Prediction

Business Analytics

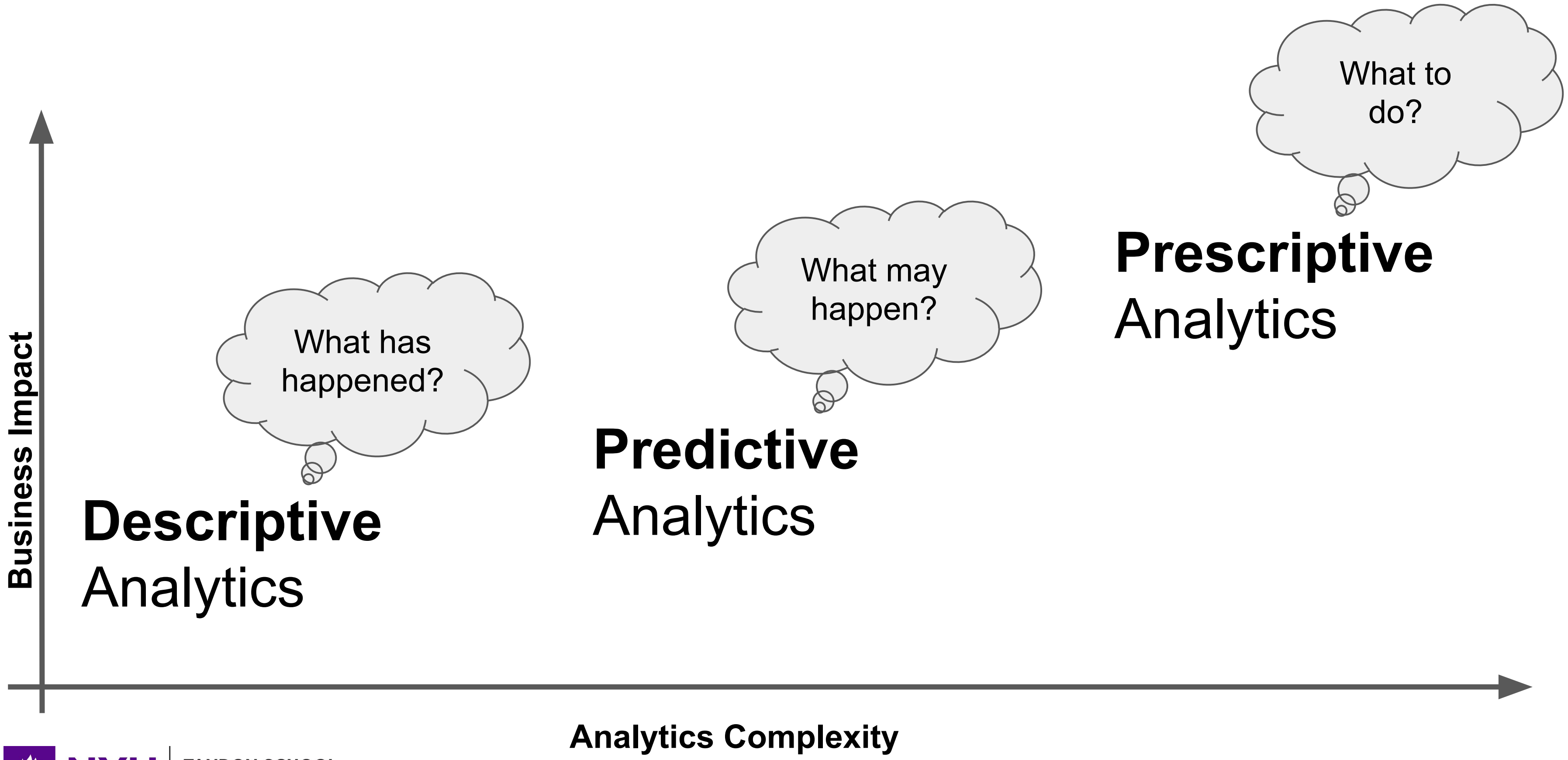
Business Analytics Definition

Business analytics refers to the application of data analysis and modeling techniques for understanding business situations and improving business decisions.

IMPLICATIONS:

- data → past business performance
- methods → statistics + mathematics + computational methods
- business decisions → actionable insight

Types of Analytics



Understanding Your Data

Exploratory analysis of data is useful for:

- understanding data properties
- detecting errors, ensuring data quality
- finding patterns in data
- determining relationships among variables
- checking assumptions
- mapping business problems into data mining tasks and suggesting modeling strategies

Homework

Part 1: Citibike Descriptive Analytics

Analytics Questions:

- Compute summary statistics for tripduration
- Compute summary statistics for age
- Compute summary statistics for tripduration in minutes (Need to transform tripduration from seconds to minutes)
- Compute the correlation between age and tripduration
- Plot the histograms and box plots for tripduration by gender

Business Questions:

- What is the total revenue assuming all users riding bikes from 0 to 45 minutes pay $3 per ride and user exceeding 45 minutes pay an additional $2 per ride.
- Looking at tripduration in minutes, what can you say about the variance in the data.
  - What does this mean for the pricing strategy?
  - What does this mean for inventory availability?
- A business manager wants to reallocate the $5M marketing budget using a gender segmentation strategy. Specifically, the manager is asking you to create two models:
  - A model that use % of male vs females in the dataset
  - A model based on average trip duration by gender

Homework

Part 2: Teach me something

This part of the assignment is fairly simple and open-ended. Your first task is to get yourself a data set that you like and teach me something about it. Anything. It doesn't have to be profound, it doesn't have to be earth changing, it should just use your skills from this lesson. Some thoughts on choosing your dataset:

- I'm assuming many of you have datasets that you're already working with for other projects (web traffic, Kinect output, Twitter feeds, biofeedback data, etc.), so feel free to use one of those.
- Don't have data already? No worries. The easiest place to get tabular data (CSV) is from Data.Gov https://catalog.data.gov/dataset?res_format=CSV
- Not everything is a CSV (the only type of data we've loaded in yet), but if you can find tabular data, that's going to fit well with the course.

To submit

Once you've got your dataset, your job is to do the following:

1. Write a couple of sentences about what your dataset contains (column names, types) and why you chose the dataset.
2. Teach me one thing about your dataset. This can (and should be) extremely basic. You don't have to find some amazing correlation in your data, just tell me one true thing. Or make one plot. You've learned how to look at maxes and mins, you can subset your data, you know how to plot it, so you should easily be able to find something to say about your data.
3. Finally, what is the business application of the findings and dataset. What possibilities do you have now as a business manager?

Lesson Objectives

1. Regression - Theory
   a. Linear Models
   b. Ordinary Least Squares
   c. Simple Linear Regression
2. Regression Applied
   a. Model strength
   b. Model interpretation
   c. Dummy variables
   d. Non-linear transformations
3. Classification
   a. Statistical classification
   b. Decision Trees

Linear Models

Regression models estimate the relationships among variables to predict outcomes.

Example: How does bike trip duration change as we introduce a new customer type, a new pricing scheme, or with different weather conditions.

In this week you will learn the basics of regression analysis and the specifics about linear regression models that example the relationship between numerical variables.

Business Case:

What is the influence of a variable (price, advertising, and etc.) on an outcome (market shares, sales, overall satisfaction)?

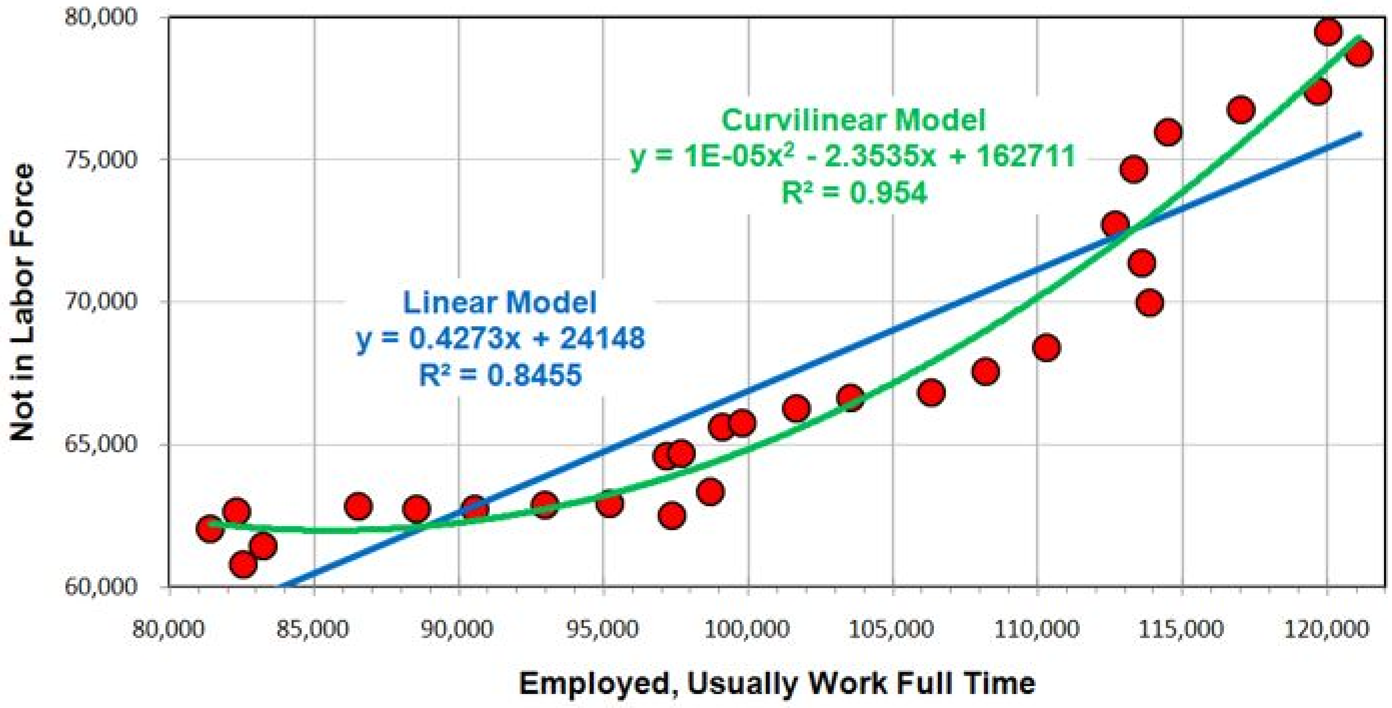
$$X \rightarrow Y$$

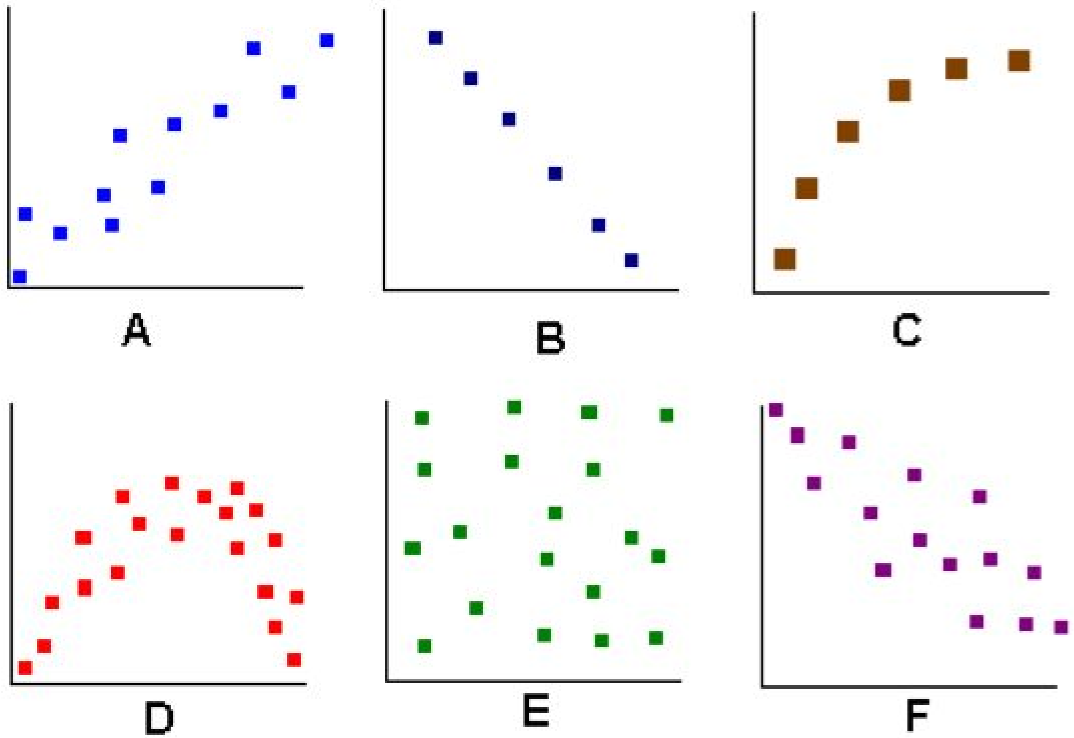
Independent variable (X) → Dependant variable (Y)

$$y=mx+b$$

NOTE IN TERMINOLOGY

- *Y is know as the dependent variable the variable that regression model seek to predict or response variable*
- *X is the independent variable, predictor or explanatory variable.*


Curvilinear Model
y = 1E-05x2 - 2.3535x + 162711
R² = 0.954
Linear Model
y = 0.4273x + 24148
R² = 0.8455
Not in Labor Force
80,000
75,000
70,000
65,000
60,000
80,000
85,000
90,000
95,000
100,000
105,000
110,000
115,000
120,000
Employed, Usually Work Full Time



A
B
C
D
E
F

Simple Linear Regression

- The "workhorse" of statistical analysis is the simple linear regression.
- Used to determine the relationship between two variables.
  - Given one variable, a regression will provide the expected value of the other variable.
- The outcome of the regression → Y: response.
- The input variable → X: predictor.

$$Y_i = b_0 + b_1 X_i + \varepsilon_i, \; i=1, \ldots, n$$

where:

Y_i = *i*th observation of the dependent variable, Y

X_i = *i*th observation of the independent variable, X

b_0 = regression intercept term

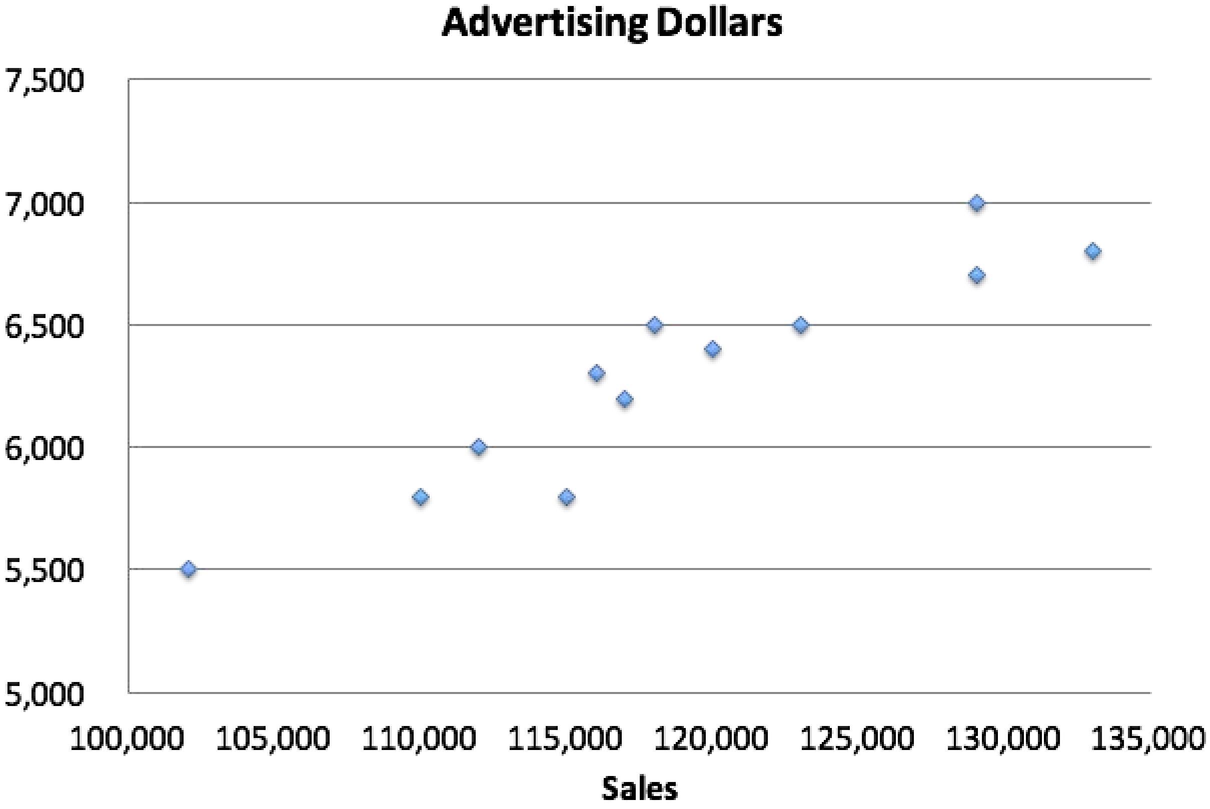
b_1 = regression slope coefficient

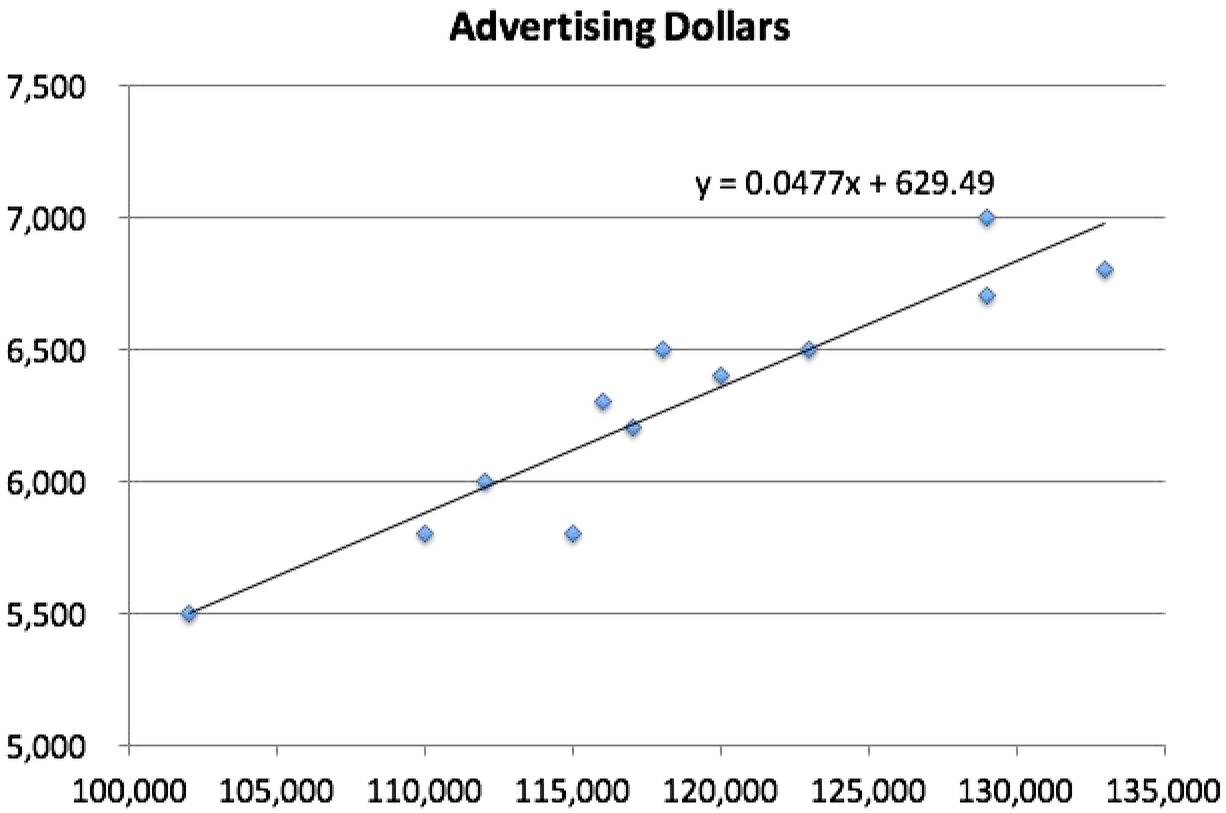
ε_i = residual for the *i*th observation (also referred to as the disturbance term or error term)

Sales vs Advertising

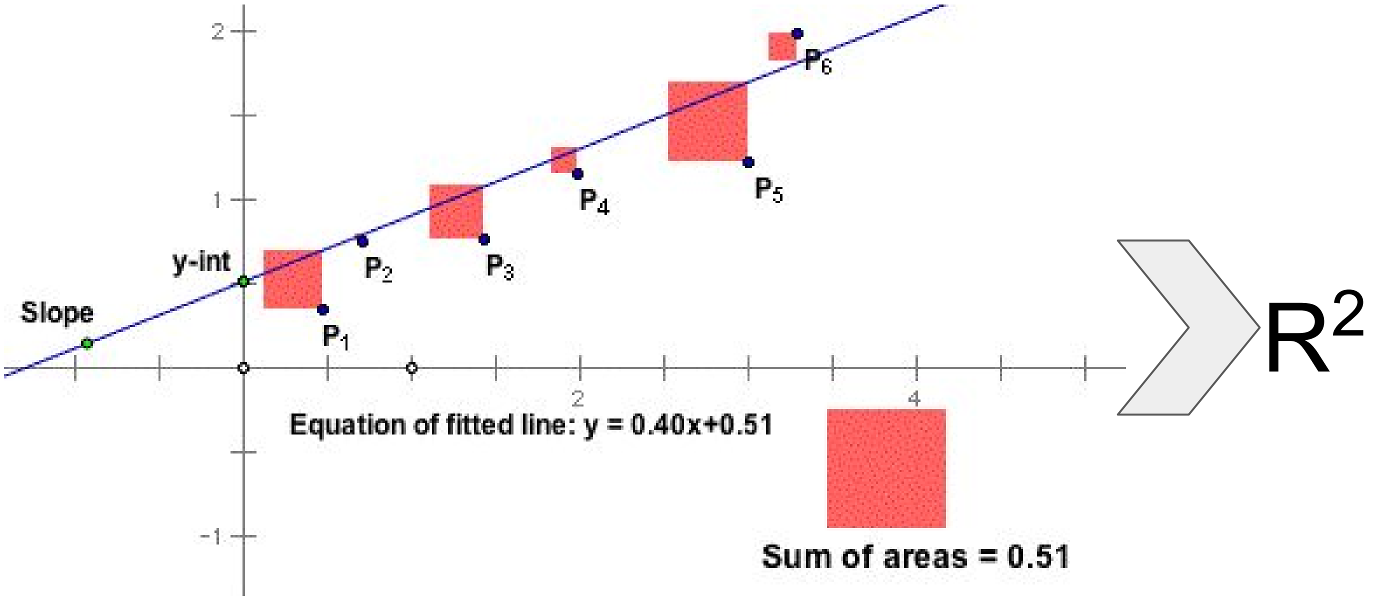
Month	Sales	Advertising Dollars
Jan	102,000	5,500
Feb	110,000	5,800
Mar	112,000	6,000
Apr	115,000	5,800
May	117,000	6,200
Jun	116,000	6,300
Jul	118,000	6,500
Aug	129,000	7,000
Sep	123,000	6,500
Oct	120,000	6,400
Nov	129,000	6,700
Dec	133,000	6,800

Scatter


Advertising Dollars
7,500
7,000
6,500
6,000
5,500
5,000
100,000
105,000
110,000
115,000
120,000
125,000
130,000
135,000
Sales


Advertising Dollars
y = 0.0477x + 629.49
7,500
7,000
6,500
6,000
5,500
5,000
100,000
105,000
110,000
115,000
120,000
125,000
130,000
135,000

Ordinary Least Squares



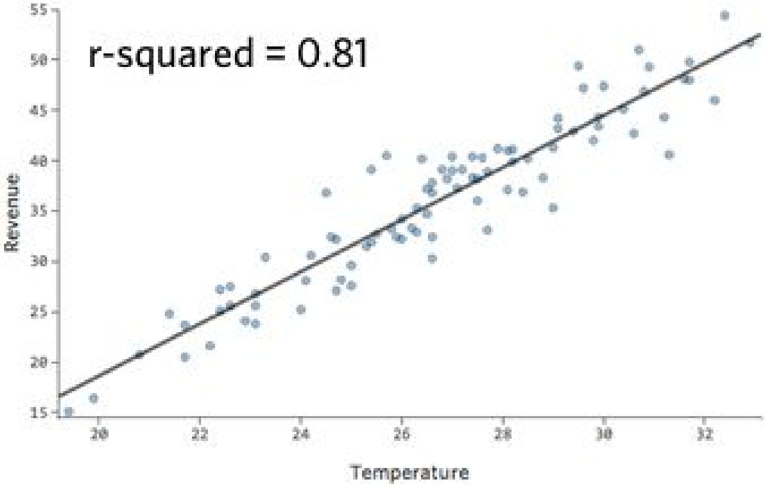
R^2 is an statistical measure of how close the data are to the fitted regression line.

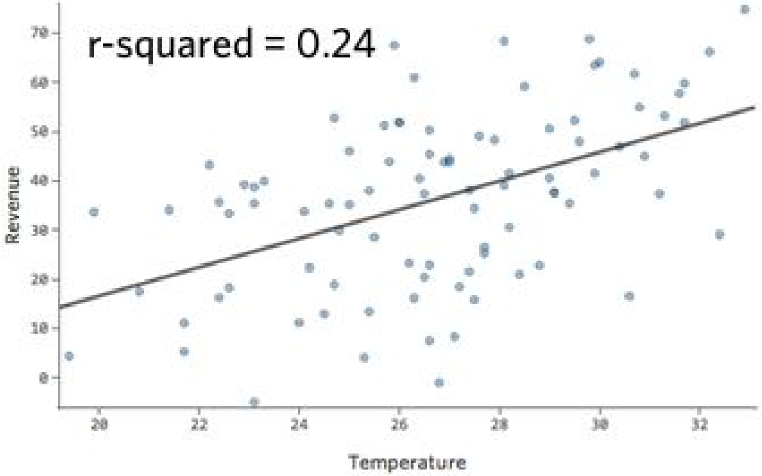
It indicates the goodness of fit of the model.

R^2 definition: Explained variation / Total variation

R^2 is always between 0 and 100%:

- 0% → model explains none of the variability
- 100% → model explains all the variability


r-squared = 0.81
Revenue
Temperature
55
50
45
40
35
30
25
20
15
20
22
24
26
28
30
32


r-squared = 0.24
Revenue
Temperature
70
60
50
40
30
20
10
0
20
22
24
26
28
30
32

Calculating b1:		0.0477	
x-mean(x)	y-mean(y)	(x-mean(x))* (y-mean(y))	(x-mean(x))^2
-16,666.67	-791.67	13,194,444.44	277,777,777.78
-8,666.67	-491.67	4,261,111.11	75,111,111.11
-6,666.67	-291.67	1,944,444.44	44,444,444.44
-3,666.67	-491.67	1,802,777.78	13,444,444.44
-1,666.67	-91.67	152,777.78	2,777,777.78
-2,666.67	8.33	-22,222.22	7,111,111.11
-666.67	208.33	-138,888.89	444,444.44
10,333.33	708.33	7,319,444.44	106,777,777.78
4,333.33	208.33	902,777.78	18,777,777.78
1,333.33	108.33	144,444.44	1,777,777.78
10,333.33	408.33	4,219,444.44	106,777,777.78
14,333.33	508.33	7,286,111.11	205,444,444.44

Calculating b0: **629.4926**

Intercept

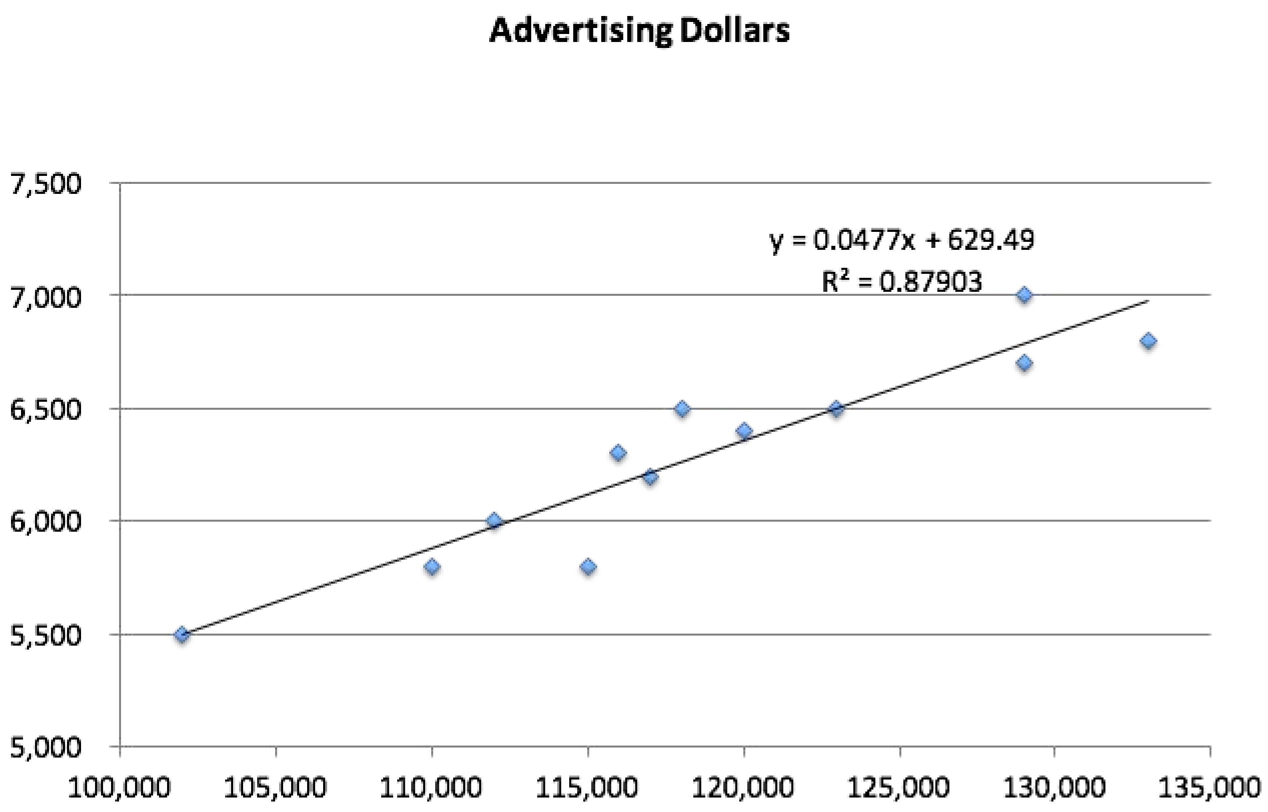
$$\hat{y}_i = b_0 + b_1 x_i$$

Slope

$$b_0 = \bar{y} - b_1 \bar{x}$$

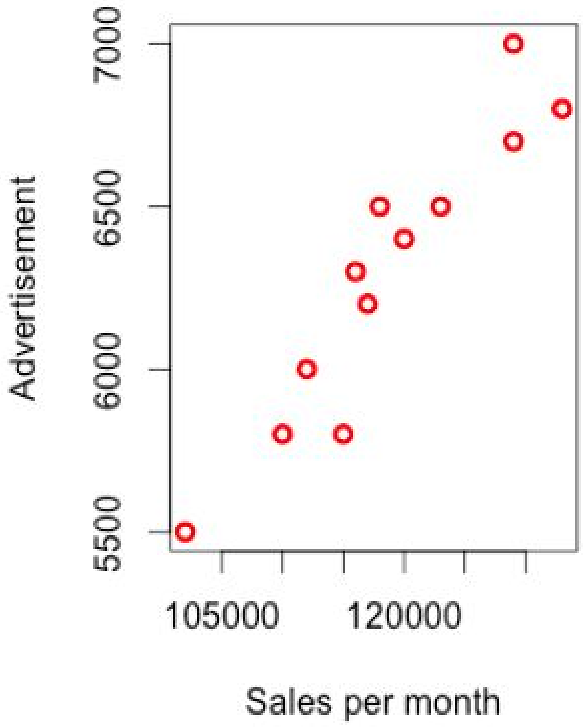
$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$R^2 = \left(\frac{\sum(xy) - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x^2)}{n}\right]\left[\sum y^2 - \frac{(\sum y^2)}{n}\right]}} \right)^2$$

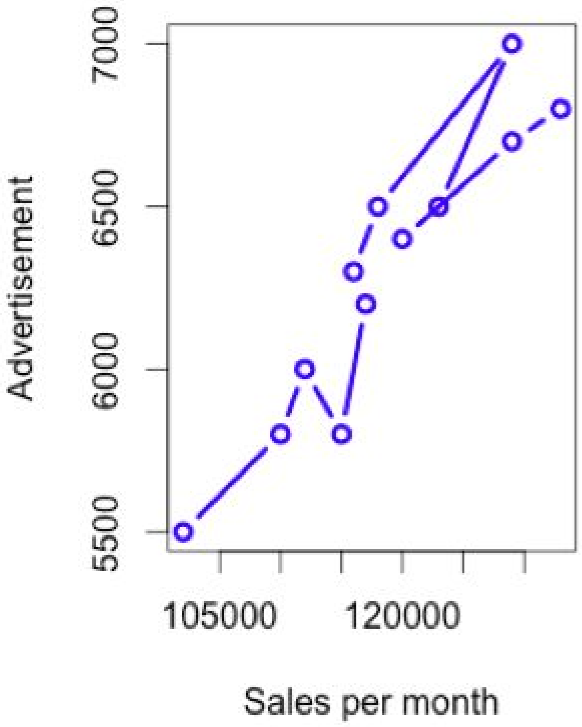

Advertising Dollars
y = 0.0477x + 629.49
R² = 0.87903
7,500
7,000
6,500
6,000
5,500
5,000
100,000
105,000
110,000
115,000
120,000
125,000
130,000
135,000

R Code for linear models

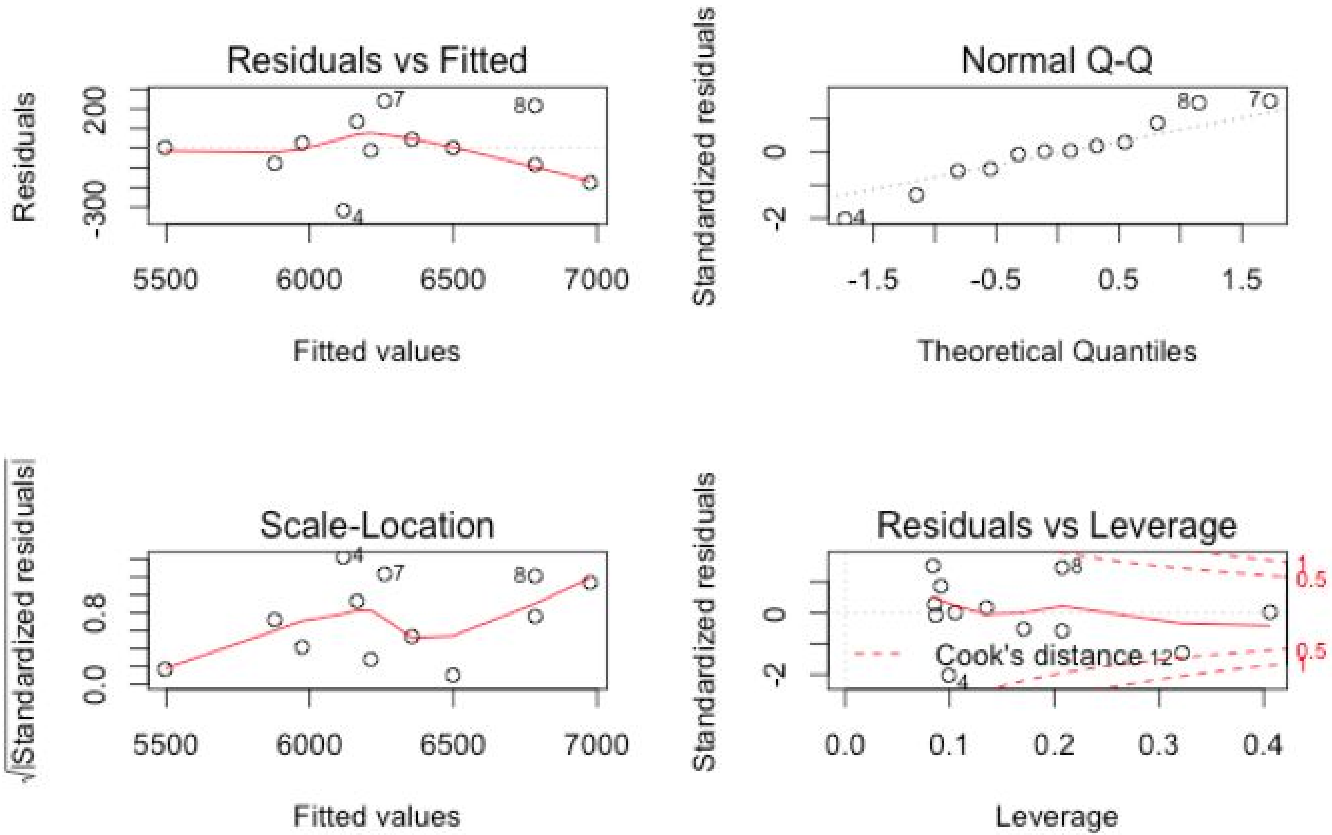
```
data <- read.csv("~/Google Drive/Business Analytics/Data/Sales vs Advertisement.csv",
     header=TRUE, stringsAsFactors=TRUE)
dim(data)
names(data)
x=data$Sales
y=data$Advertising.Dollars
par(mfrow=c(1,2))
 plot(x,y , col="red", lwd=3,
     ylab="Advertisement", xlab="Sales per month")
plot(x,y, type="b", col="blue", lwd=3,
     ylab="Advertisement", xlab="Sales per month")

model<-lm(y ~ x)
model
summary(model)

par(mfrow=c(2,2))
plot(model)
```



Advertisement
7000
6500
6000
5500
105000
120000
Sales per month



Advertisement
7000
6500
6000
5500
105000
120000
Sales per month



Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

R Session

Build a linear model for Zagat using “Food” as and predictors and “Price” as a response.

Build a linear model for Zagat using “Food” and “Decor” as and predictors and “Price” as a response. Hint use *lm(y ~ x1+x2)*

Build a linear model for Zagat using “Food”, “Decor”, and “Service” as and predictors and “Price” as a response. Hint use *lm(y ~ x1+x2+x3)*

What is a Model?

A model is a representation or simplified version of a concept, phenomenon, relationship, or system of the real world.

The objectives of a model include:

1. to facilitate understanding
2. to aid in decision making by simulating 'what if' scenarios
3. to explain, control, and predict events on the basis of past observations.

Since most objects and phenomenon are very complicated and much too complex to be comprehended in their entirety, a model is “simplified” based on some assumptions about what is and is not important for a specific purpose.

Predictive Modeling

- The model describes a relationship between a set of selected variables and the predefined target variable.
- How do we find or select important, informative variables or attributes of the entities described by the data??
- e.g. Will a customer churn soon after her contract expires?
  - Are there one or more variables that reduce the uncertainty around the value of the target, i.e., the customer churning?
  - Build a model of the propensity to churn as a function of customer attributes

Modeling Concepts

- The creation of models from data is known as **model induction**.
  - Philosophical term that refers to generalizing from specific cases to general rules.
  - Models are general rules in a statistical sense -- they do not hold 100% of the time.
- The procedure that creates the model is called the **induction algorithm or learner**.
- The input data for the induction algorithm are called the **training data**.
  - The value of the target variable is known.

Regression Analysis

The uses of a regression model include:

- Determining whether a relationship exists between variables
- Determining the strength of the relationship
- Assessing the marginal effect of a specific variable
- Forecasting/predicting the values of the dependent variable

Case Study

Suppose you are helping Warner Bros in developing a model for forecasting Box Office revenues for a new movie.

Variable	Description
Movie	Name of the movie
Opening_Week_Revenue	Opening week revenue in Millions of $
Num_Theaters	Number of movie theaters each movie was initially released at
Overall_Rating	Critic ratings for each movie (higher the number, more favorable the rating)
Genre	1:Action, 2:Comedy, 3:Kids, 4: Other

Case Study

Movie	Opening_Week_Revenue	Num_Theaters	Overall_Rating	Genre
Van Helsing	51.7	3575	36	1
Collateral	24.7	3188	71	1
Alien Vs. Predator	38.3	3395	29	1
Man on Fire	22.8	2980	47	1
Sex and the City	57	3285	53	2
Marley and Me	36.4	3480	53	2
Four Christmases	31.1	3310	41	2
Tropic Thunder	25.8	3319	71	2

Roadmap

- Understand the data
  - descriptive statistics for variables of interest
  - plotting your dependent variable to check for any outliers, presence of trends or seasonality
- Selection of Variables
  - statistical methods
  - judgement
    - The variable’s importance in making a managerial decision
    - The variable helps to control for important factors
  - data availability

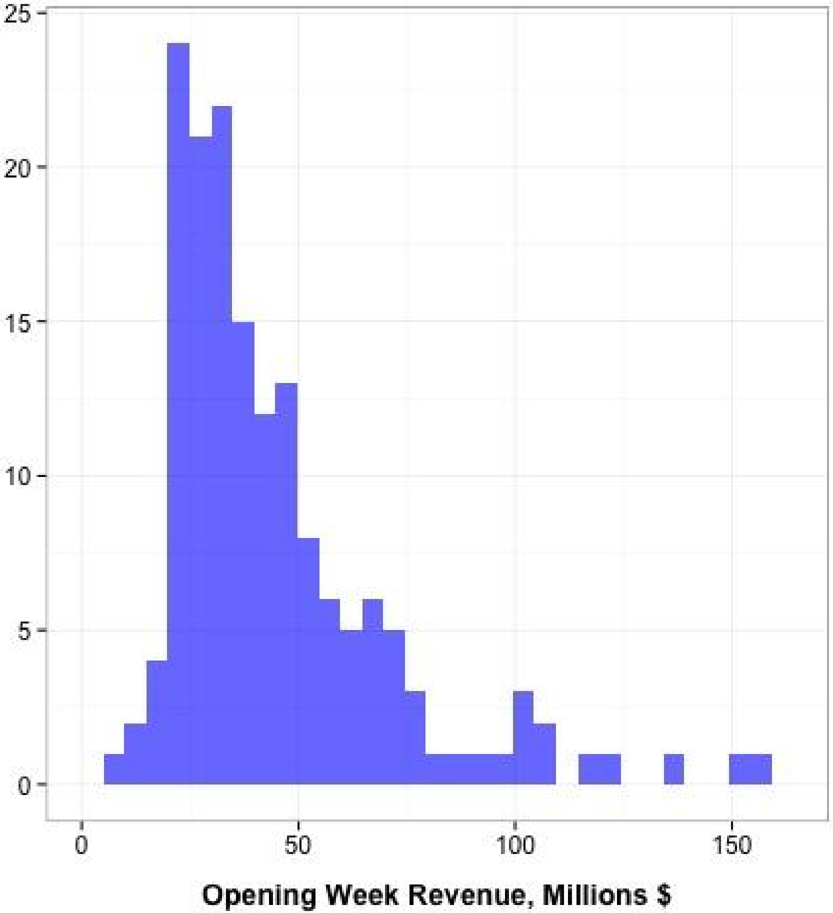
Objective

- Develop a regression model for “Opening week Revenues” using the remaining variables as predictors. Interpret your parameters.
- The attributes for the new movie “You Name It” are as follows:

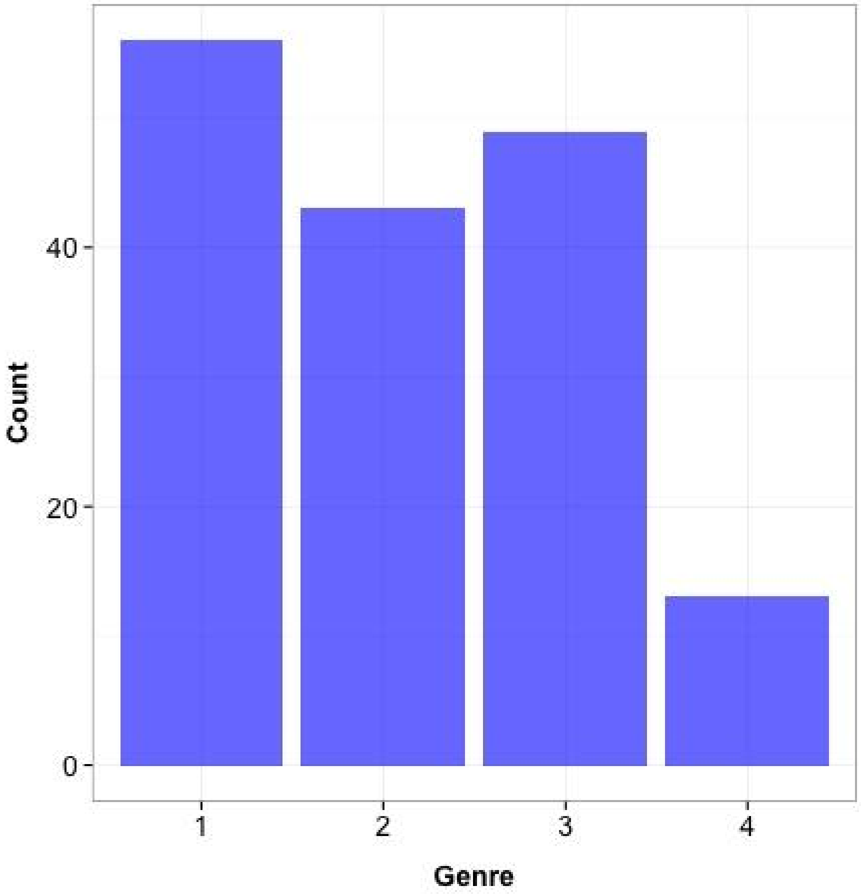
  Theaters= 3611, Rating= 57, Action= 1

- Given this information, what are the predicted first week revenues for the new movie?

Distribution of Opening Week Revenues

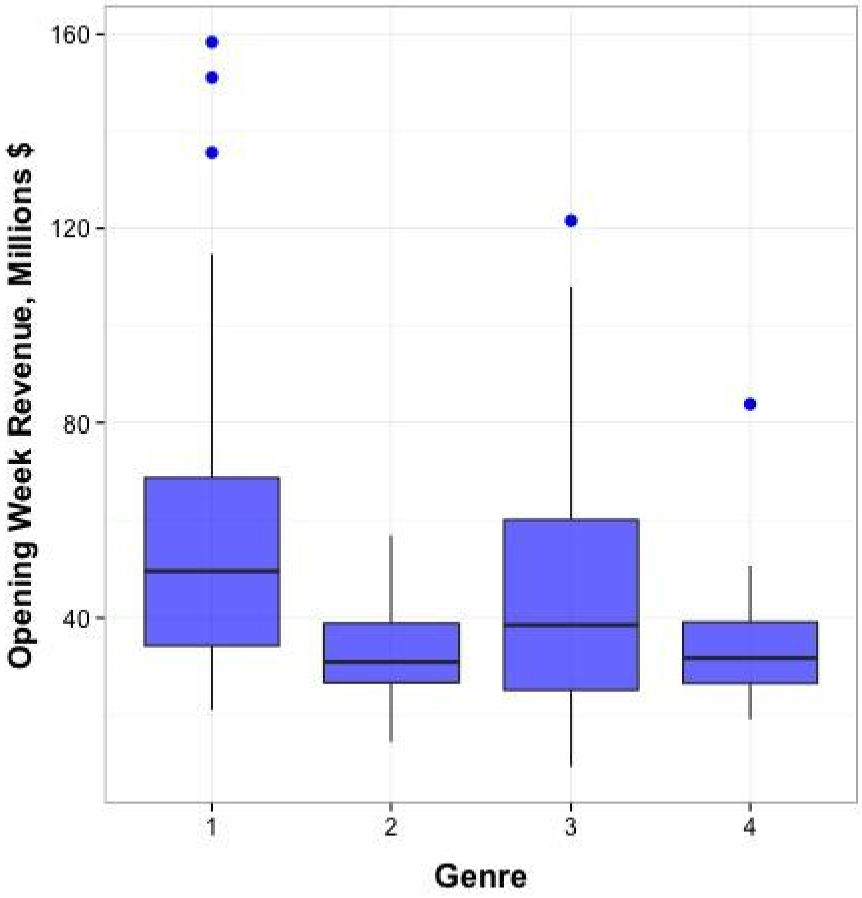


Distribution of Movies by Genre



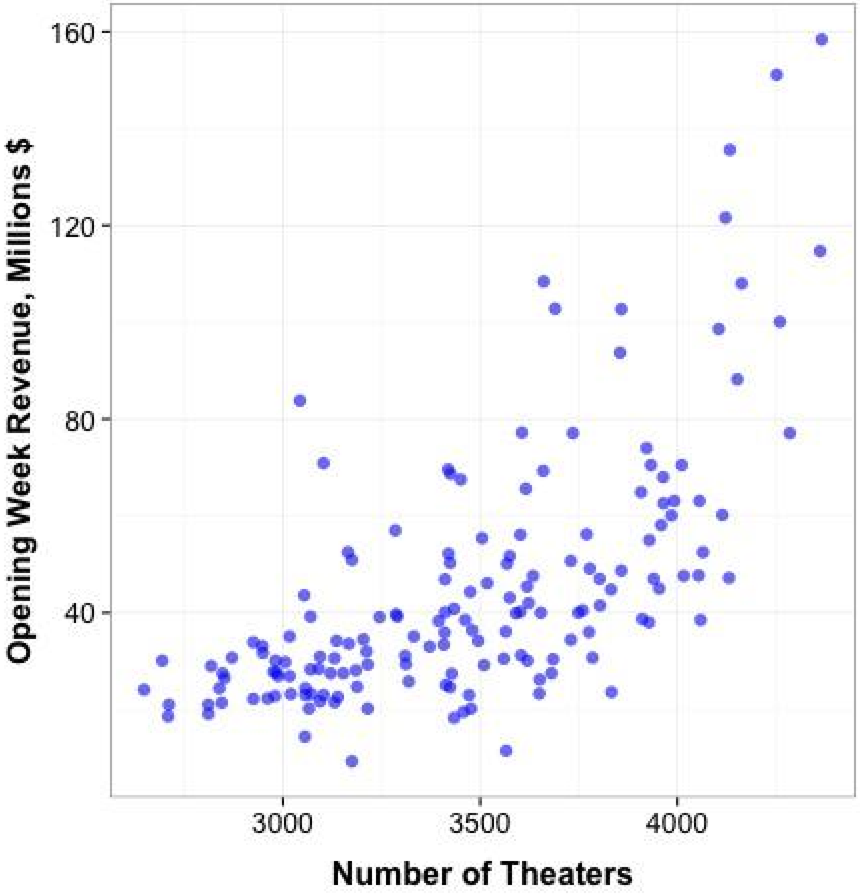
1:Action, 2:Comedy, 3:Kids, 4: Other

Distribution of Opening Revenues by Genre



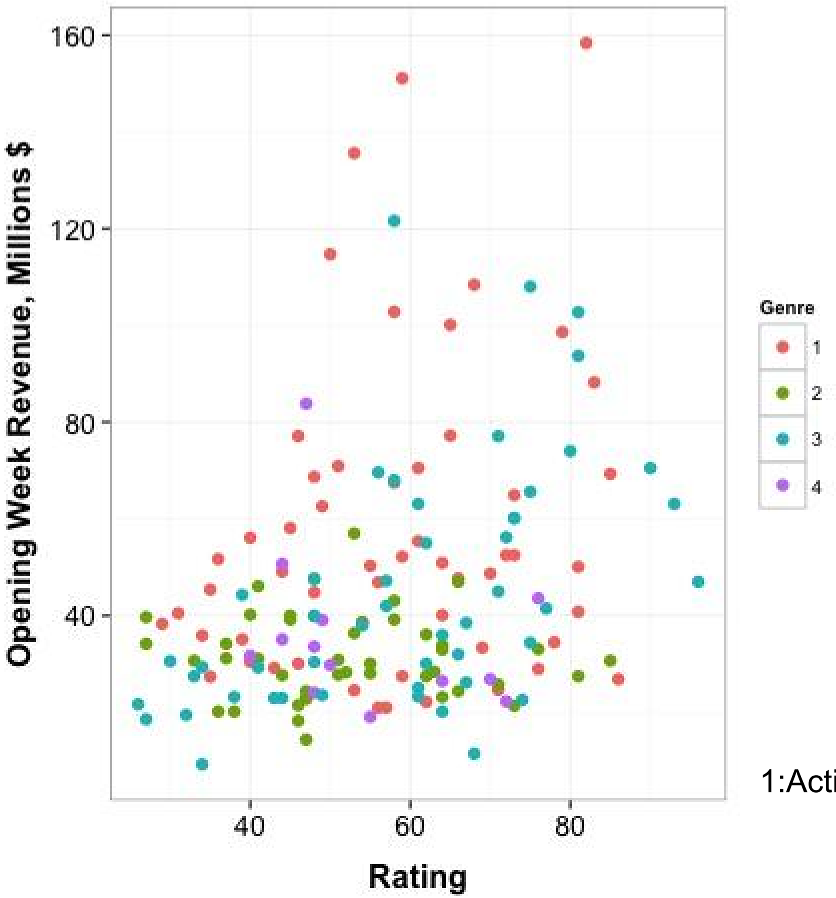
1:Action, 2:Comedy, 3:Kids, 4: Other

Distribution of Opening Revenues by Number of Theaters



NYU TANDON SCHOOL OF ENGINEERING

Distribution of Opening Revenues by Genre



1:Action, 2:Comedy, 3:Kids, 4: Other

Correlation Matrix

Correlation Coefficients Matrix						
Sample size	161	Critical value (10%)	1.65449			
			Opening_Week_Revenue	Num_Theaters	Overall_Rating	Genre
Opening_Week_Revenue	**Pearson Correlation Coefficient**		1.			
	R Standard Error					
	t					
	p-value					
	H0 (10%)					
Num_Theaters	**Pearson Correlation Coefficient**		0.65722	1.		
	R Standard Error		0.00357			
	t		10.99545			
	p-value		0.E+0			
	H0 (10%)		rejected			
Overall_Rating	**Pearson Correlation Coefficient**		0.30326	0.22071	1.	
	R Standard Error		0.00571	0.00598		
	t		4.013	2.85335		
	p-value		0.00009	0.0049		
	H0 (10%)		rejected	rejected		
Genre	**Pearson Correlation Coefficient**		-0.21327	-0.09862	0.00124	1.
	R Standard Error		0.006	0.00623	0.00629	
	t		-2.75262	-1.24969	0.01559	
	p-value		0.0066	0.21325	0.98758	
	H0 (10%)		rejected	accepted	accepted	

R

Variable vs. Variable	R
Num_Theaters vs. Opening_Week_Revenue	0.65722
Overall_Rating vs. Opening_Week_Revenue	0.30326
Overall_Rating vs. Num_Theaters	0.22071
Genre vs. Opening_Week_Revenue	-0.21327
Genre vs. Num_Theaters	-0.09862
Genre vs. Overall_Rating	0.00124

Interpreting Correlation Coefficients

Exactly –1 → A perfect downhill (negative) linear relationship

–0.70 → A strong downhill (negative) linear relationship

–0.50 → A moderate downhill (negative) relationship

–0.30 → A weak downhill (negative) linear relationship

0 → No linear relationship

+0.30 → A weak uphill (positive) linear relationship

+0.50 → A moderate uphill (positive) relationship

+0.70 → A strong uphill (positive) linear relationship

Exactly +1 → A perfect uphill (positive) linear relationship

Linear Regression in R

```
lm(formula = Opening_Week_Revenue ~ Num_Theaters + Overall_Rating +
    Genre, data = movies)

Residuals:
    Min      1Q  Median      3Q     Max
-32.163 -11.710  -2.718   7.488  64.794

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -110.31172   14.99351  -7.357 1.02e-11 ***
Num_Theaters     0.04238    0.00411  10.310  < 2e-16 ***
Overall_Rating   0.27838    0.09620   2.894  0.00436 **
Genre2         -10.21687    3.92821  -2.601  0.01020 *
Genre3         -16.19055    3.60622  -4.490 1.39e-05 ***
Genre4           1.34393    5.99047   0.224  0.82279
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.32 on 155 degrees of freedom
Multiple R-squared:  0.5307,    Adjusted R-squared:  0.5156
F-statistic: 35.06 on 5 and 155 DF,  p-value: < 2.2e-16
```

Linear Regression in R

	Coefficients	Standard Error	LCL	UCL	t Stat	p-level	H0 (10%) rejected?
Intercept	-97.63324	13.93958	-120.6979	-74.56858	-7.00403	6.86752E-11	Yes
Num_Theaters	0.0389	0.00381	0.03259	0.0452	10.2088	0.E+0	Yes
Overall_Rating	0.28838	0.09994	0.12302	0.45374	2.88551	0.00446	Yes
Genre	-4.11685	1.54532	-6.67377	-1.55994	-2.66407	0.00853	Yes
T (10%)	1.65462						

LCL - Lower value of a reliable interval (LCL)
UCL - Upper value of a reliable interval (UCL)

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -110.31172   14.99351  -7.357 1.02e-11 ***
Num_Theaters    0.04238    0.00411  10.310  < 2e-16 ***
Overall_Rating  0.27838    0.09620   2.894  0.00436 ** 
Genre2        -10.21687    3.92821  -2.601  0.01020 *  
Genre3        -16.19055    3.60622  -4.490 1.39e-05 ***
Genre4          1.34393    5.99047   0.224  0.82279    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.32 on 155 degrees of freedom
Multiple R-squared:  0.5307,    Adjusted R-squared:  0.5156 
F-statistic: 35.06 on 5 and 155 DF,  p-value: < 2.2e-16
```

Data Manipulation

Convert the “Genre” variable into a series of dummy variables.

- A dummy variable is an artificial variable created to represent an attribute with two or more distinct categories/levels.
- The total number of dummy variables needed is 1 less than the number of categories. The left out category is absorbed in the intercept.
- It does not matter what you leave out — all included dummy variables will be interpreted with respect to what you leave out.

Movie	Opening_Week _Revenue	Num_Theaters	Overall _Rating	Genre1	Genre2	Genre3	Genre4
Van Helsing	51.7	3575	36	1	0	0	0
Collateral	24.7	3188	71	1	0	0	0
Alien Vs. Predator	38.3	3395	29	1	0	0	0
Man on Fire	22.8	2980	47	1	0	0	0
Sex and the City	57	3285	53	0	1	0	0
Marley and Me	36.4	3480	53	0	1	0	0
Four Christmases	31.1	3310	41	0	1	0	0
Tropic Thunder	25.8	3319	71	0	1	0	0

Dummy Variables

- Compare averages to regression with dummy variables only.
- We left out “Action” in the model.

Opening Week Revenue

Genre	Mean	N	Std Deviation
Action	56.664	56	32.09
Comedy	31.981	43	8.87
Kids	45.104	49	25.59
Other	35.869	13	16.88

Model	Estimate	Std Error	t value
(Intercept)	56.664	3.284	17.256
Comedy	-24.683	4.983	-4.954
Kids	-11.56	4.807	-2.405
Other	-20.795	7.565	-2.749

Model	Estimate	Std Error	t value
(Intercept)	31.981	3.747	8.534
Action	24.683	4.983	4.954
Kids	13.123	5.135	2.556
Other	3.888	7.778	0.5

- Or leave out “Comedy”
- The model fit doesn’t change. The coefficients get adjusted based on the left out category.

```
Call:
lm(formula = Opening_Week_Revenue ~ Num_Theaters + Overall_Rating +
    Comedy + Kids + Other, data = movies)

Residuals:
    Min      1Q  Median      3Q     Max
-32.163 -11.710  -2.718   7.488  64.794

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -110.31172   14.99351  -7.357 1.02e-11 ***
Num_Theaters      0.04238    0.00411  10.310  < 2e-16 ***
Overall_Rating    0.27838    0.09620   2.894  0.00436 **
ComedyTRUE      -10.21687    3.92821  -2.601  0.01020 *
KidsTRUE        -16.19055    3.60622  -4.490 1.39e-05 ***
OtherTRUE         1.34393    5.99047   0.224  0.82279
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.32 on 155 degrees of freedom
Multiple R-squared:  0.5307,    Adjusted R-squared:  0.5156
F-statistic: 35.06 on 5 and 155 DF,  p-value: < 2.2e-16
```

Understanding Model Strength

- R^2 / Multiple R-squared is called the coefficient of determination.
  - represents the proportion of the total variation explained by the linear relationship
- It is always between 0 and 1.
- A larger R^2 value indicates that the linear regression model has more explaining power.
- Rule of thumb:

  $.65 \leq R^2 \leq 1$: strong model

  $.25 \leq R^2 < .65$: the model has moderate strength

  $0 \leq R^2 < .25$: the model is weak; hardly worth considering in its present form

Significance of Variables

```
Call:
lm(formula = Opening_Week_Revenue ~ Num_Theaters + Overall_Rating +
    Comedy + Kids + Other, data = movies)

Residuals:
    Min      1Q  Median      3Q     Max
-32.163 -11.710  -2.718   7.488  64.794

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -110.31172   14.99351  -7.357 1.02e-11 ***
Num_Theaters      0.04238    0.00411  10.310  < 2e-16 ***
Overall_Rating    0.27838    0.09620   2.894  0.00436 **
ComedyTRUE      -10.21687    3.92821  -2.601  0.01020 *
KidsTRUE        -16.19055    3.60622  -4.490 1.39e-05 ***
OtherTRUE         1.34393    5.99047   0.224  0.82279
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.32 on 155 degrees of freedom
Multiple R-squared:  0.5307,    Adjusted R-squared:  0.5156
F-statistic: 35.06 on 5 and 155 DF,  p-value: < 2.2e-16
```

Statistical Significance

- Statistical significance is the likelihood that the difference in conversion rates between a given variation and the baseline is not due to random chance.
- A result of an experiment is statistically significant if it is likely not caused by chance for a given statistical significance level.
- Your statistical significance level reflects your risk tolerance and confidence level. For example, if results of your analysis has a significance level of 95%, this means that you can be 95% confident that the observed results are real and not caused by randomness. It also means that there is a 5% chance that you could be wrong.

Statistical Significance

- Null hypothesis (**H_0**) indicates that there is no significant difference between specified populations, any observed difference is due to sampling, experimental error, or randomness.
- The **p** value, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis is true.
- We use p values to determine statistical significance in a hypothesis test.
- A low p value suggests that your sample provides enough evidence that you can reject the null hypothesis for the entire population.

```
Call:
lm(formula = Opening_Week_Revenue ~ Num_Theaters + Overall_Rating +
    Comedy + Kids + Other, data = movies)

Residuals:
    Min      1Q  Median      3Q     Max
-32.163 -11.710  -2.718   7.488  64.794

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -110.31172   14.99351  -7.357 1.02e-11 ***
Num_Theaters     0.04238    0.00411  10.310  < 2e-16 ***
Overall_Rating   0.27838    0.09620   2.894  0.00436 **
ComedyTRUE     -10.21687    3.92821  -2.601  0.01020 *
KidsTRUE       -16.19055    3.60622  -4.490 1.39e-05 ***
OtherTRUE        1.34393    5.99047   0.224  0.82279
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.32 on 155 degrees of freedom
Multiple R-squared:  0.5307,    Adjusted R-squared:  0.5156
F-statistic: 35.06 on 5 and 155 DF,  p-value: < 2.2e-16
```

- **t-value**: comparing our sample populations and determining if there is a significant difference between their means.
- **p-value**: the probability that ‘t’ falls into a certain range (confidence intervals).
    - a p-value $\leq$ 0.05 suggests a significant difference between the means of our sample population and we would reject our null hypothesis.
- **Null Hypothesis:** Usually written in the following form: “There is no significant difference between population A and population B.“

Interpretation

Each additional theater the movie is shown in increases the opening week revenue by $0.04MM ($40K).

```
lm(formula = Opening_Week_Revenue ~ Num_Theaters + Overall_Rating +
    Comedy + Kids + Other, data = movies)

Residuals:
    Min      1Q  Median      3Q     Max
-32.163 -11.710  -2.718   7.488  64.794

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -110.31172   14.99351  -7.357 1.02e-11 ***
Num_Theaters     0.04238    0.00411  10.310  < 2e-16 ***
Overall_Rating   0.27838    0.09620   2.894  0.00436 **
ComedyTRUE     -10.21687    3.92821  -2.601  0.01020 *
KidsTRUE       -16.19055    3.60622  -4.490 1.39e-05 ***
OtherTRUE        1.34393    5.99047   0.224  0.82279
```

Interpretation

Each additional rating point increases the opening week revenue by $0.28MM ($280K).

```
lm(formula = Opening_Week_Revenue ~ Num_Theaters + Overall_Rating +
    Comedy + Kids + Other, data = movies)

Residuals:
    Min      1Q  Median      3Q     Max
-32.163 -11.710  -2.718   7.488  64.794

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -110.31172   14.99351  -7.357 1.02e-11 ***
Num_Theaters     0.04238    0.00411  10.310  < 2e-16 ***
Overall_Rating   0.27838    0.09620   2.894  0.00436 **
ComedyTRUE     -10.21687    3.92821  -2.601  0.01020 *
KidsTRUE       -16.19055    3.60622  -4.490 1.39e-05 ***
OtherTRUE        1.34393    5.99047   0.224  0.82279
```

Interpretation

Comedies bring in $10.2MM less in opening week revenue than action films.

```
lm(formula = Opening_Week_Revenue ~ Num_Theaters + Overall_Rating +
    Comedy + Kids + Other, data = movies)

Residuals:
    Min      1Q  Median      3Q     Max
-32.163 -11.710  -2.718   7.488  64.794

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -110.31172   14.99351  -7.357 1.02e-11 ***
Num_Theaters      0.04238    0.00411  10.310  < 2e-16 ***
Overall_Rating    0.27838    0.09620   2.894  0.00436 **
ComedyTRUE      -10.21687    3.92821  -2.601  0.01020 *
KidsTRUE        -16.19055    3.60622  -4.490 1.39e-05 ***
OtherTRUE         1.34393    5.99047   0.224  0.82279
```

Interpretation

Num_Theaters: Each additional theater the movie is shown in increases the opening week revenue by $0.04MM ($40K).

Overall_Rating: Each additional rating point increases the opening week revenue by $0.28MM ($280K)

Comedy: Comedies bring in $10.2MM less in opening week revenue than action films.

Kids: Kids films bring in $16.2MM less revenue than action films.

Other: Other movie category brings in $1.34MM more in operating week revenue than action films. However, this effect is not statistically significant.

Prediction

- The attributes for the new movie “You Name It” are as follows:

  Theaters= 3611, Rating= 57, Action= 1

- Given this information, what are the predicted first week revenues for the new movie?

Other Type of Regressions

Generalized linear model (GLM): is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The term general linear model (GLM) usually refers to conventional linear regression models for a continuous response variable given continuous and/or categorical predictors. GLM allows to specify a link function "family")

- binomial(link = "logit")
- gaussian(link = "identity")
- Gamma(link = "inverse")
- inverse.gaussian(link = "1/mu^2")
- poisson(link = "log")

```
x1<-c(56.1, 26.8, 23.9, 46.8, 34.8, 42.1, 22.9, 55.5, 56.1, 46.9, 26.7,
33.9, 37.0, 57.6, 27.2, 25.7, 37.0, 44.4, 44.7, 67.2, 48.7, 20.4, 45.2,
22.4, 23.2, 39.9, 51.3, 24.1, 56.3, 58.9, 62.2, 37.7, 36.0, 63.9, 62.5,
44.1, 46.9, 45.4, 23.7, 36.5, 56.1, 69.6, 40.3, 26.2, 67.1, 33.8, 29.9,
25.7, 40.0, 27.5)

x2<-c(12.29, 11.42, 13.59, 8.64, 12.77, 9.9, 13.2, 7.34, 10.67, 18.8, 9.84,
16.72, 10.32, 13.67, 7.65, 9.44, 14.52, 8.24, 14.14, 17.2, 16.21, 6.01,
14.23, 15.63, 10.83, 13.39, 10.5, 10.01, 13.56, 11.26, 4.8, 9.59, 11.87, 11,
12.02, 10.9, 9.5,  10.63, 19.03, 16.71, 15.11, 7.22, 12.6, 15.35, 8.77,
9.81, 9.49, 15.82, 10.94, 6.53)

y<-c(1.54, 0.81, 1.39, 1.09, 1.3, 1.16, 0.95, 1.29, 1.35, 1.86, 1.1, 0.96,
1.03, 1.8, 0.7, 0.88, 1.24, 0.94, 1.41, 2.13, 1.63, 0.78, 1.55, 1.5, 0.96,
1.21, 1.4, 0.66, 1.55, 1.37, 1.19, 0.88, 0.97, 1.56, 1.51, 1.09, 1.23, 1.2,
1.62, 1.52, 1.64, 1.77, 0.97, 1.12, 1.48, 0.83, 1.06, 1.1, 1.21, 0.75)

lm(y ~ x1 + x2)
glm(y ~ x1 + x2, family=gaussian)
glm(y ~ x1 + x2, family=gaussian(link="log"))
```

Other Type of Regressions

Logistic regression: Used extensively in clinical trials, scoring and fraud detection, when the *response is binary* (chance of succeeding or failing, e.g. for a new tested drug or a credit card transaction).

Can be well approximated by linear regression after transforming the response (logit transform). Some versions (Poisson or Cox regression) have been designed for a non-binary response, for categorical data (classification), ordered integer response (age groups), and even continuous response (regression trees).

```r
mydata <- read.csv("http://www.ats.ucla.edu/stat/data/binary.csv")
## view the first few rows of the data
str(mydata)
dim(mydata)
summary(mydata)
sapply(mydata, sd)
mylogit <- glm(admit ~ gre + gpa + rank, data = mydata, family = "binomial")
mylogit
summary(mylogit)
```

Other Type of Regressions

- **Ridge regression**: A more robust version of linear regression, putting constraints on regression coefficients to make them much more natural, less subject to overfitting, and easier to interpret.
- **Lasso regression**: Similar to ridge regression, but automatically performs variable reduction (allowing regression coefficients to be zero).
- **Ecologic regression**: Consists in performing one regression per strata, if your data is segmented into several rather large core strata, groups, or bins.
- **Bayesian regression**: the statistical analysis is undertaken within the context of Bayesian inference
- **Quantile regression:** Used in connection with extreme events,
- **Jackknife regression:** New type of regression. It solves all the drawbacks of traditional regression. Requires advanced parameter setting

Classification

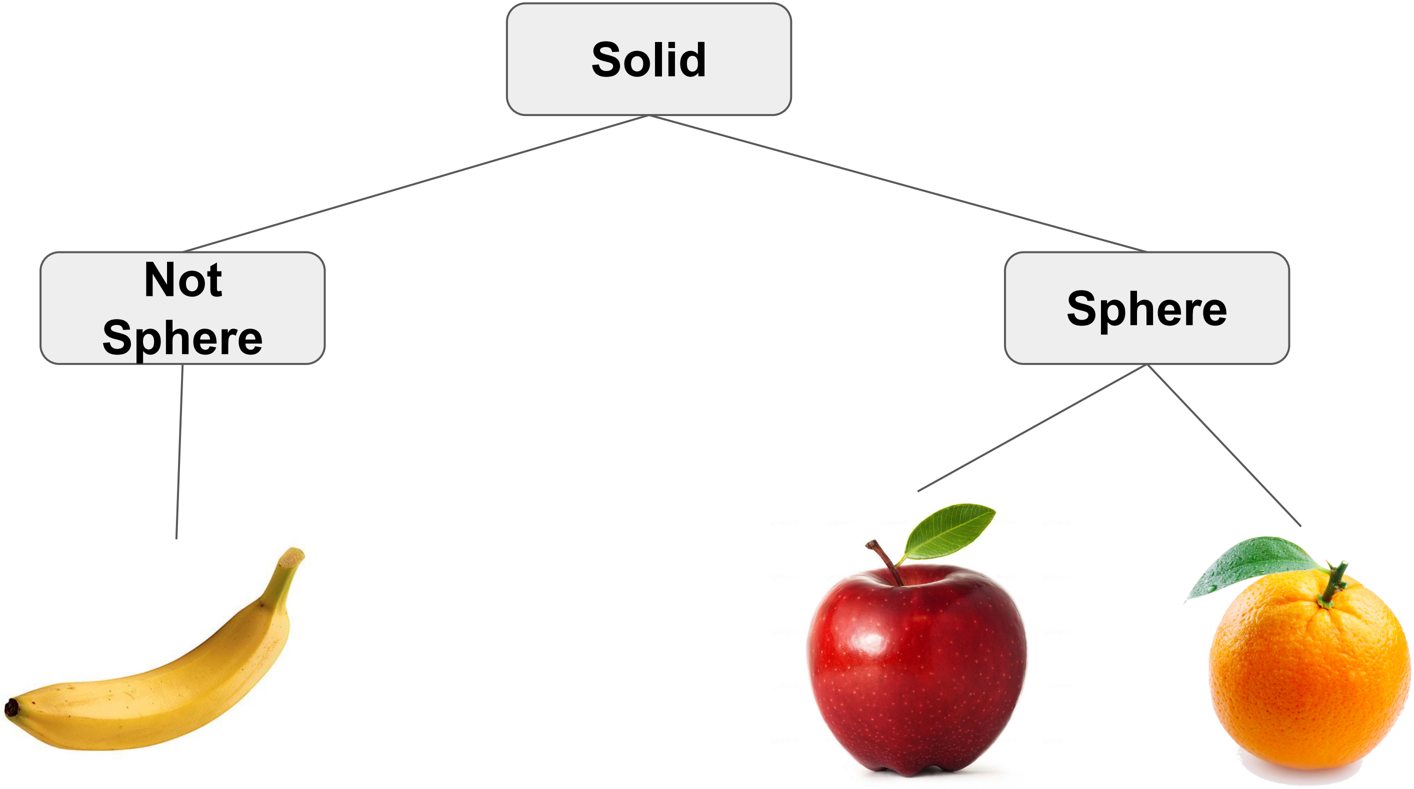
Classification is the task of assigning objects to one of several predefined categories.

- detecting spam emails based on message header and content
- segmenting customers based on their response to an offer
- categorizing loan applications according to their risk level

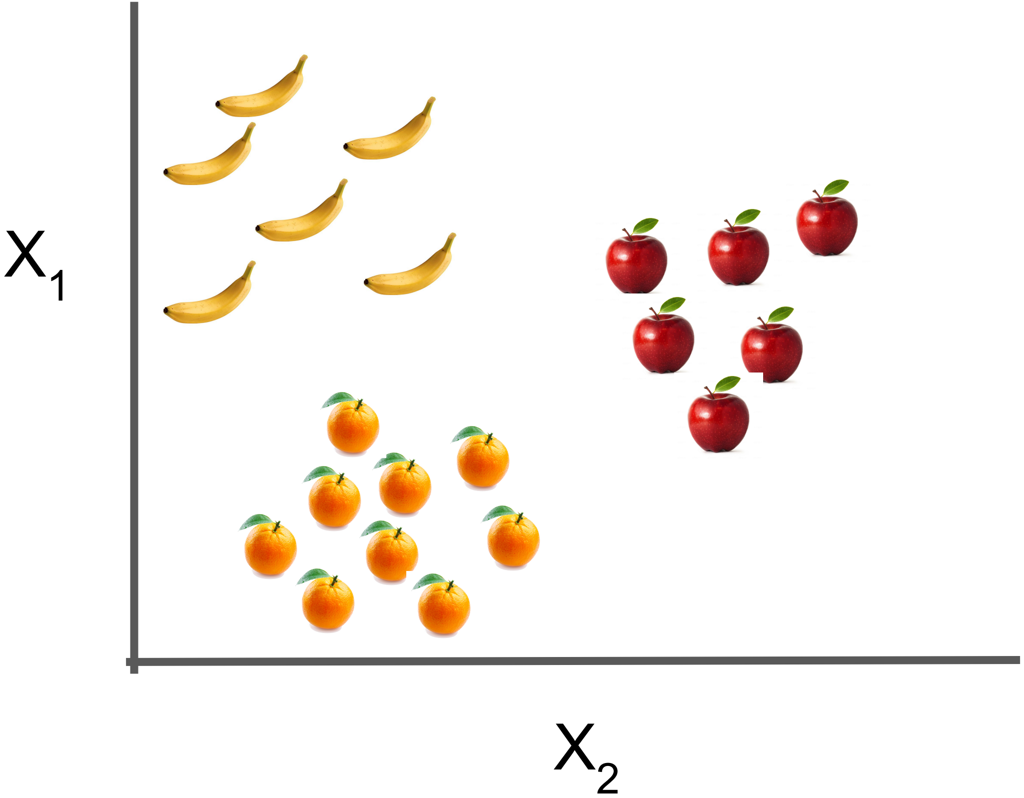
Classification

- A classification model can serve as an explanatory tool to distinguish between objects of different classes -- descriptive analytics
- It can also be used to predict the class label of unknown records -- predictive analytics
- Classification techniques are most suited for predicting or describing data sets with binary or nominal categories.
    - They are less effective for ordinal categories (e.g.: classify a person as a member of high-, medium-, or low-income group) because they do not consider the implicit order among the categories.
- Examples of classification techniques include decision tree classifiers, neural networks, support vector machines, naive Bayes classifiers...

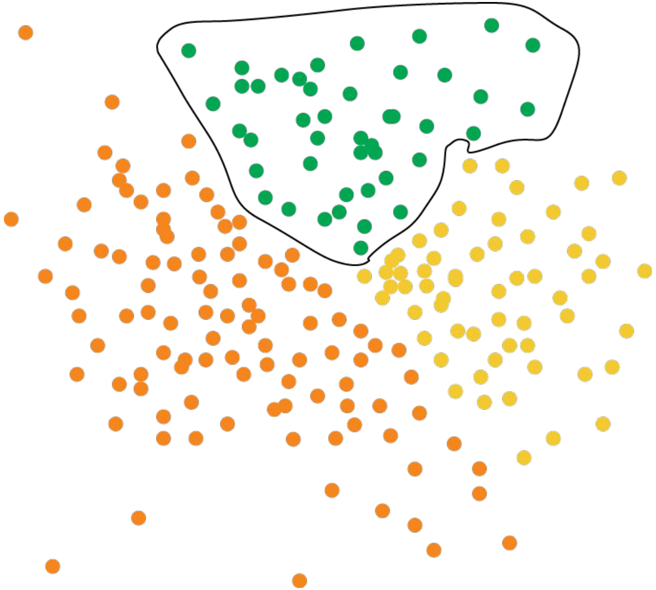
Classification Example

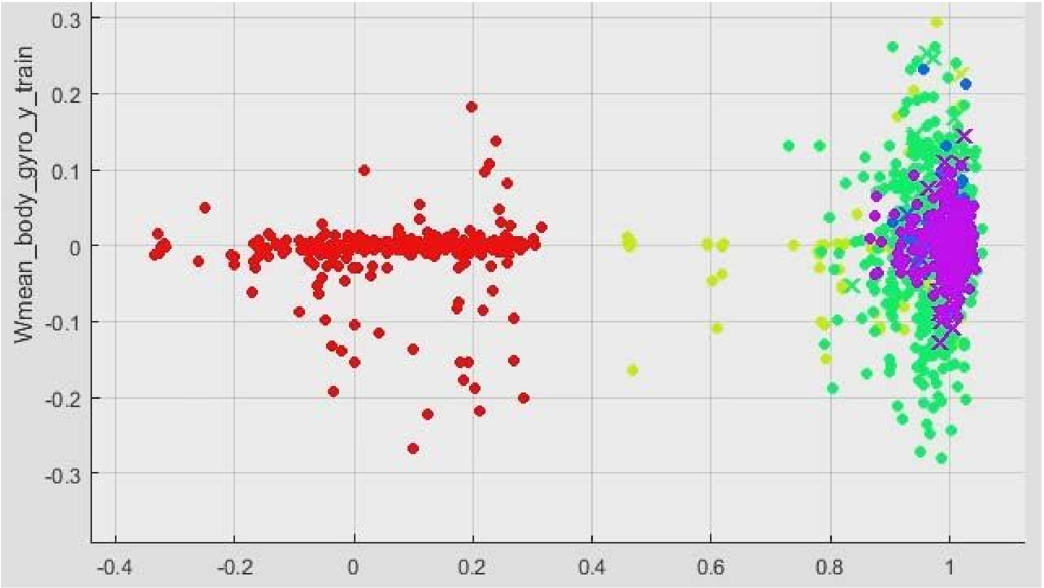


Statistical Classification

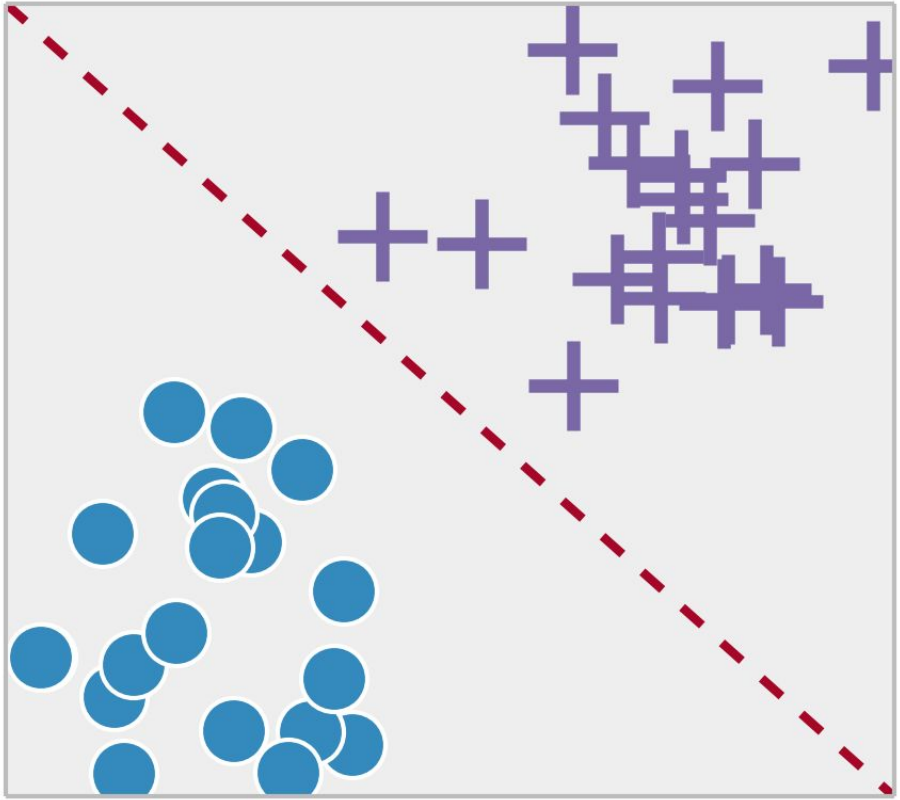


Statistical Classification Examples



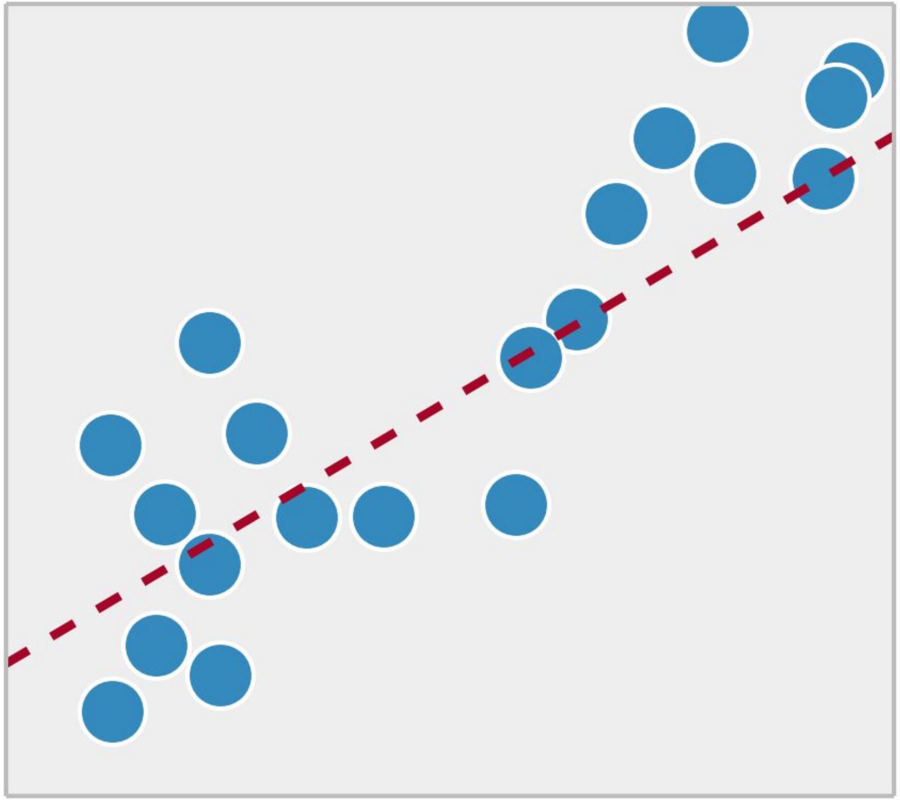


Classification



Output: Discrete (labels); Decision boundary
Evaluation: Accuracy;

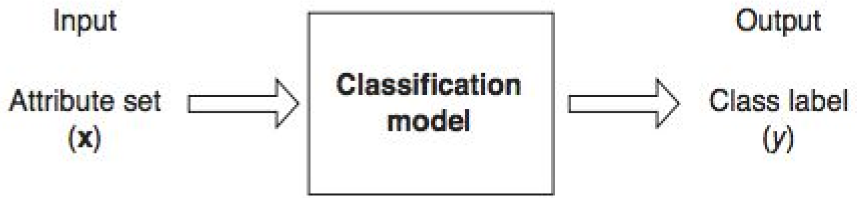
Regression



Output: Continuous (number); best fit line
Evaluation: Sum of Errors; R^2

Modeling Process

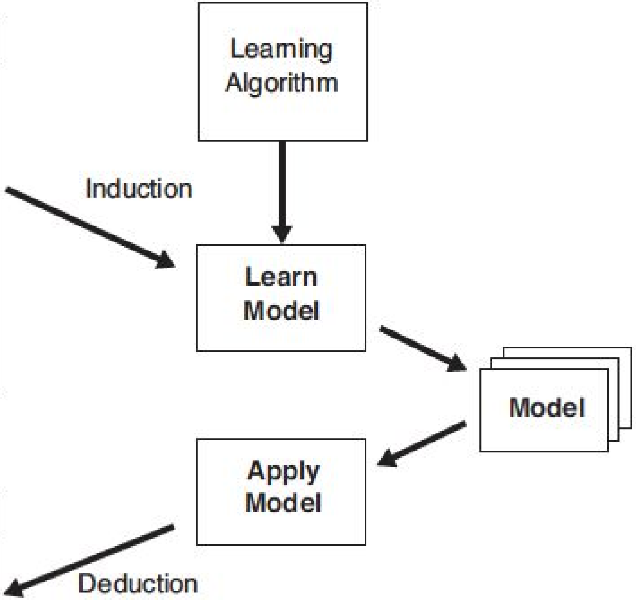
- Employ a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of the input data.
- The model should both fit the input data well and correctly predict the class labels of records it has never seen before.
  - training set
  - test set
- Key objective is to build models with good generalization capability.



Recap of Modeling Process

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes



Test Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Classification

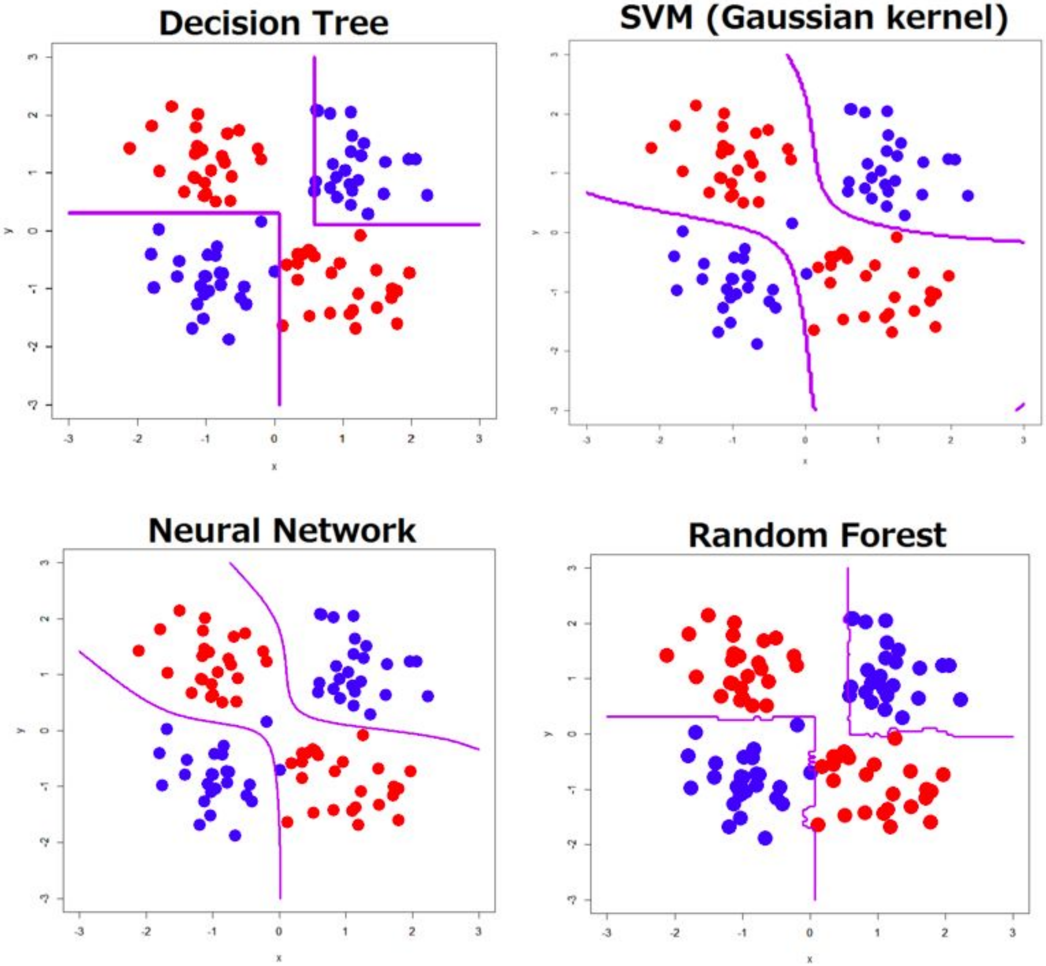
- Evaluation of the performance of a classification model is based on the counts of test records correctly and incorrectly predicted by the model.
    - confusion matrix

		Predicted Class	
		$Class = 1$	$Class = 0$
Actual Class	$Class = 1$	f_{11}	f_{10}
	$Class = 0$	f_{01}	f_{00}

- Other performance metrics:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \qquad \text{Error rate} = \frac{\text{Number of wrong predictions}}{\text{Total number of predictions}}$$

- Base rate: how well would a classifier perform by simply choosing that class for every instance

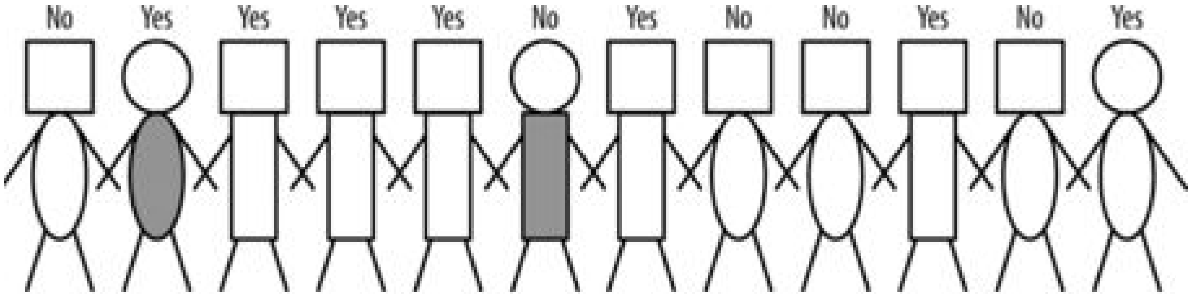

Decision Tree
SVM (Gaussian kernel)
Neural Network
Random Forest
x
y
-3 -2 -1 0 1 2 3

Decision Trees

- Pose a series of questions about the characteristics of the target variable.
  - A follow-up question is asked until a conclusion is reached about the class label of the record
- The series of questions and their possible answers can be organized in the form of a decision tree.
  - nodes -- root node, internal nodes, leaf or terminal nodes
  - directed edges
- Each leaf node is assigned a class label.
- Non-terminal nodes contain attribute test conditions to separate records that have different characteristics.

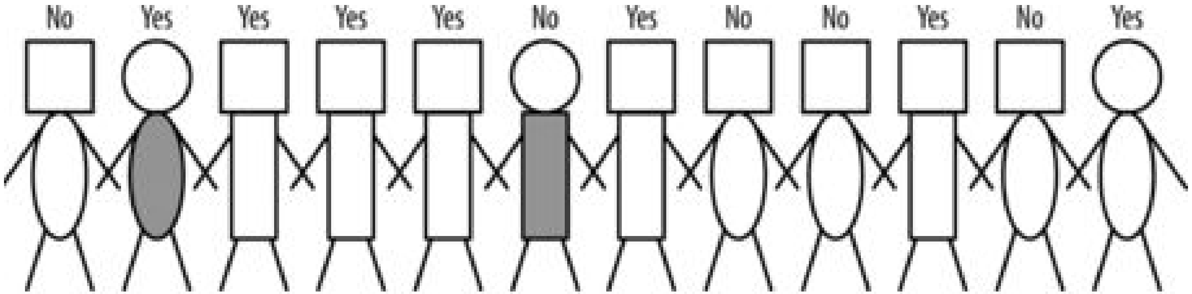
Classification Problem

- Determining whether a customer becomes a loan write-off
  - Binary classification problem with target variable “yes” or “no”



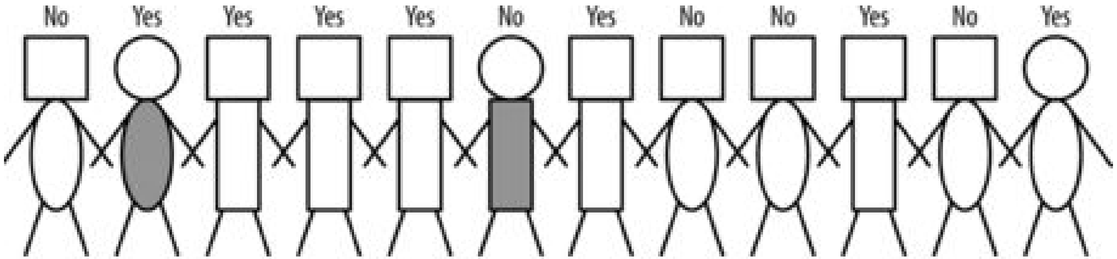
Classification Problem

- Determining whether a customer will default on a loan
  - Binary classification problem with target variable “yes” or “no”
  - Customers represented as stick figures with three attributes

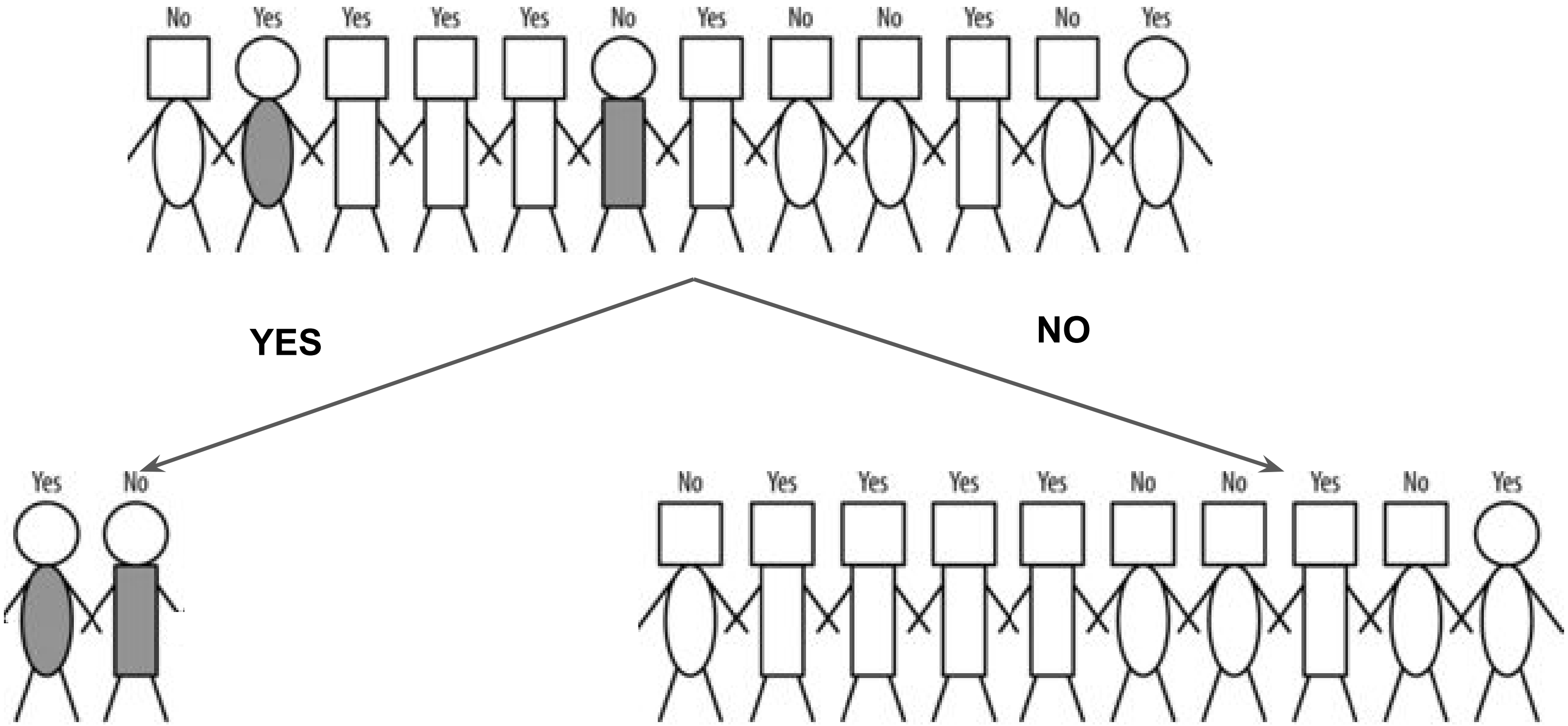


Classification Problem

- Determining whether a customer will default on a loan
  - Binary classification problem with target variable “yes” or “no”
  - Customers represented as stick figures with three attributes
    - head shape
    - body shape
    - body color
  - Which of the attributes would be best to segment these people into groups to distinguish defaults from non defaults?
  - We would like the resultant groups to be as pure as possible with respect to the target variable.

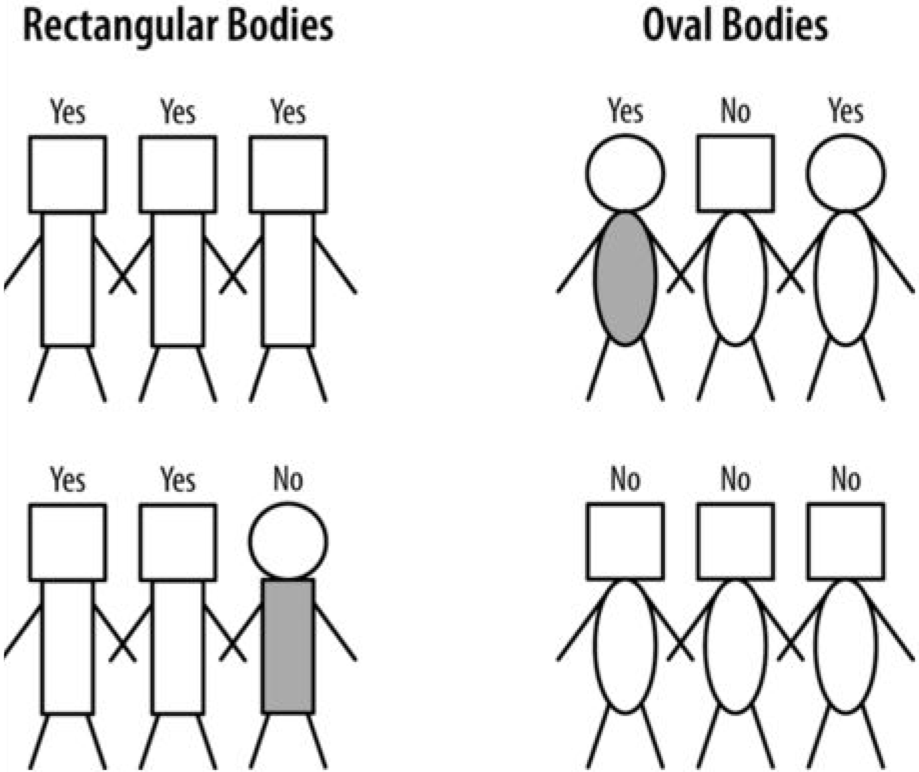


body-color = gray

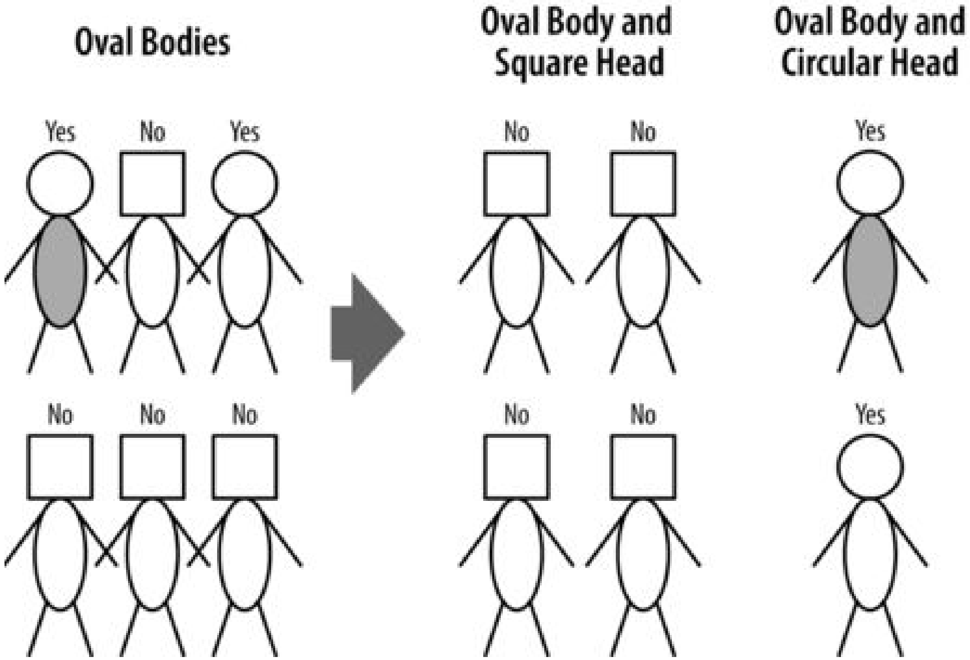


Are these groups pure?

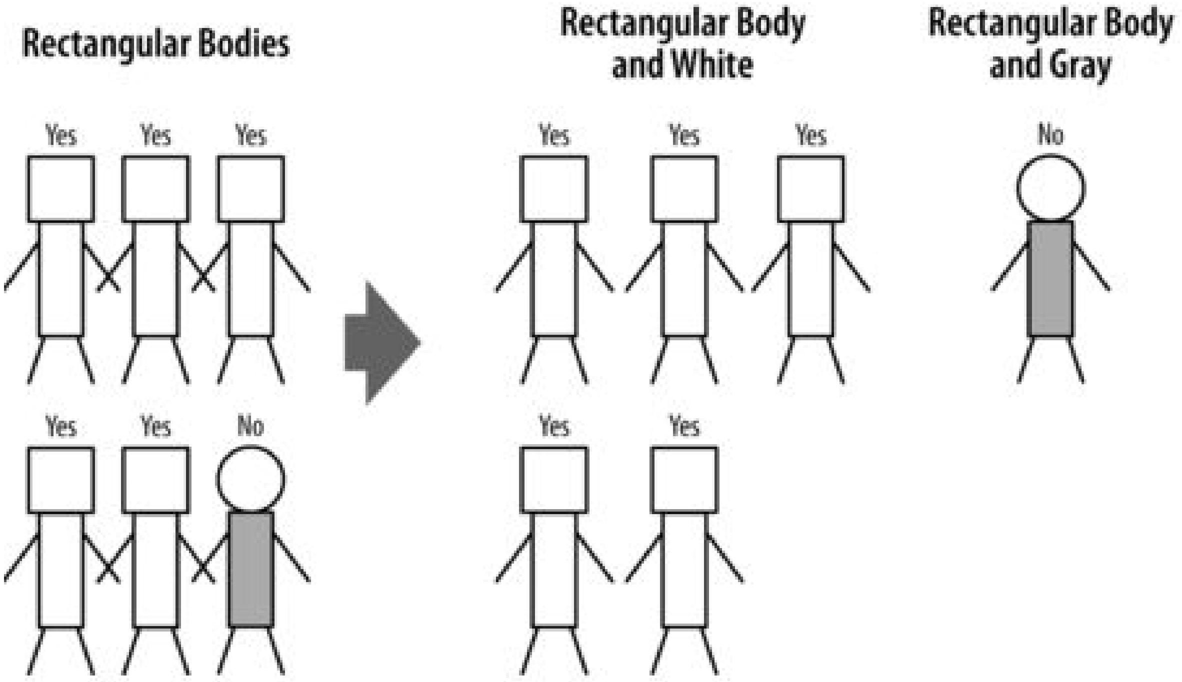
First partitioning: Body shape



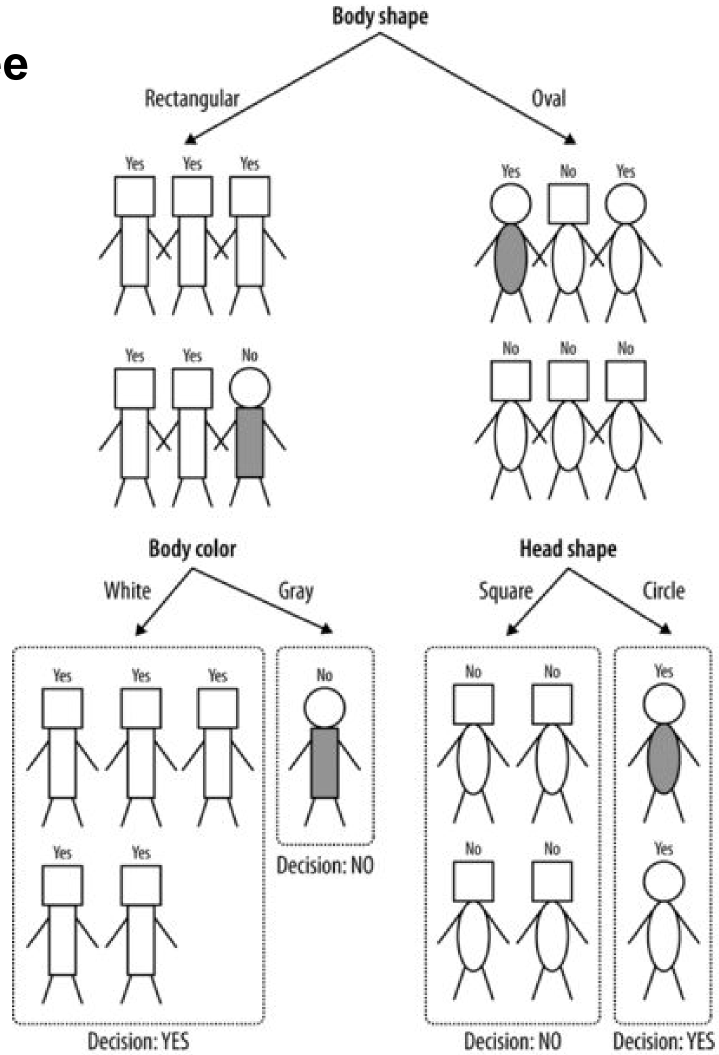
Second partitioning: Oval body people subgrouped by head type



Third partitioning: Rectangular body people subgrouped by body color

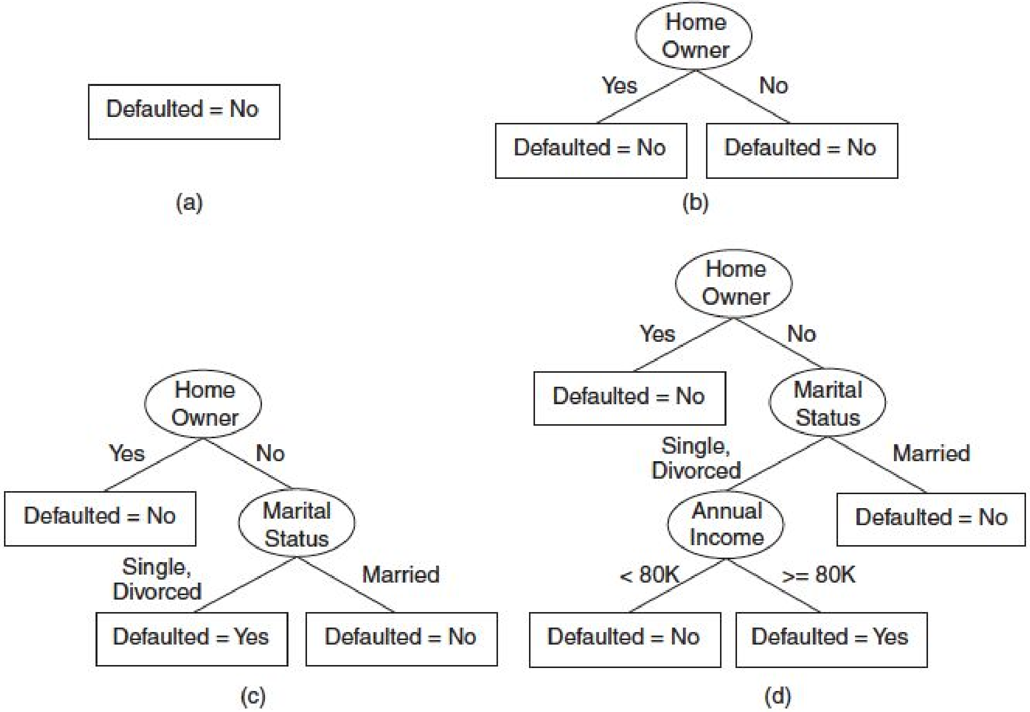


The classification tree resulting from the splits



Classification Problem

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Classification Example

- Data from direct marketing campaigns of a Portuguese banking institution
- The marketing campaigns were based on phone calls.
- Often, more than one contact to the same client was required
- Outcome: customer signed up for a bank term deposit or not
  - subscribe = yes/no
- The classification goal is to predict if a given client will subscribe (yes/no) for a term deposit.
- https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Data Attributes

 # bank client data:

1 - age (numeric)
2 - job : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services")
3 - marital : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
4 - education (categorical: "unknown","secondary","primary","tertiary")
5 - default: has credit in default? (binary: "yes","no")
6 - balance: average yearly balance, in euros (numeric)
7 - housing: has housing loan? (binary: "yes","no")
8 - loan: has personal loan? (binary: "yes","no")

 # related with the last contact of the current campaign:

9 - contact: contact communication type (categorical: "unknown","telephone","cellular")
10 - day: last contact day of the month (numeric)
11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
12 - duration: last contact duration, in seconds (numeric)

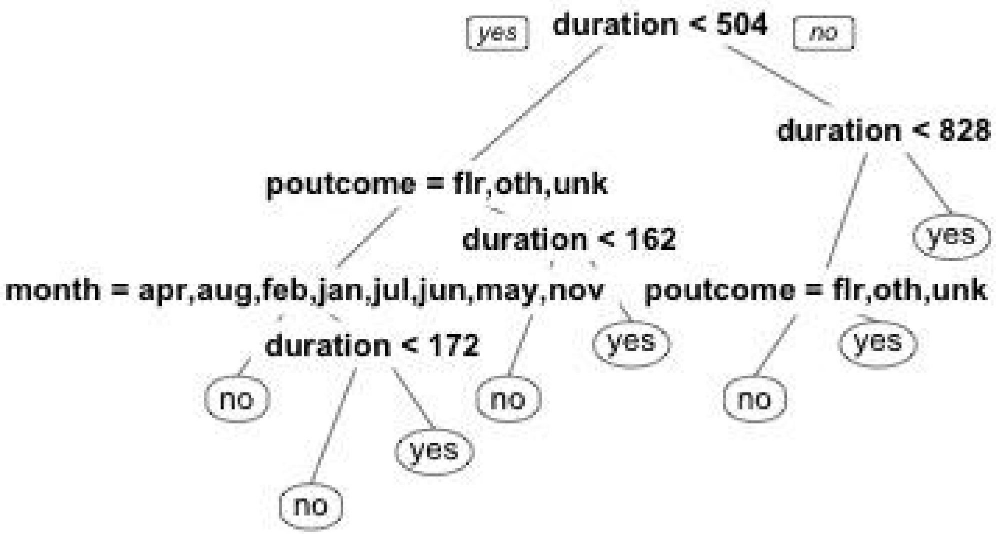
 # other attributes:

13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
15 - previous: number of contacts performed before this campaign and for this client (numeric)
16 - poutcome: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

Output variable (desired target):

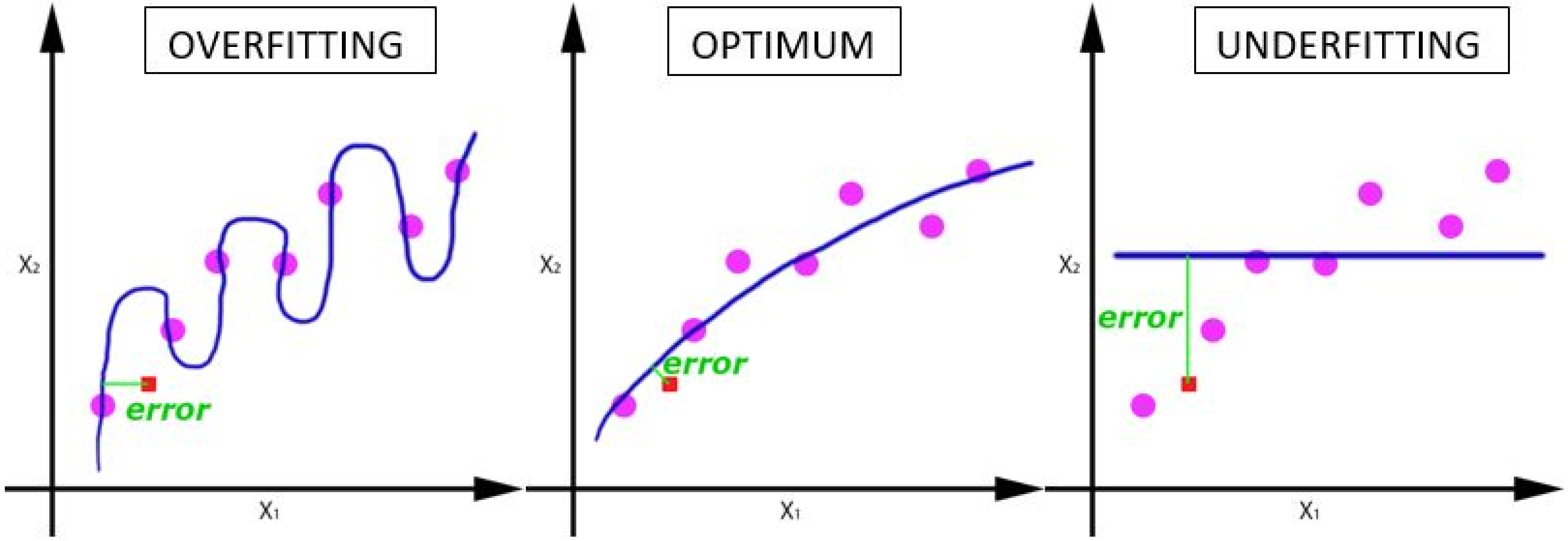
17 - subscribe - has the client subscribed a term deposit? (binary: "yes","no")

Subscribe for Deposit?



Model selection is about goodness of fit

The **goodness of fit** of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question



Performance of a logistic regression (Confusion Matrix): It is nothing but a tabular representation of Actual vs Predicted values. This helps us to find the accuracy of the model and avoid overfitting. This is how it looks like:

		Predicted	
		Good	Bad
Actual	Good	True Positive (d)	False Negative (c)
	Bad	False Positive (b)	True Negative (a)

You can calculate the accuracy of your model with:

$$\frac{\text{True Positive} + \text{True Negatives}}{\text{True Positive} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}}$$

Performance of a logistic regression (ROC Curve): Receiver Operating Characteristic (ROC) summarizes the model's performance by evaluating the trade offs between true positive rate (sensitivity) and false positive rate (1-specificity).

ROC summarizes the predictive power for all possible values of p > 0.5. The area under curve (AUC), referred to as index of accuracy(A) or concordance index, is a perfect performance metric for ROC curve. Higher the area under curve, better the prediction power of the model.