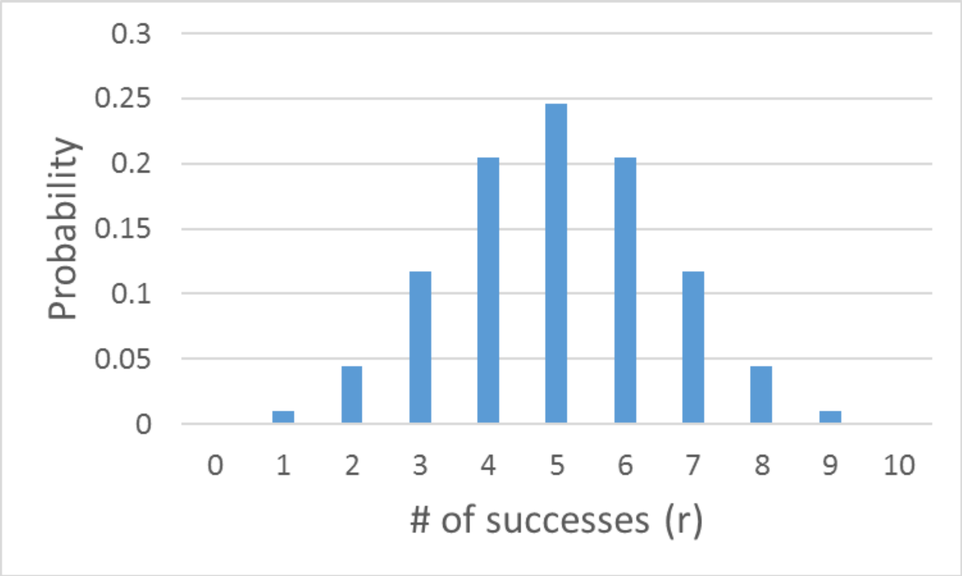
- Linear regression slopes can be much larger than 1 or much smaller than zero and hence thresholding becomes difficult.

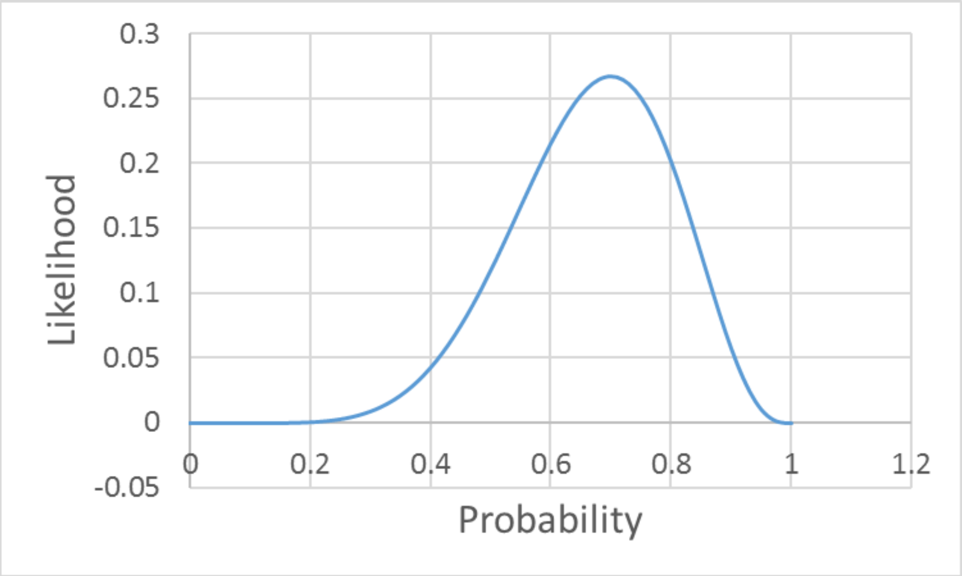
- Error terms do not follow normal distribution.
- Error terms are not independent.
- Error variances are heteroscedastic.

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

Probability vs Likelihood - Excel

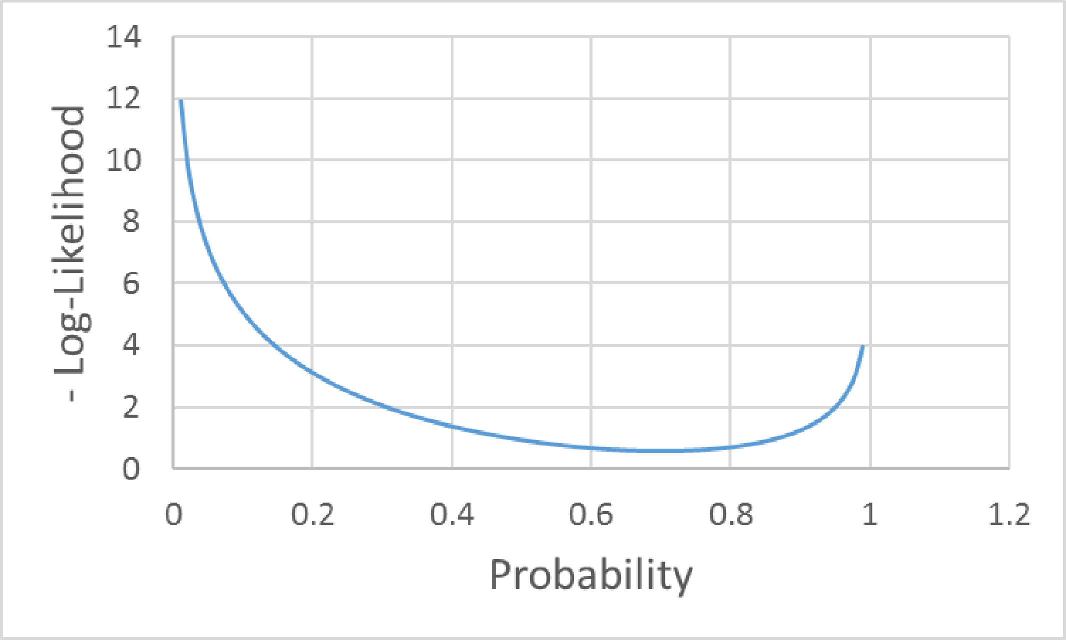
- Likelihood is also known as reverse probability.
- In Probability, we **predict data** based on **known parameters**. *(Recall B(n,p), Geo(p), Po(λ), N(μ,σ^2), etc.)*
- In Likelihood, we **predict parameters** based on **known data**.





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.

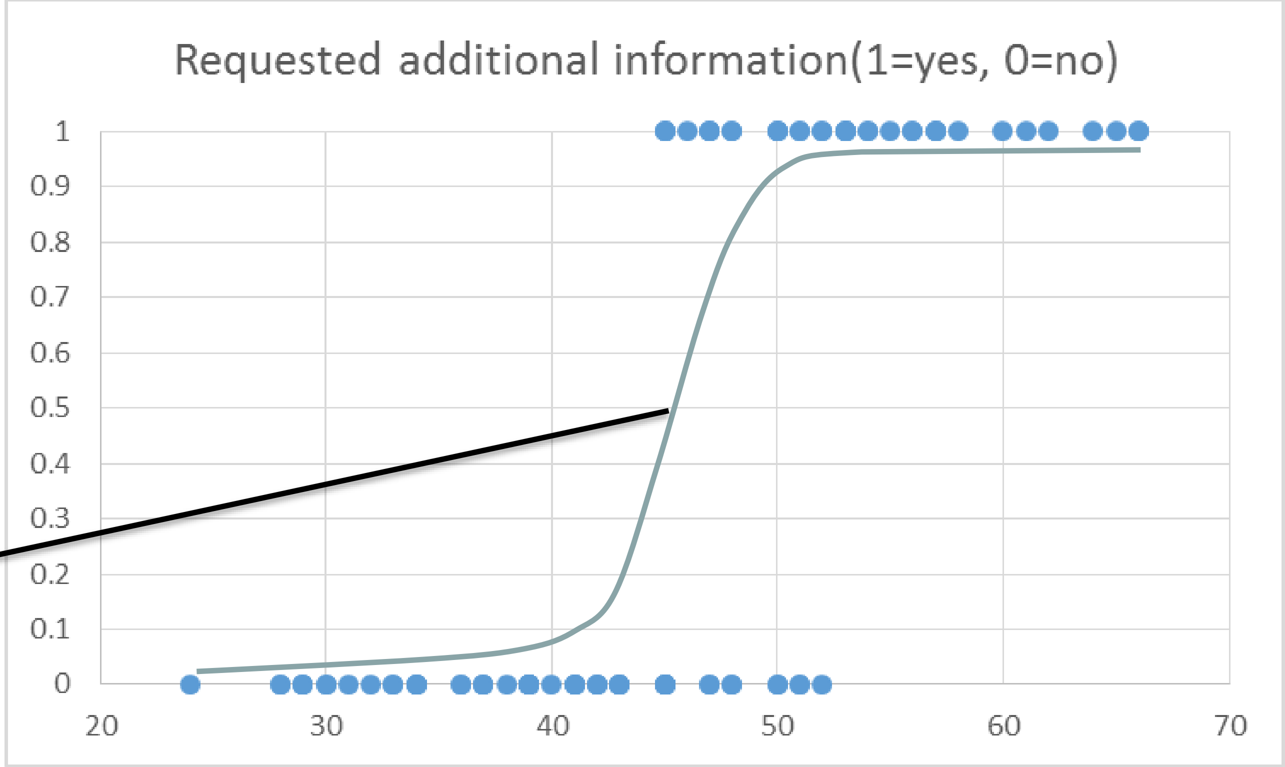


Example

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

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

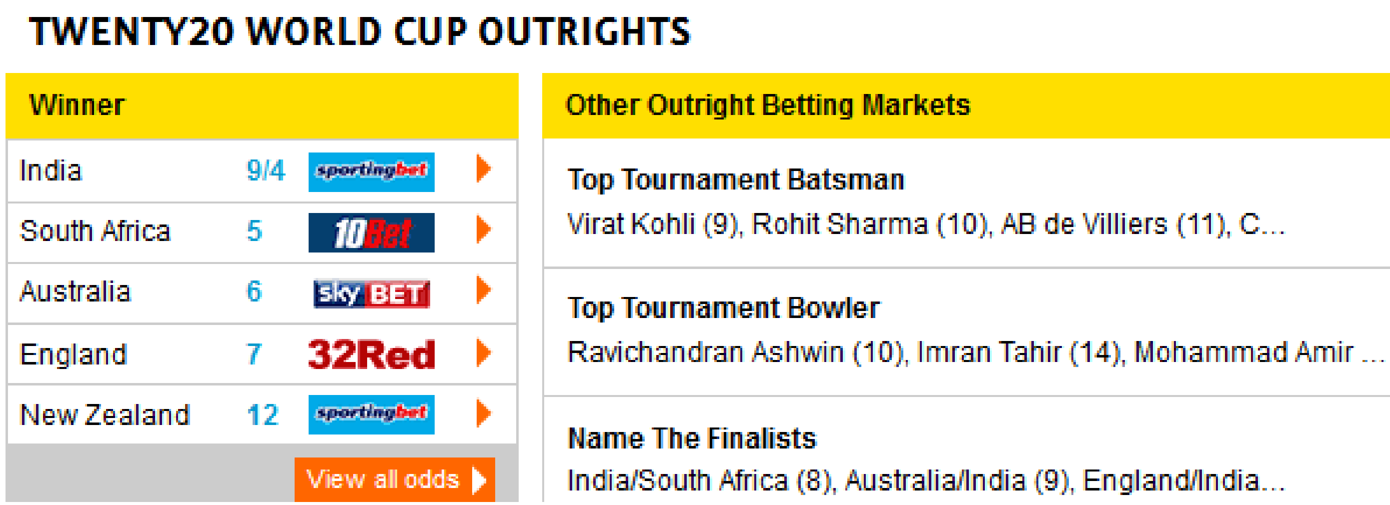
Odds Ratio is obtained by the probability of an event occurring divided by the probability that it will not occur.

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

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

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

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

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

Logistic model

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

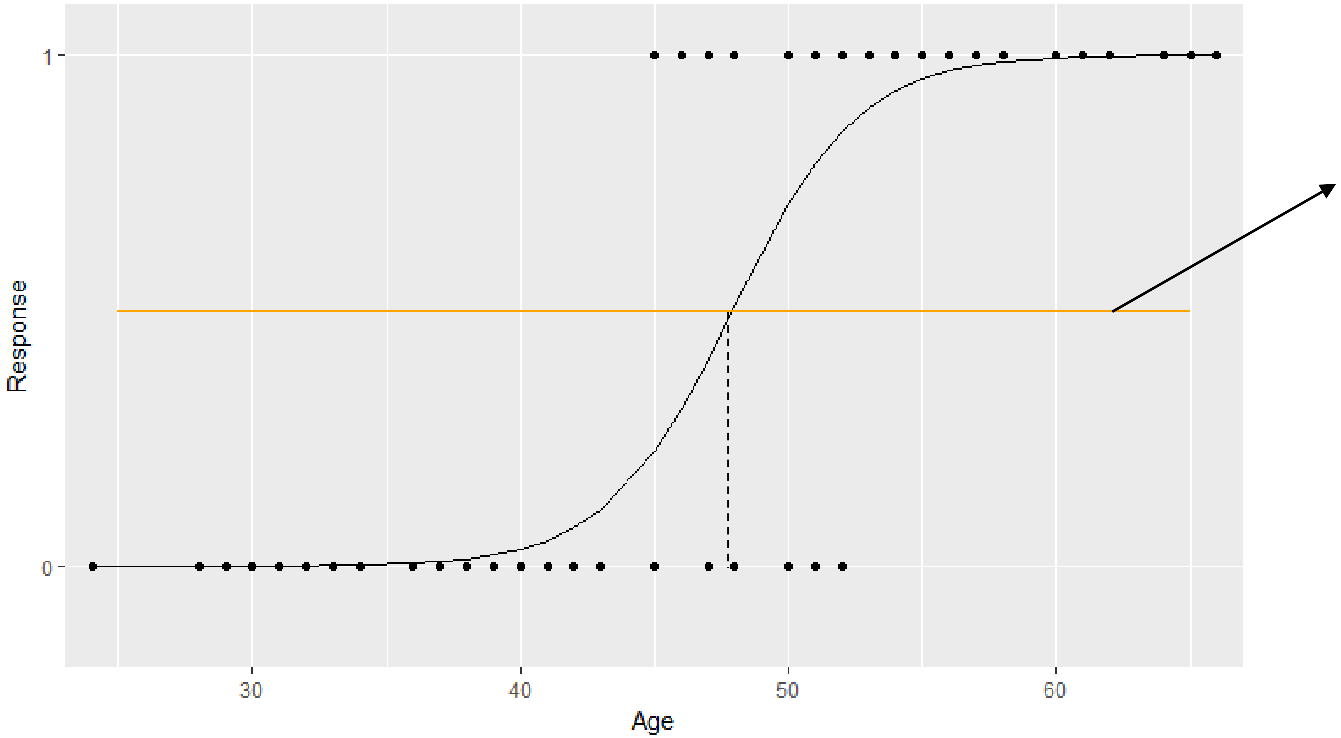
R and Interpreting the output

```
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?

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

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

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
```

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.

```
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.
- 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

Logistic Regression -Pseudo R^2

- Note that R^2 is not defined for Logistic Regressions

- McFadden's Pseudo R^2

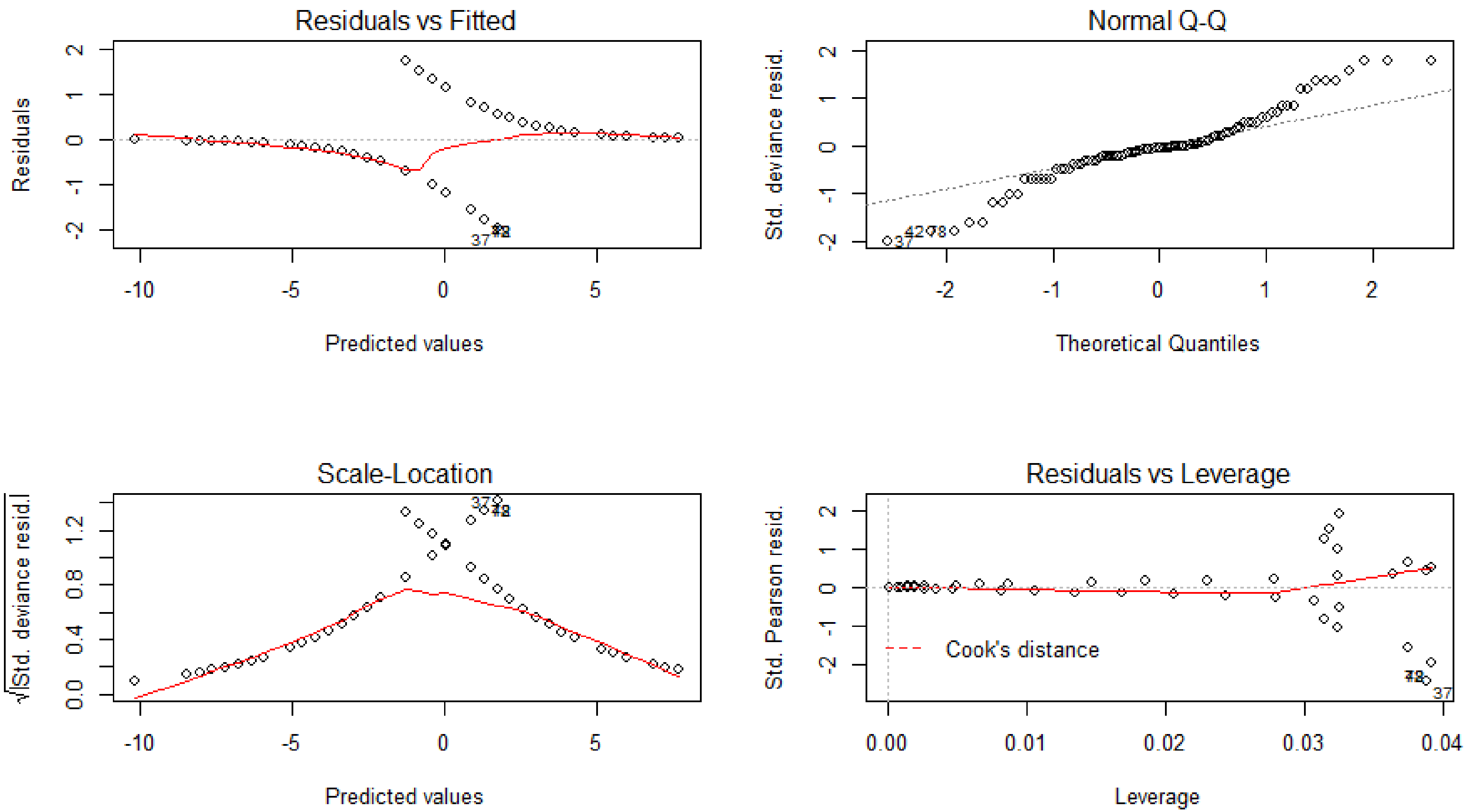
- Pseudo R^2 = 1 - $\frac{Residual\ Dev}{Null\ Dev}$

```
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
```

$$\text{Pseudo } R^2 = 1 - \frac{49.937}{123.156} = 0.59$$

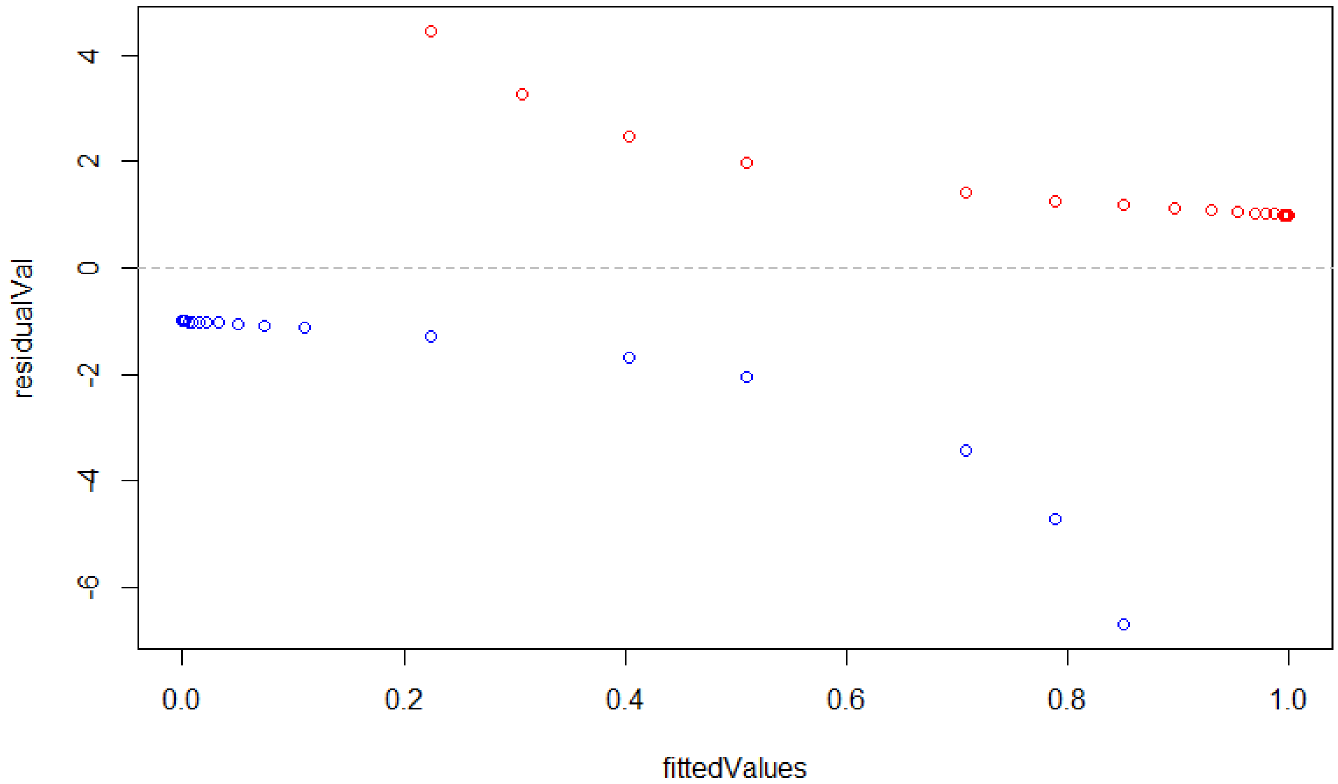
Caution: This Pseudo R-Squared does not have the same interpretation as in Least Squares Regression and is rarely used.

Residual Plots



Why does the *Residuals vs Fitted* graph show a two-line pattern?

Understanding Residual Plot




```
> plot(fittedValues,residualVal,col=c("blue","red")[ActualResponse])
> abline(h=0,lty=2,col="grey")
```

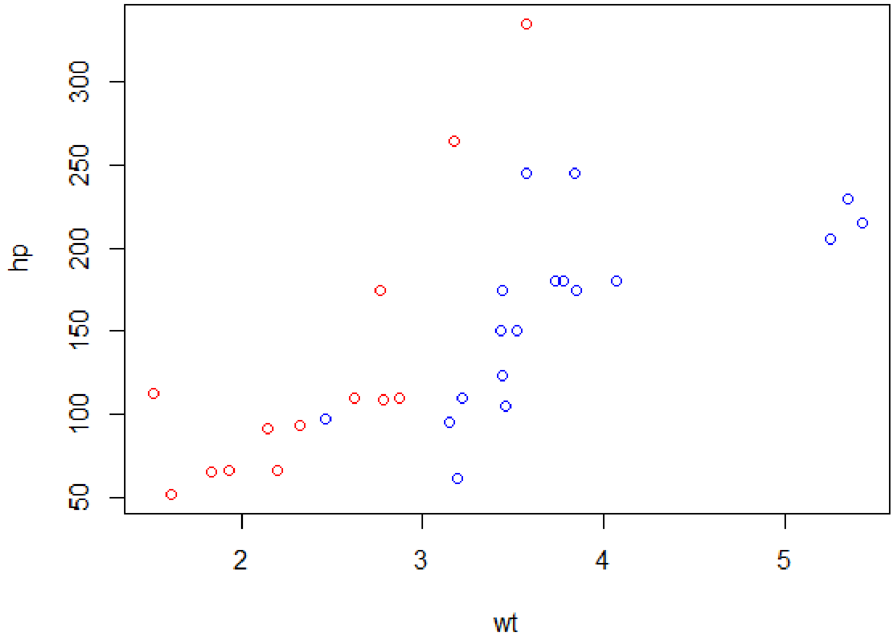
Example: Automatic or Manual Transmission

model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1
Duster 360	14.3	8	360	245	3.21	3.57	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.19	20	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.44	18.9	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.07	17.4	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.73	17.6	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.78	18	0	0	3	3
Cadillac Fleetwood	10.4	8	472	205	2.93	5.25	17.98	0	0	3	4
Lincoln Continental	10.4	8	460	215	3	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.2	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.9	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.7	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318	150	2.76	3.52	16.87	0	0	3	2
AMC Javelin	15.2	8	304	150	3.15	3.435	17.3	0	0	3	2
Camaro Z28	13.3	8	350	245	3.73	3.84	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79	66	4.08	1.935	18.9	1	1	4	1
Porsche 914-2	26	4	120.3	91	4.43	2.14	16.7	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.9	1	1	5	2
Ford Pantera L	15.8	8	351	264	4.22	3.17	14.5	0	1	5	4
Ferrari Dino	19.7	6	145	175	3.62	2.77	15.5	0	1	5	6
Maserati Bora	15	8	301	335	3.54	3.57	14.6	0	1	5	8
Volvo 142E	21.4	4	121	109	4.11	2.78	18.6	1	1	4	2

mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (1000 lbs)
qsec	1/4 mile time
vs	V/S
am	Transmission (0 = automatic, 1 =
gear	Number of forward gears
carb	Number of carburetors

Example: Automatic or Manual Transmission

Using the MTcars dataset, estimate the probability of a vehicle being fitted with a manual transmission if it has a 120hp engine and weights 2800 lbs.




```
> with(mtcars,plot(wt,hp, col=c("blue","red")[am+1]))
```

Example: Automatic or Manual Transmission

```
> mtOut <- glm(am ~wt+hp ,data=mtcars, family=binomial) #Lets Build model
> summary(mtOut) #Check its significance

Call:
glm(formula = am ~ wt + hp, family = binomial, data = mtcars)

Deviance Residuals:
    Min       1Q    Median       3Q      Max
-2.2537  -0.1568  -0.0168   0.1543   1.3449

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 18.86630    7.44356   2.535  0.01126 *
wt          -8.08348    3.06868  -2.634  0.00843 **
hp           0.03626    0.01773   2.044  0.04091 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43.230  on 31  degrees of freedom
Residual deviance: 10.059  on 29  degrees of freedom
AIC: 16.059

Number of Fisher Scoring iterations: 8

> #Now compute probability on the new data
> nd <- data.frame(hp=120, wt=2.8)
> predict(mtOut,newdata=nd,type="response")
        1
0.6418125
```

There is a 64% probability of the car being fitted with an Automatic transmission.

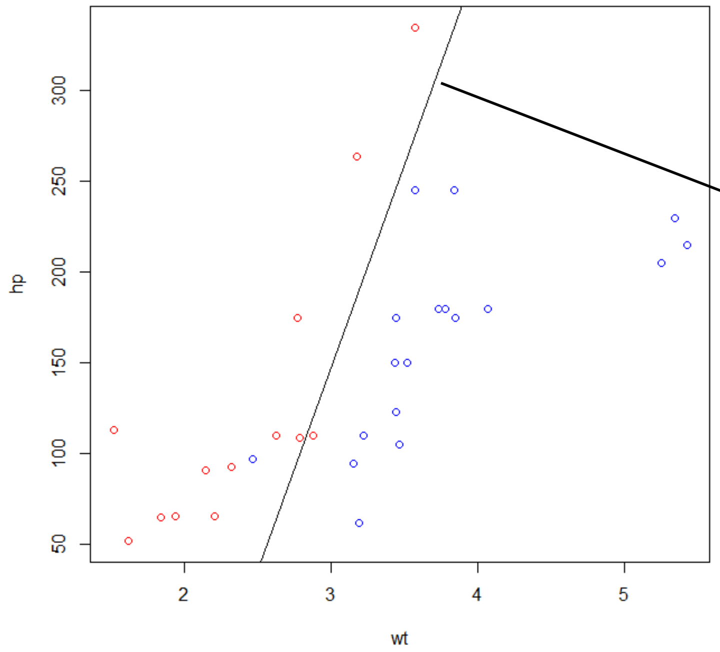
Example: Automatic or Manual Transmission

```
Call:
glm(formula = am ~ wt + hp, family = binomial, data = mtcars)

Deviance Residuals:
    Min        1Q    Median        3Q       Max
-2.2537   -0.1568   -0.0168    0.1543    1.3449

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 18.86630    7.44356   2.535  0.01126 *
wt          -8.08348    3.06868  -2.634  0.00843 **
hp           0.03626    0.01773   2.044  0.04091 *
---
```

Equation of the fit

Ln(S) = 18.8663 – 8.080348 wt + 0.03636 hp

Setting Ln(S) = 0 in the above equation gives the equation of dividing line between the two classes. This line marks the set of points for which prob=0.5



$$\boxed{0 = 18.8663 - 8.080348\ wt + 0.03636\ hp}$$

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 and R

 # 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?

```
Call:  glm(formula = y ~ job + marital + education + balance + housing +
    loan + contact + day + month + duration + campaign + previous +
    poutcome, family = "binomial", data = subscribetermdeposit)

Coefficients:
      (Intercept)      jobblue-collar     jobentrepreneur        jobhousemaid
       -2.555e+00          -3.103e-01          -3.573e-01          -5.028e-01
    jobmanagement          jobretired   jobself-employed         jobservices
       -1.652e-01           2.552e-01          -2.981e-01          -2.241e-01
       jobstudent       jobtechnician       jobunemployed          jobunknown
        3.819e-01          -1.758e-01          -1.771e-01          -3.124e-01
    maritalmarried       maritalsingle  educationsecondary   educationtertiary
       -1.792e-01           9.171e-02           1.832e-01           3.790e-01
  educationunknown             balance          housingyes             loanyes
        2.506e-01           1.289e-05          -6.767e-01          -4.259e-01
  contacttelephone      contactunknown                 day            monthaug
       -1.629e-01          -1.622e+00           9.976e-03          -6.931e-01
         monthdec            monthfeb            monthjan            monthjul
        6.920e-01          -1.458e-01          -1.260e+00          -8.305e-01
         monthjun            monthmar            monthmay            monthnov
        4.544e-01           1.591e+00          -4.001e-01          -8.706e-01
         monthoct            monthsep            duration            campaign
        8.828e-01           8.741e-01           4.194e-03          -9.082e-02
         previous       poutcomeother     poutcomesuccess     poutcomeunknown
        1.022e-02           2.049e-01           2.298e+00          -6.803e-02

Degrees of Freedom: 45210 Total (i.e. Null);  45171 Residual
Null Deviance:      32630
Residual Deviance: 21560   AIC: 21640
```

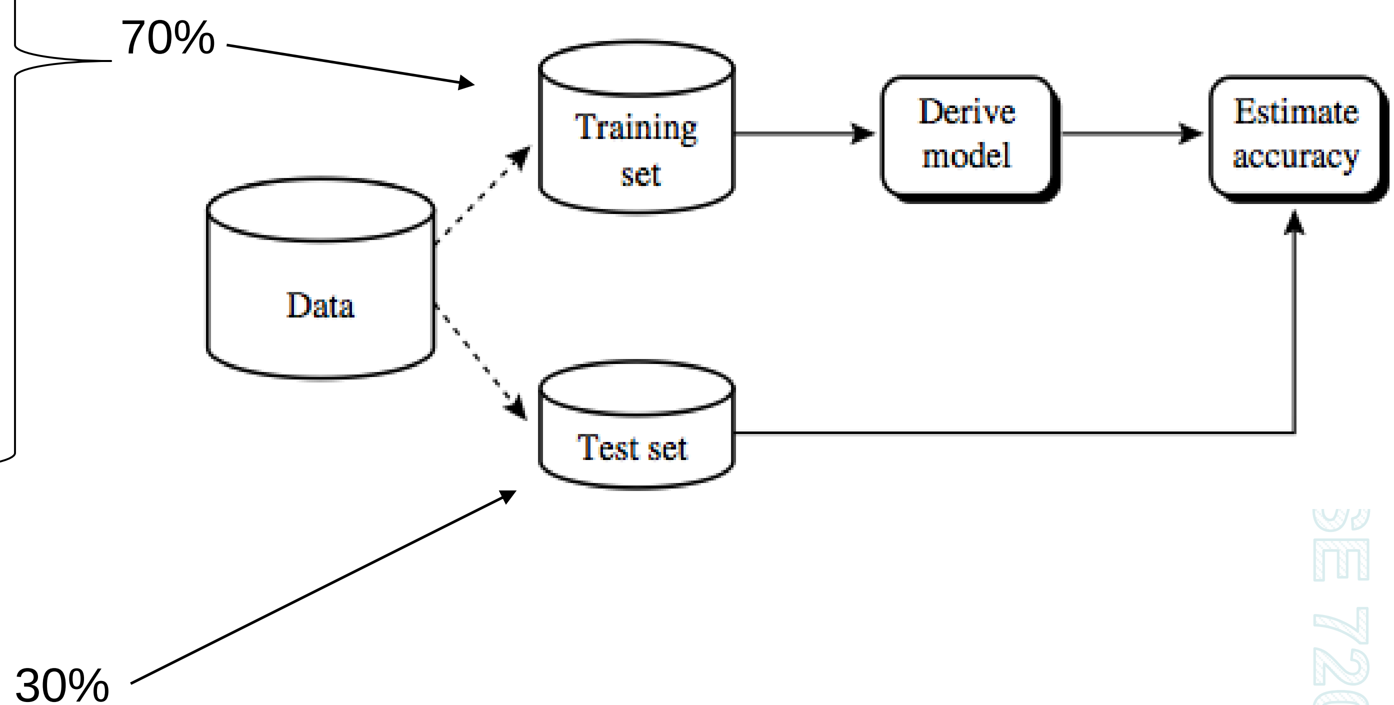
Applications

- 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

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

mpg	cyl	disp	hp	drat	wt
21	6	160	110	3.9	
21	6	160	110	3.9	
22.8	4	108	93	3.85	
21.4	6	258	110	3.08	
18.7	8	360	175	3.15	
18.1	6	225	105	2.76	
14.3	8	360	245	3.21	
24.4	4	146.7	62	3.69	
22.8	4	140.8	95	3.92	
19.2	6	167.6	123	3.92	
17.8	6	167.6	123	3.92	
16.4	8	275.8	180	3.07	
17.3	8	275.8	180	3.07	
15.2	8	275.8	180	3.07	
10.4	8	472	205	2.93	
10.4	8	460	215	3	
14.7	8	440	230	3.23	
32.4	4	78.7	66	4.08	
30.4	4	75.7	52	4.93	
33.9	4	71.1	65	4.22	
21.5	4	120.1	97	3.7	
15.5	8	318	150	2.76	
15.2	8	304	150	3.15	
13.3	8	350	245	3.73	
19.2	8	400	175	3.08	
27.3	4	79	66	4.08	
26	4	120.3	91	4.43	
30.4	4	95.1	113	3.77	
15.8	8	351	264	4.22	
19.7	6	145	175	3.62	
15	8	301	335	3.54	
21.4	4	121	109	4.11	



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 is 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	
		True	False
Actual	True	30	357
	False	9	2170

Testing Set

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

Some More Performance Measures for Regression and Classification Models

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

Kappa Metric

10-year CHD risk		Predicted	
		True	False
Actual	True	30	357
	False	9	2170

- Total= 30+357+9+2170=2566
- TotalAccuracy=(30+2170)/2566=0.857

- PercTrue=(30+357)/2566 = 0.15 ; PercFalse=(9+2170)/2566 = 0.85
- PredTrue=(30+9)/2566=0.015 ; PredFalse=(357+2170)/2566 = 0.985

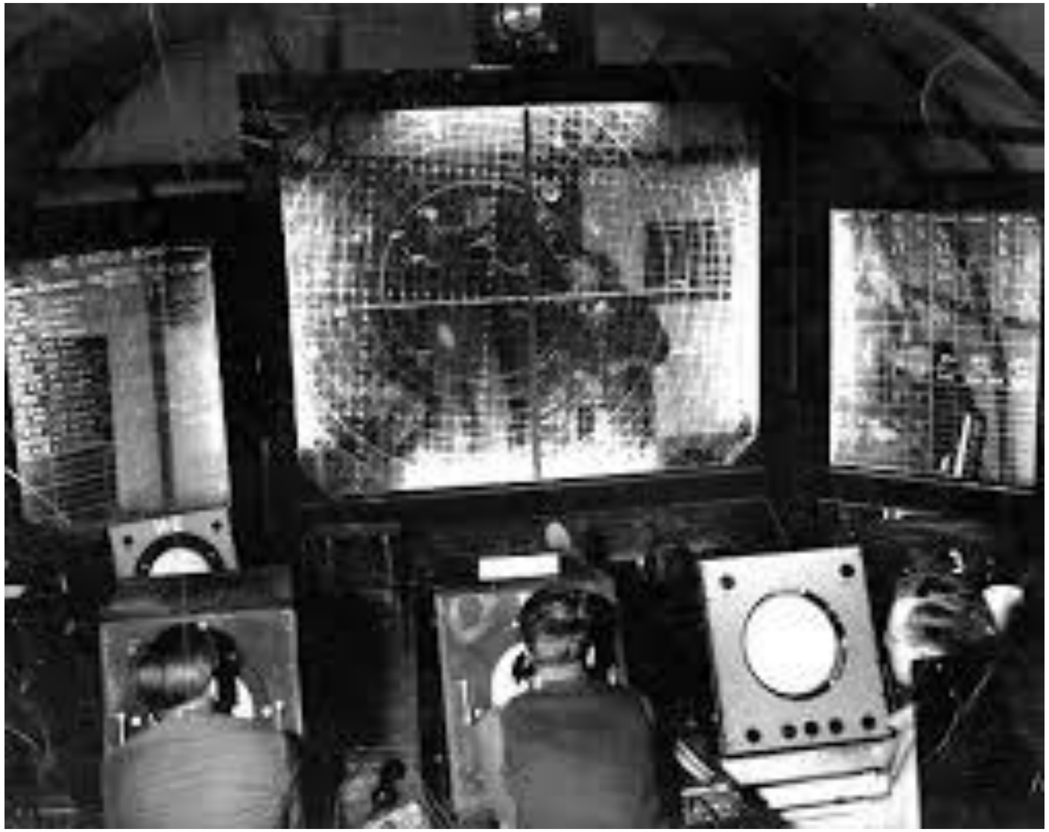
- randomAccuracy= 0.15*0.015 + 0.85*0.985 = 0.84

- Kappa = $\frac{TotalAccur - randomAccur}{1 - randomAccur} = \frac{0.857 - 0.84}{1 - 0.84}$ = 0.10

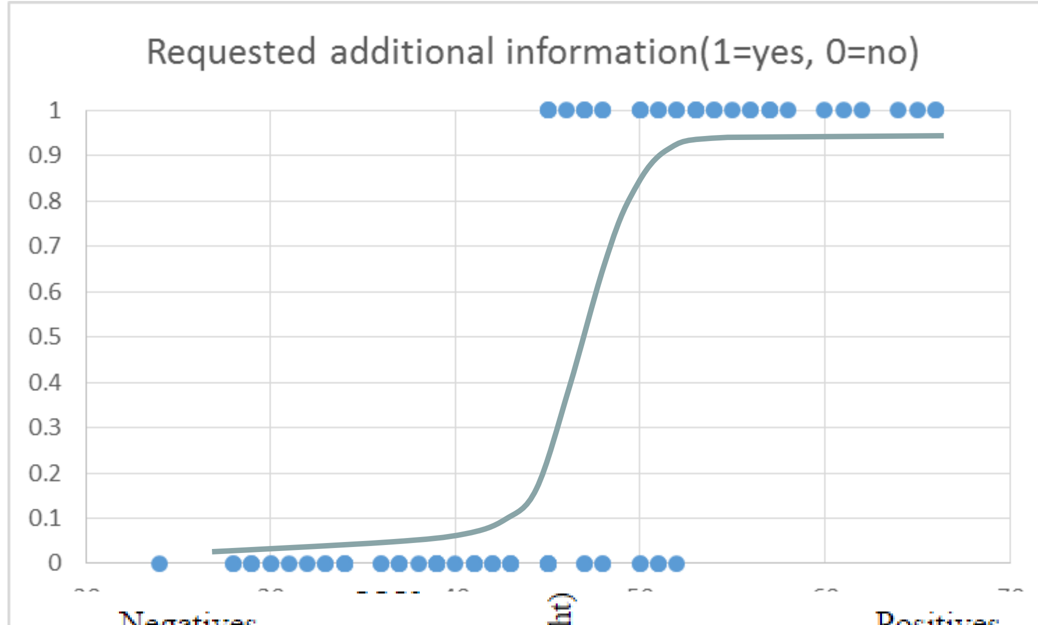
Slightly better than Random!

ROC Curves and AUC

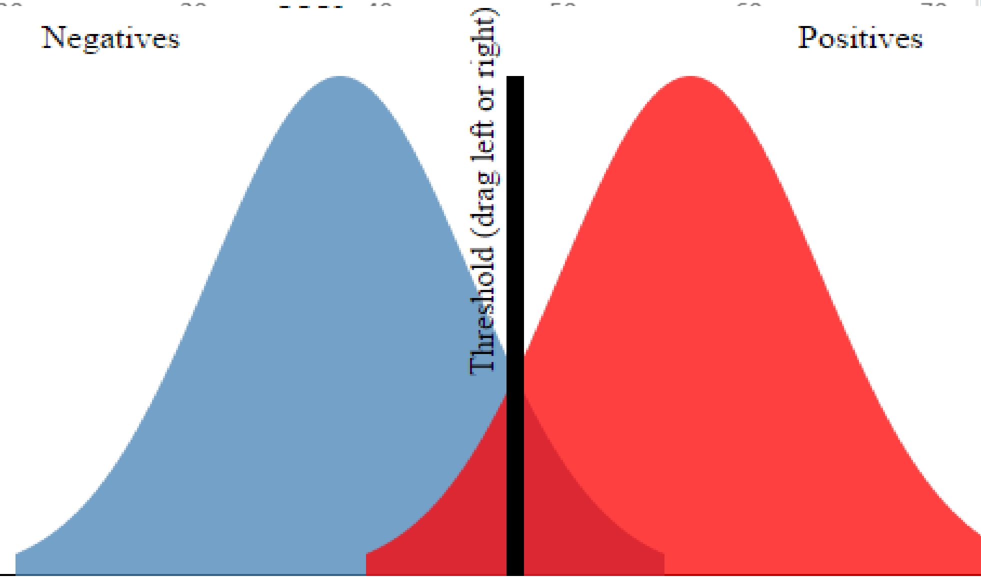
- ROC – Receiver Operating Characteristics
- AUC – Area Under the ROC Curve



- 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

- 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

	Sensitivity		Specificity	
Probability Threshold for Discriminating Between High Risk and Low Risk of Having Ten Year CHD	True Positive Rate	False Positive Rate	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

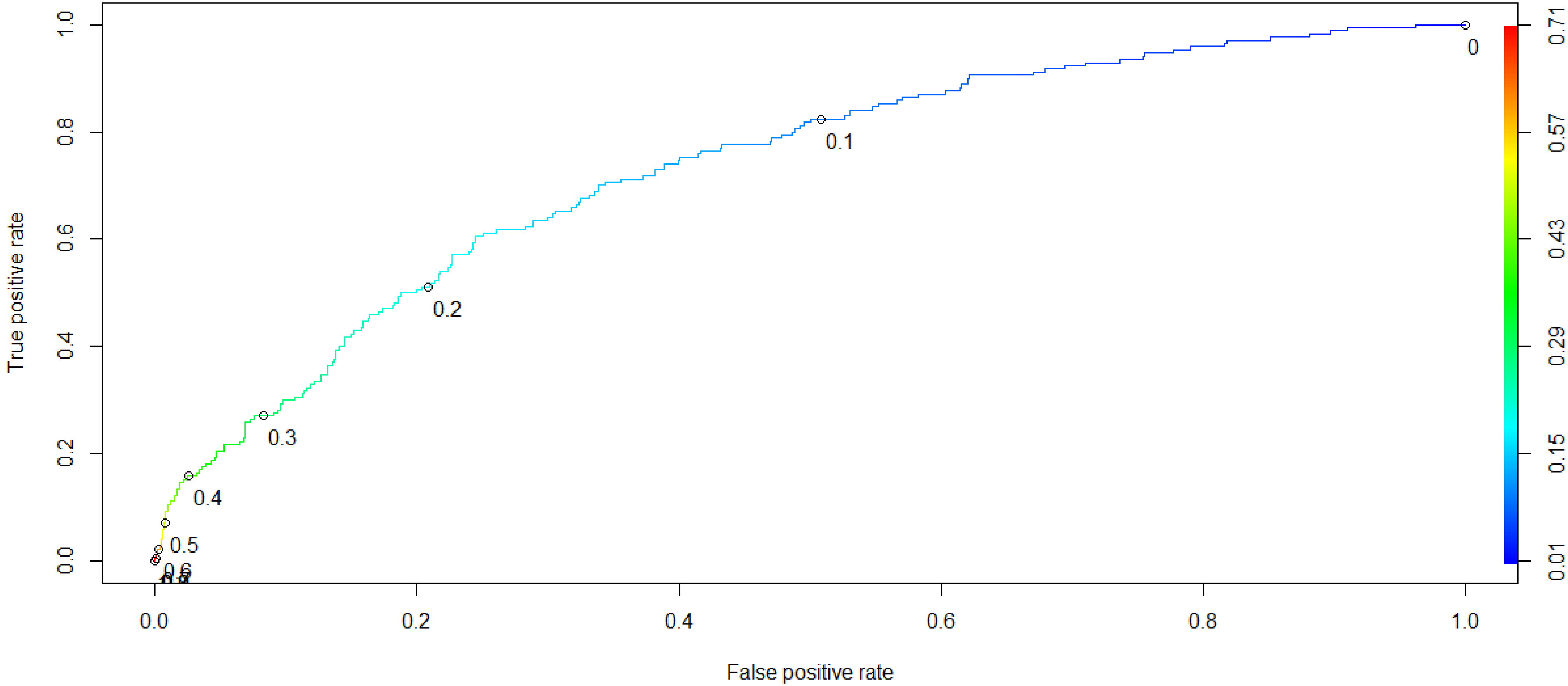
Probability Threshold for Discriminating Between High Risk and Low Risk of Having Ten Year CHD	True Positive Rate (Sensitivity)	False Positive Rate
0.9	0/170	0/922
0.7	1/170	1/922
0.5	12/170	7/922
0.3	46/170	76/922
0.1	140/170	468/922

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



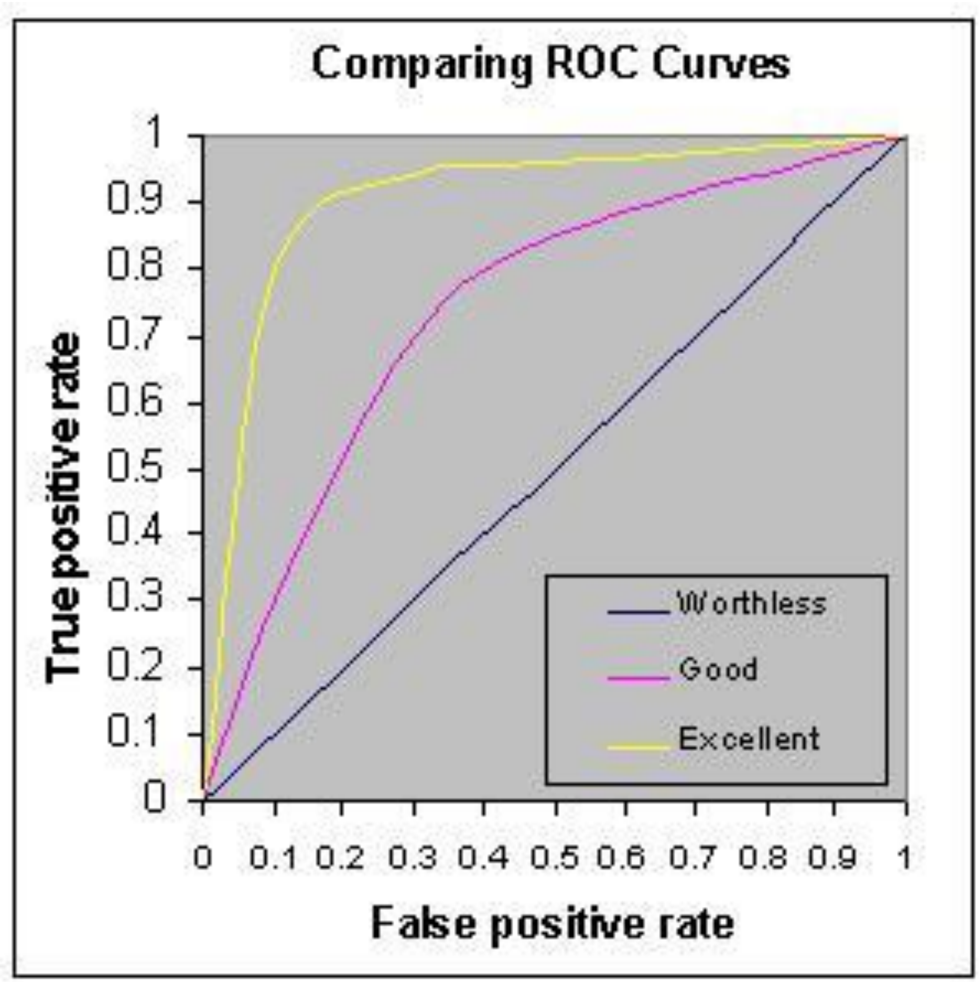
ROC Curves and AUC

- 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



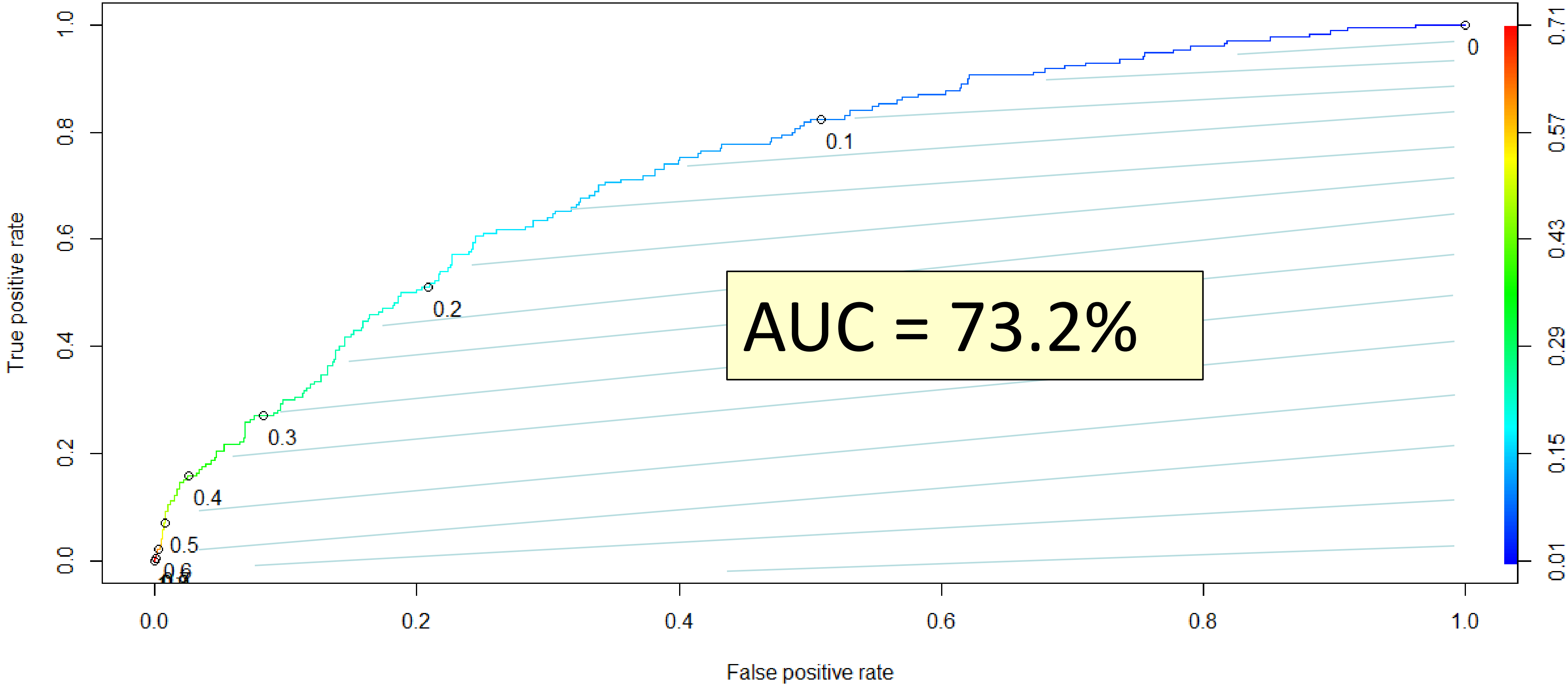
Rough rule of thumb:

- 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.



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

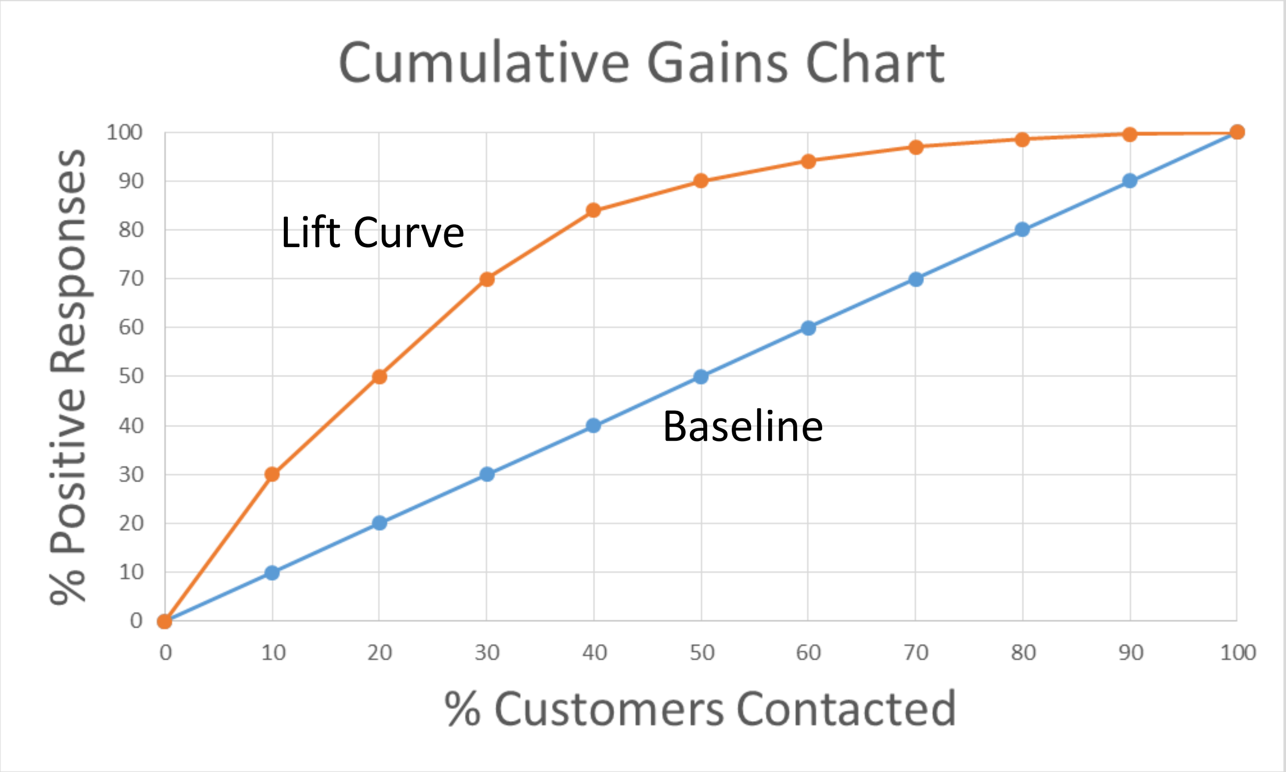
- With a predictive model, where the model assigns a probability to each customer, the customers are ordered and divided into deciles (or any other quantiles). They are then called in decreasing order of probability to buy.

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

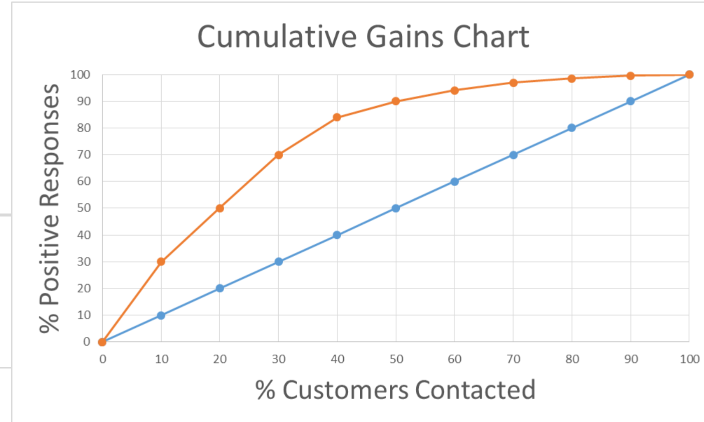
Gains and Lift Charts

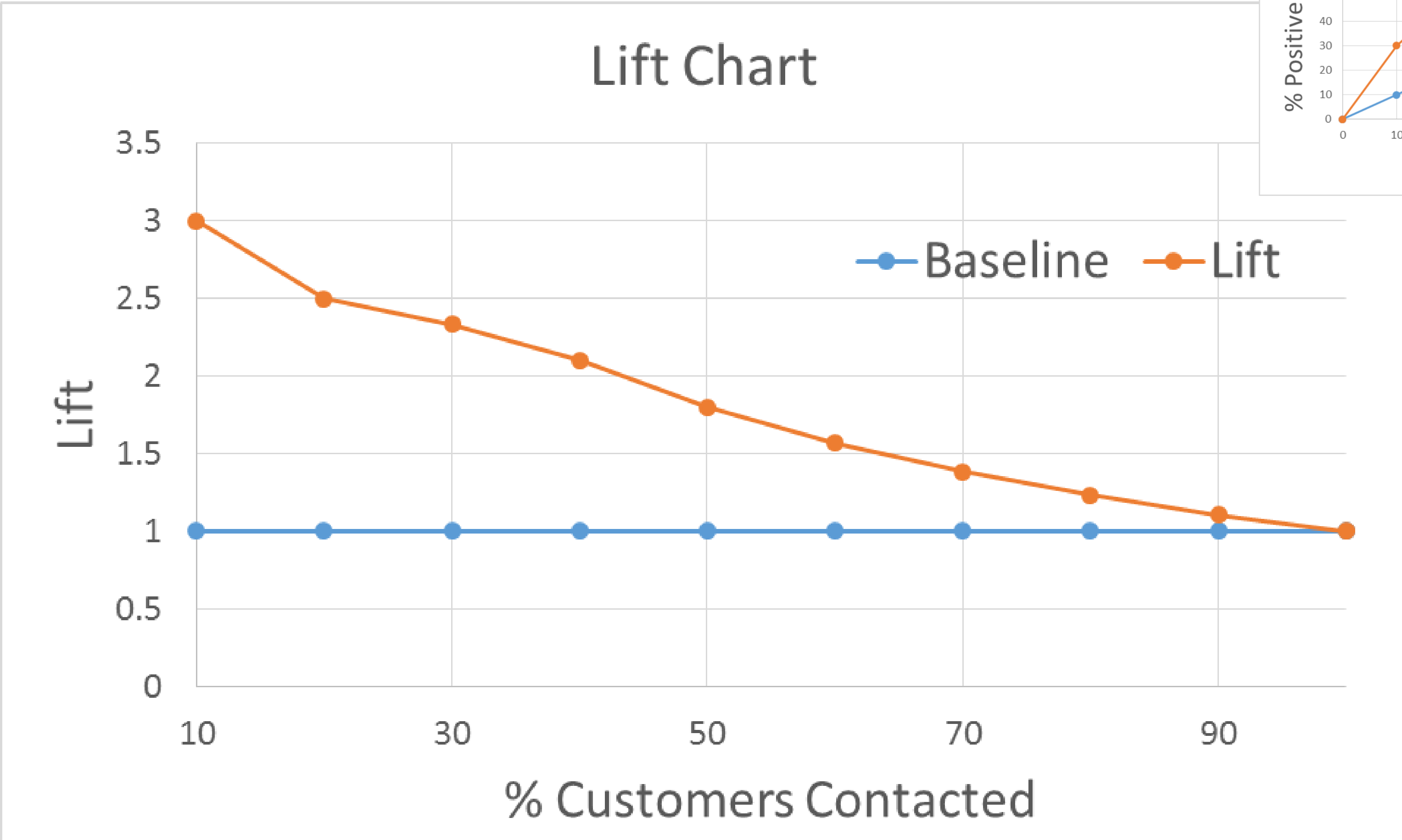
% Called	Called at Random	Called According to Model Score
0	0	0
10	10	30
20	20	50
30	30	70
40	40	84
50	50	90
60	60	94
70	70	97
80	80	98.5
90	90	99.5
100	100	100

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

Evaluating Model Accuracy and Bias-Variance Tradeoff

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: Same as holdout but useful when the data size is small.

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.

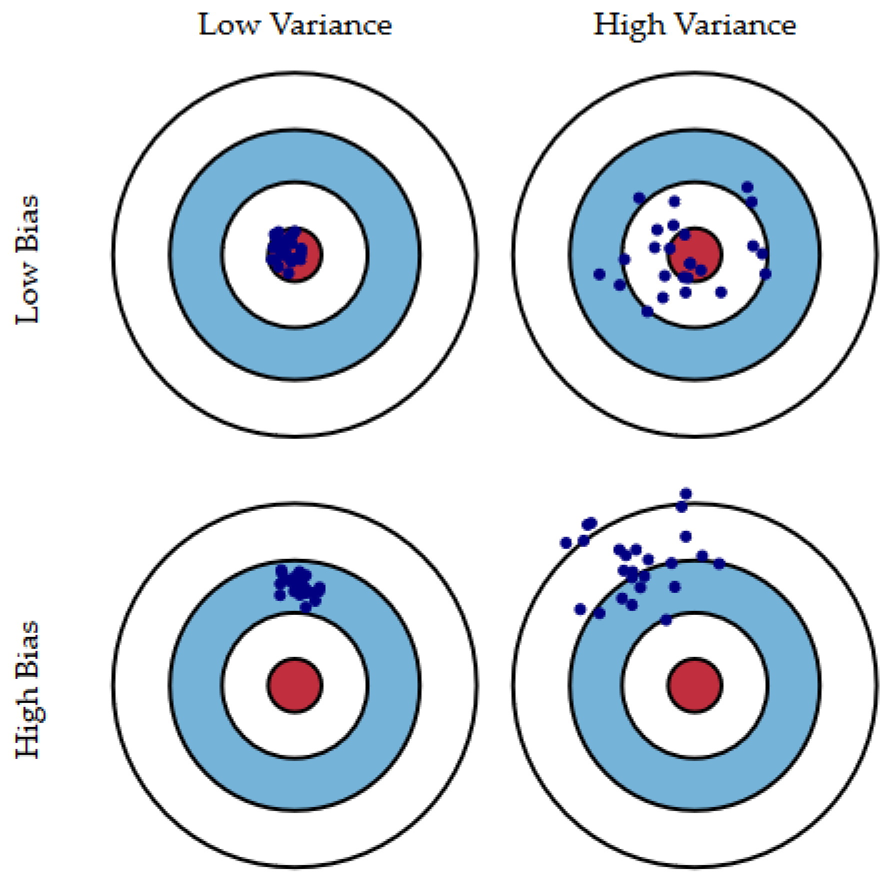
Bias-Variance Tradeoff

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

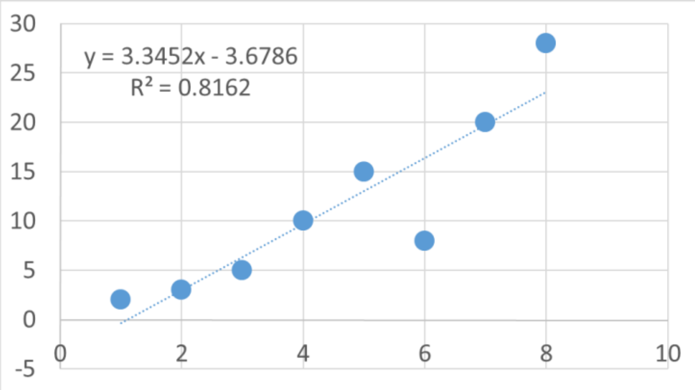
- If the model performance on training and testing data sets is inconsistent, it indicates a problem either with Bias or Variance.

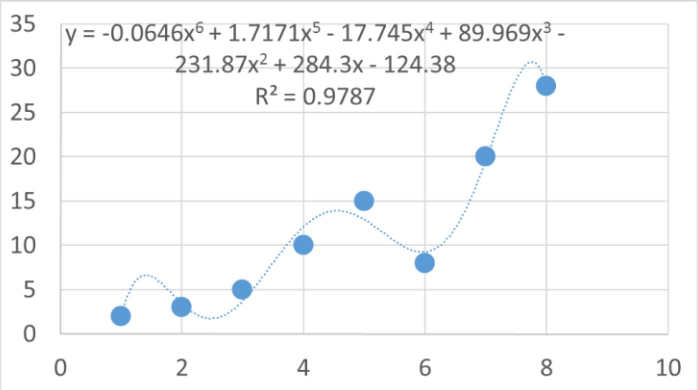
Bias-Variance Tradeoff

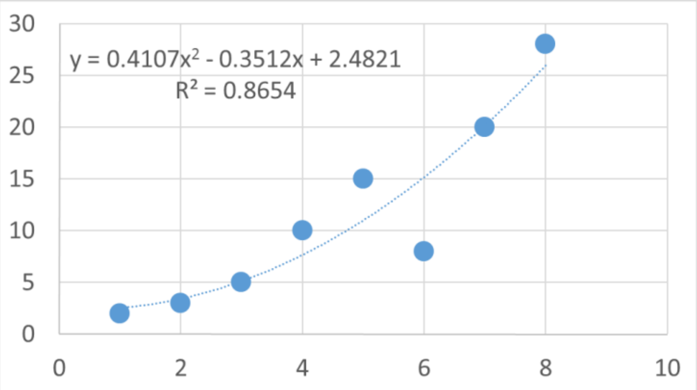
- Bulls-eye is a model that correctly predicts the real values.
- Each hit is a model based on chance variability in training datasets.



Bias-Variance Tradeoff and Underfitting vs Overfitting Excel

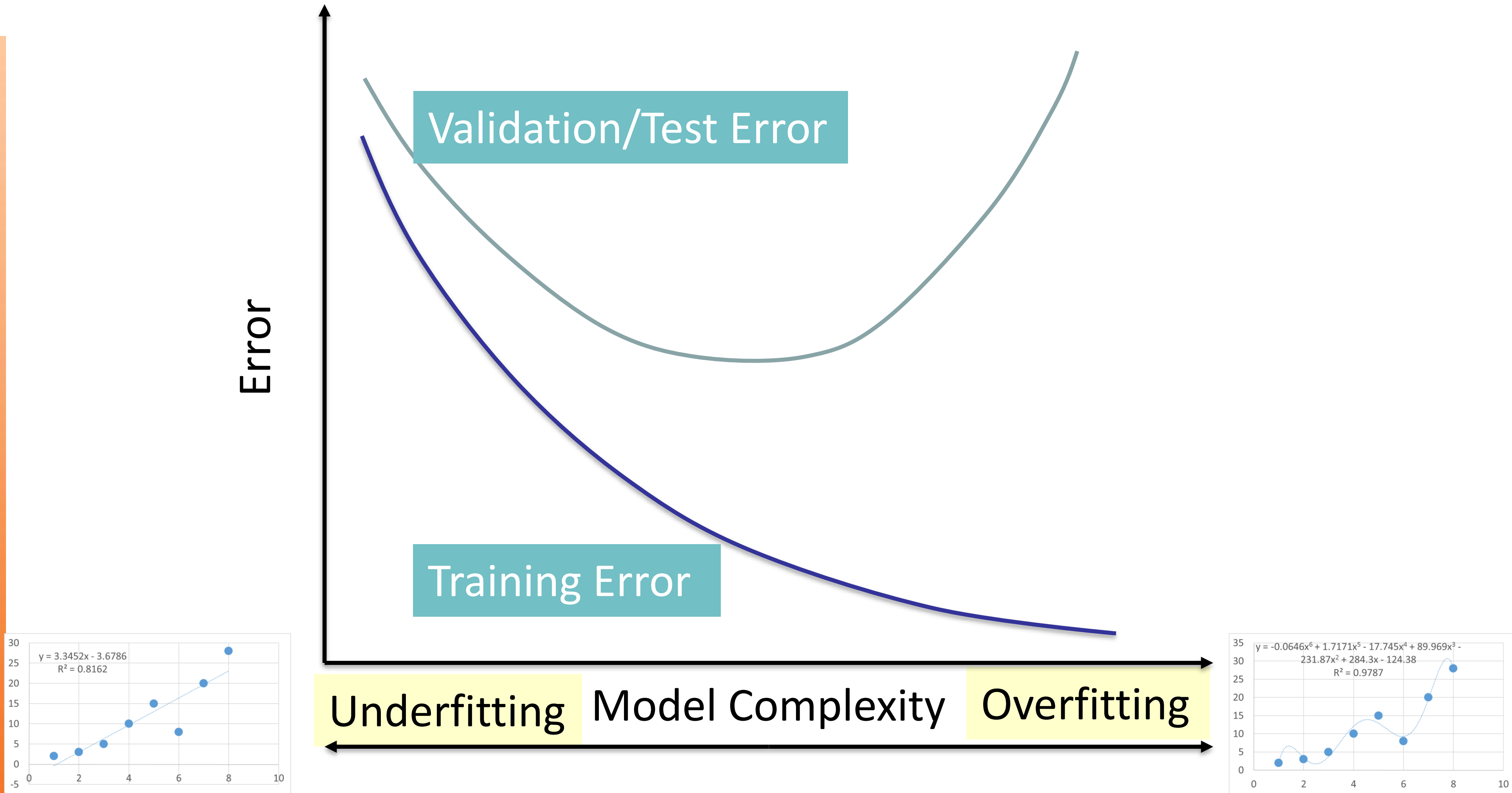




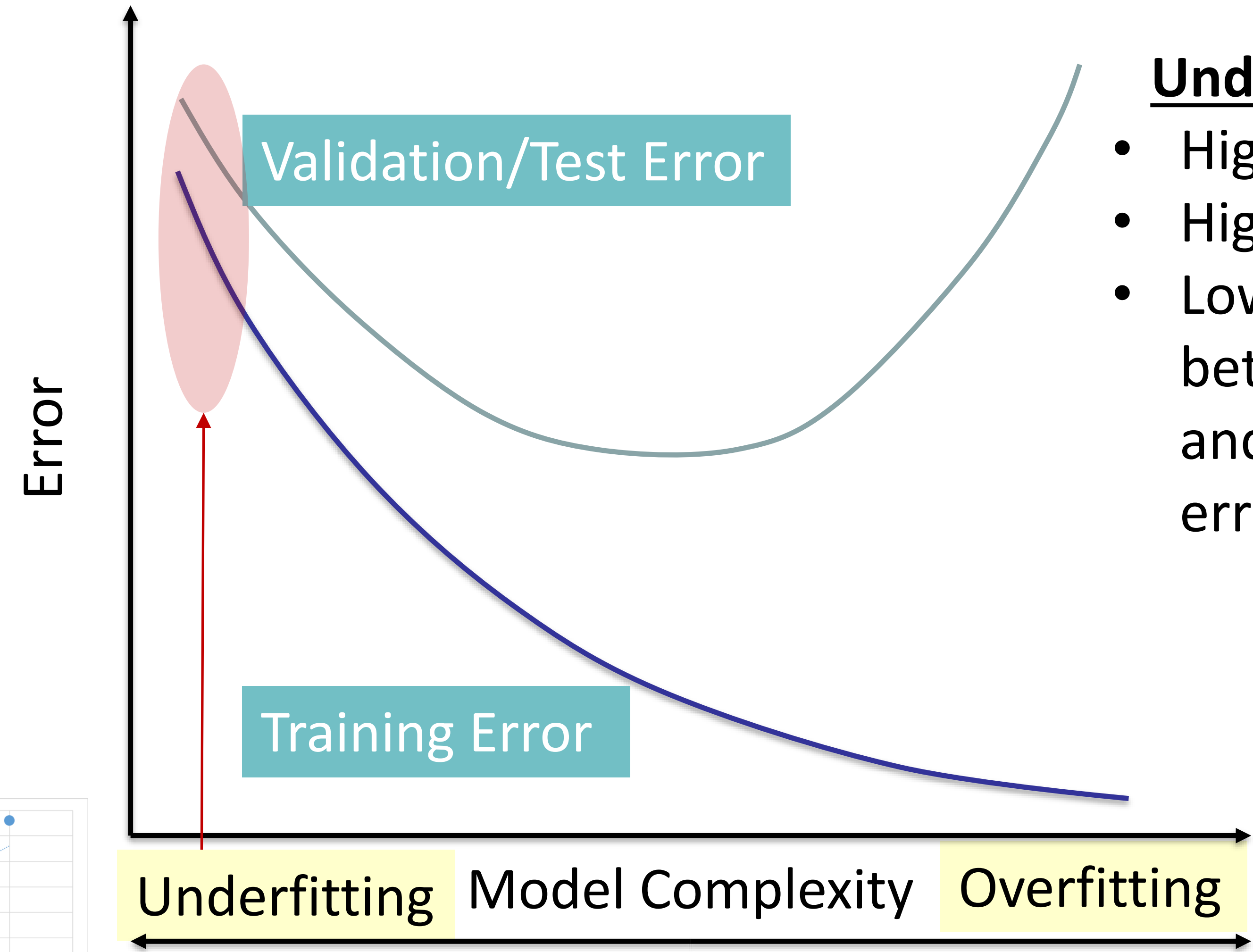


Too Simple a Model	Too Complex a Model	Right Model
Underfit	Overfit	Reasonable fit

Bias-Variance Tradeoff and Underfitting vs Overfitting

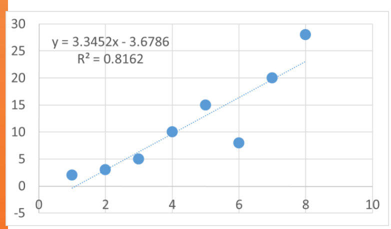


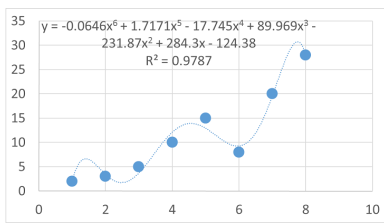
Diagnosing Bias and Variance



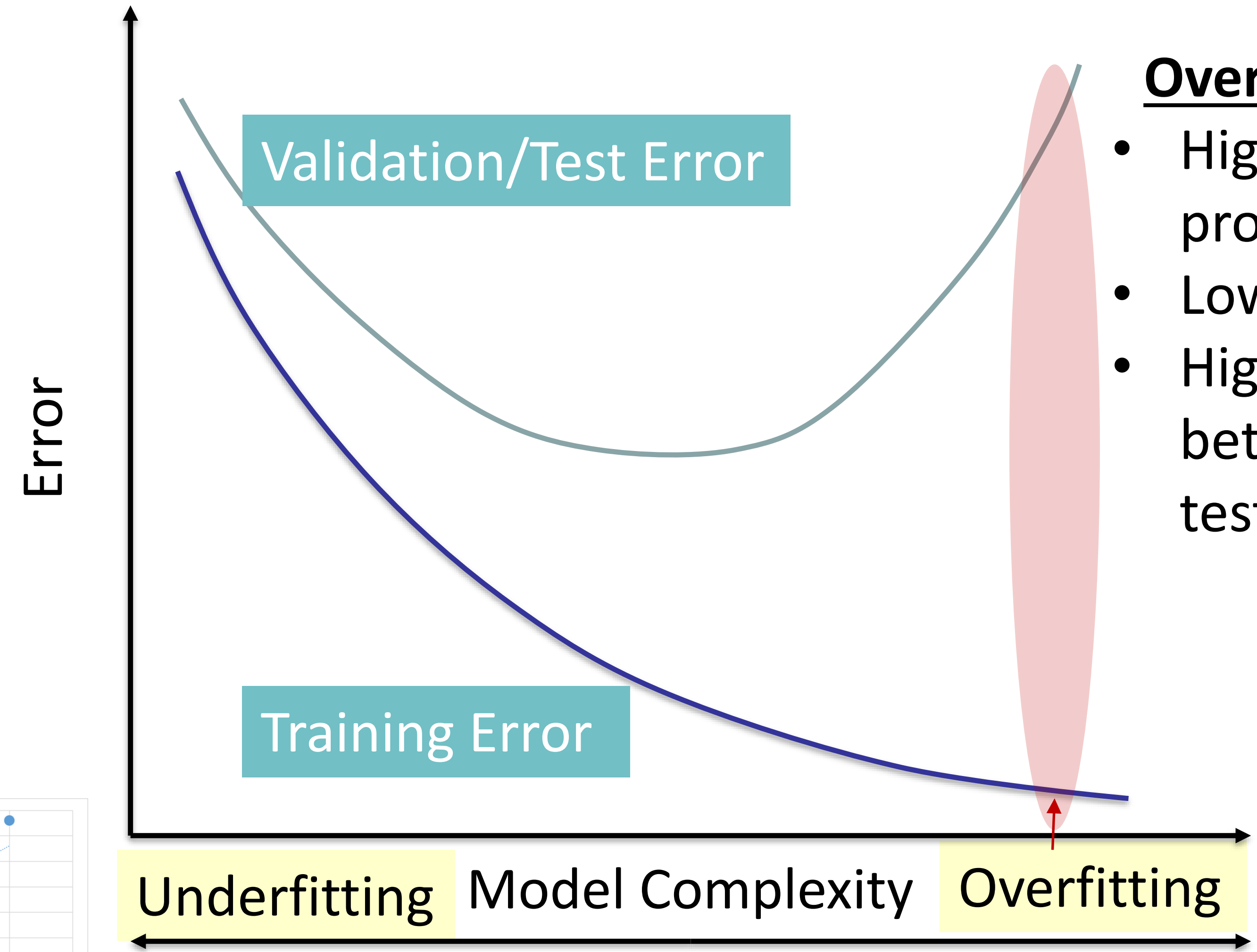
<u>Underfitting (Bias)</u>

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



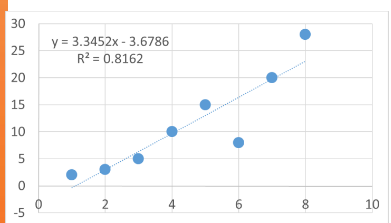


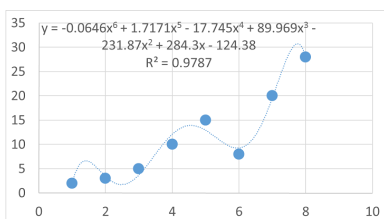
Diagnosing Bias and Variance



Overfitting (Variance)

- 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.

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

Bagging (bootstrap aggregating) techniques *(taught later in the program)* can be used to minimize errors.

International School of Engineering

Plot 63/A, Floors 1&2, Road # 13, Film Nagar, Jubilee Hills, Hyderabad - 500 033

For Individuals:	+91-9502334561/63 or 040-65743991
For Corporates:	+91-9618483483
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.