Prediction

Business Analytics

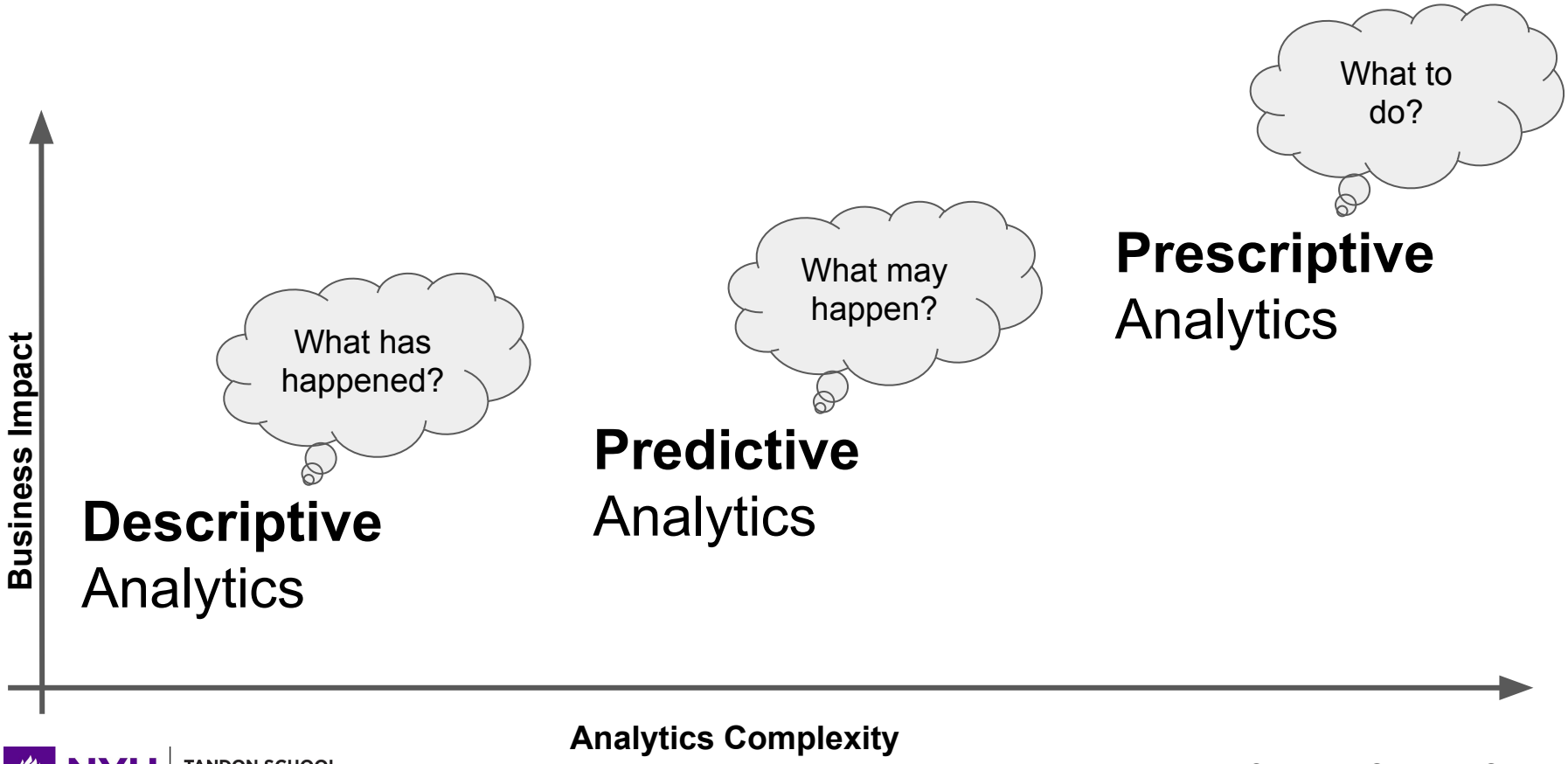
Business Analytics Definition

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



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

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 Models

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

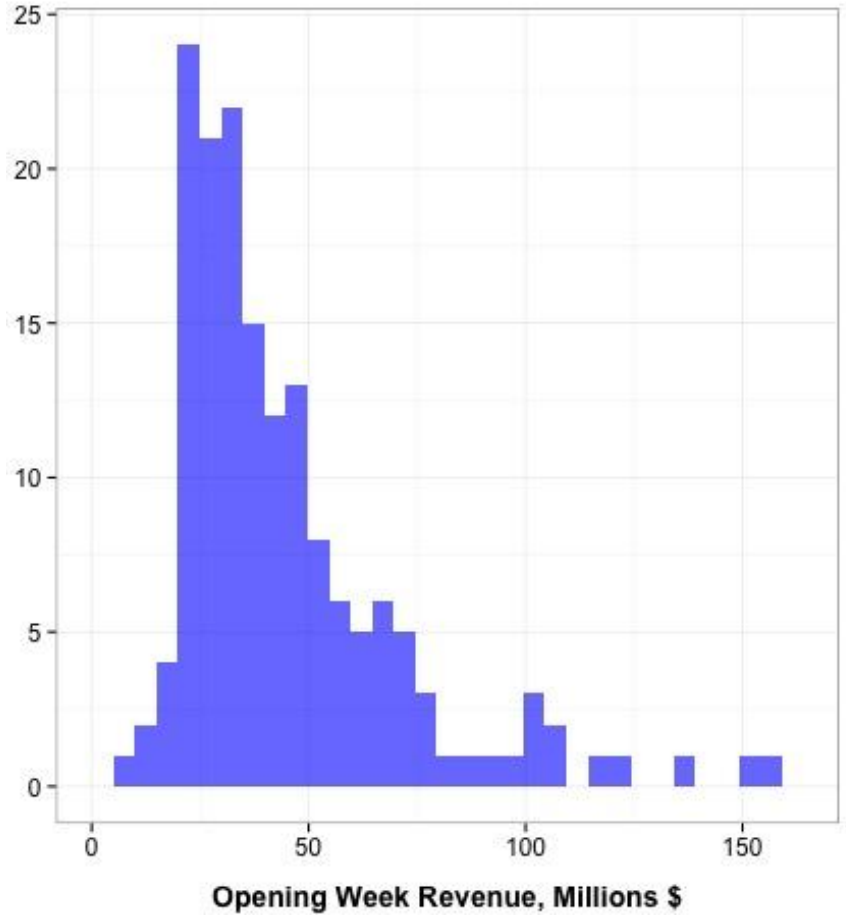
Objective

- 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:

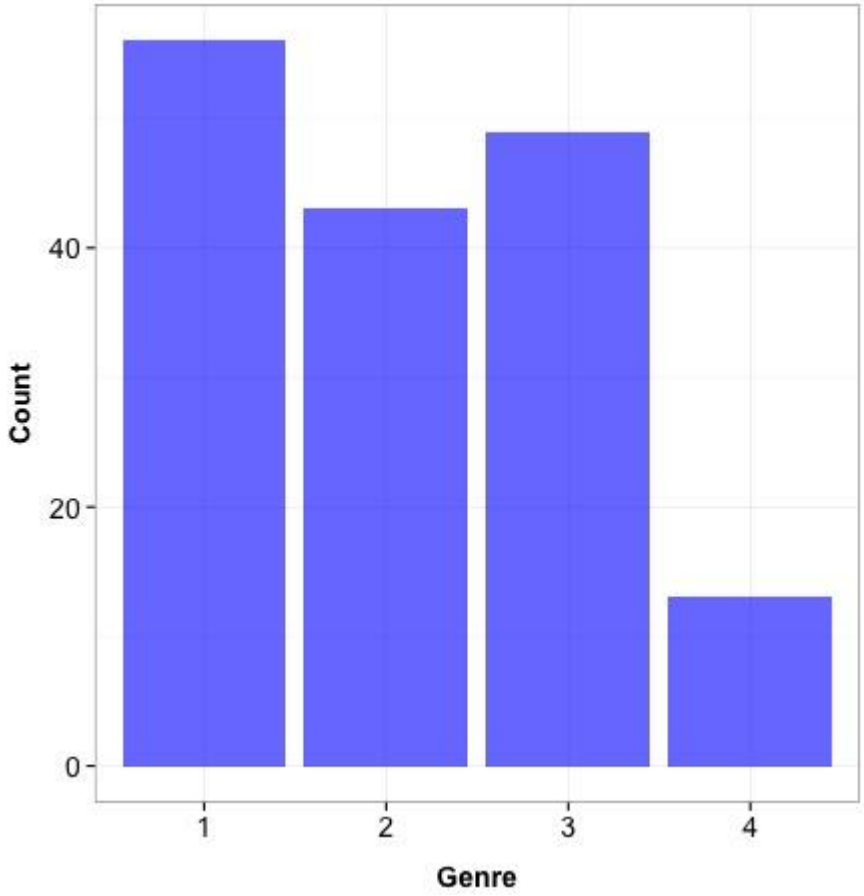
  Theaters= 3611, Rating= 57, Action= 1

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

Distribution of Opening Week Revenues

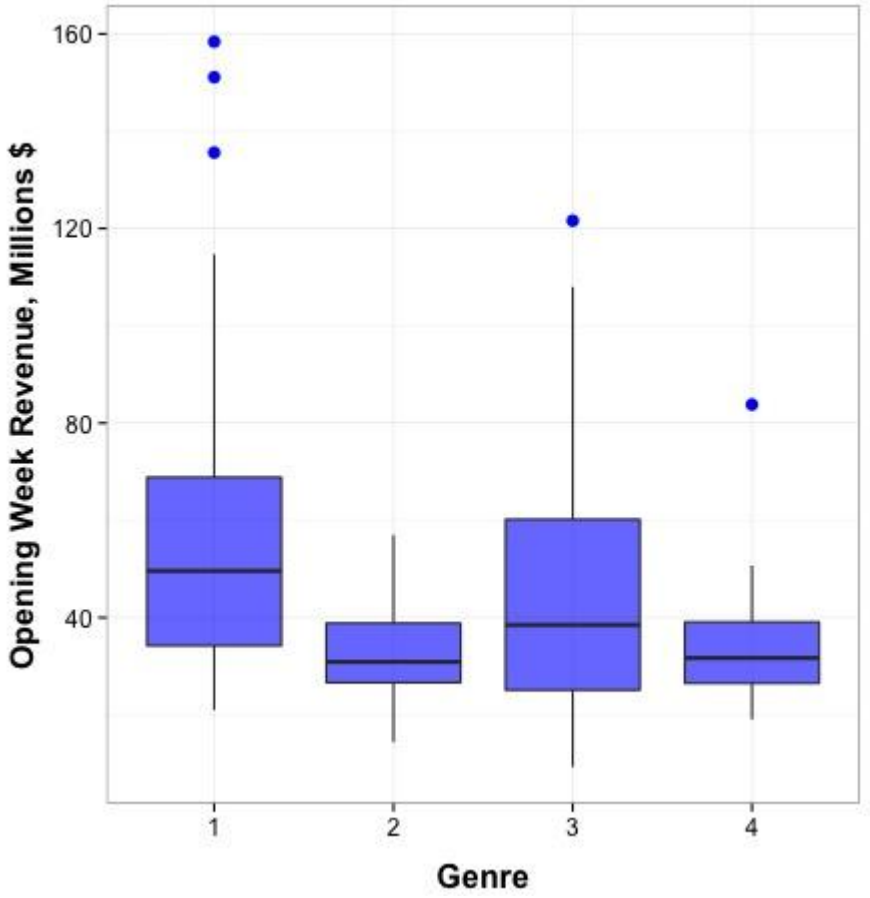


Distribution of Movies by Genre



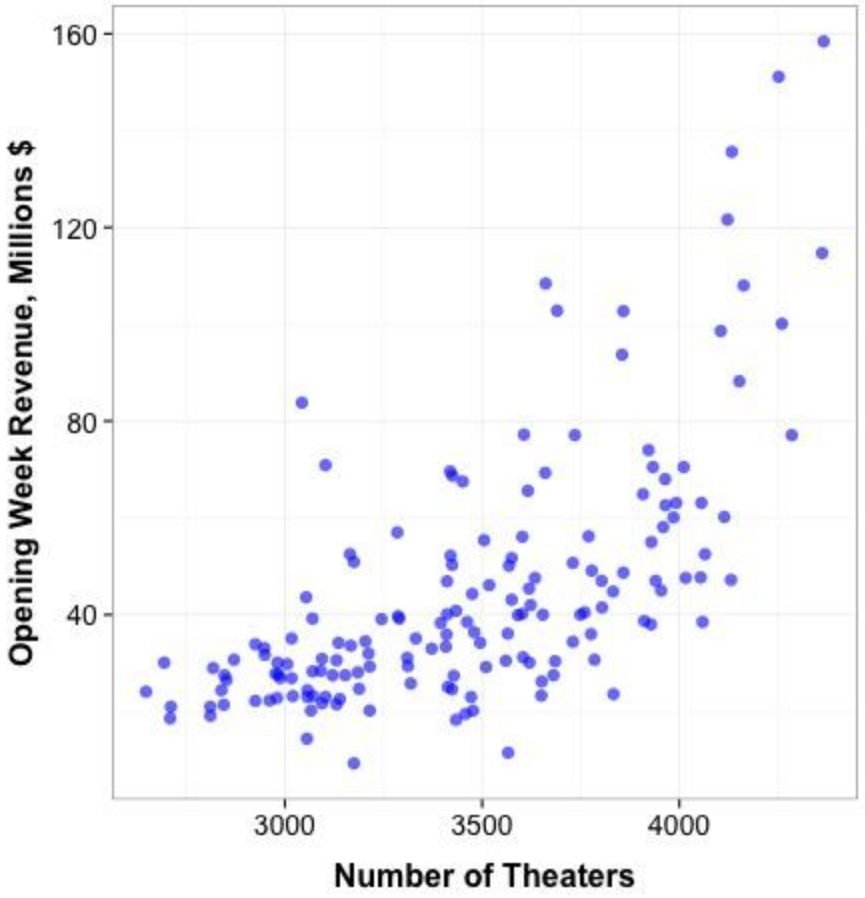
1:Action, 2:Comedy, 3:Kids, 4: Other

Distribution of Opening Revenues by Genre

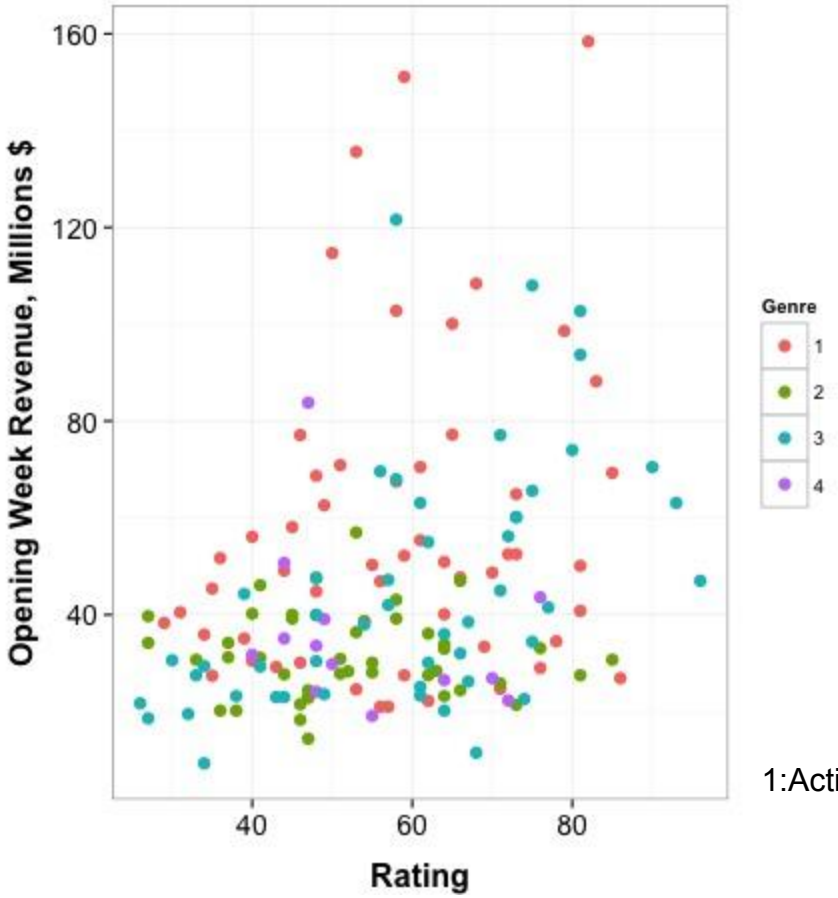


1:Action, 2:Comedy, 3:Kids, 4: Other

Distribution of Opening Revenues by Number of Theaters



Distribution of Opening Revenues by Genre



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

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

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

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

```
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 ($\mathbf{H_0}$) 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 ≤ 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:

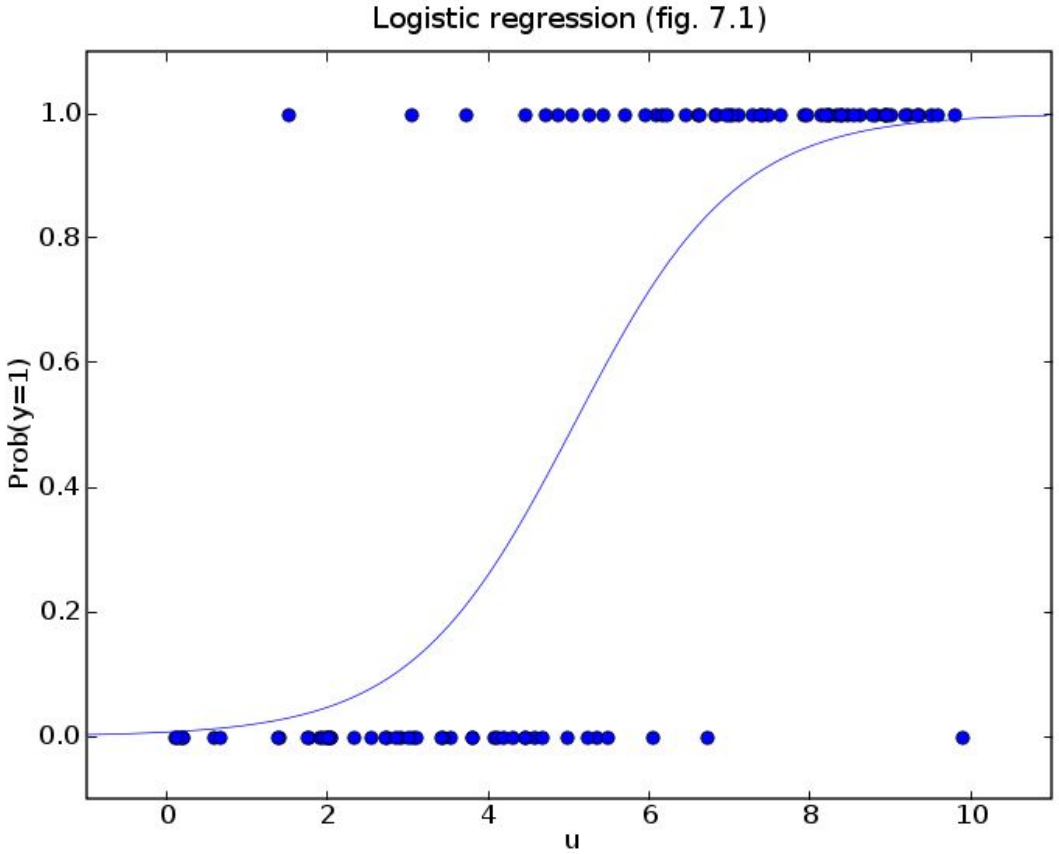
  Theaters= 3611, Rating= 57, Action= 1

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

Other Type of Regressions

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).


Logistic regression (fig. 7.1)
Prob(y=1)
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
u

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")

```
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

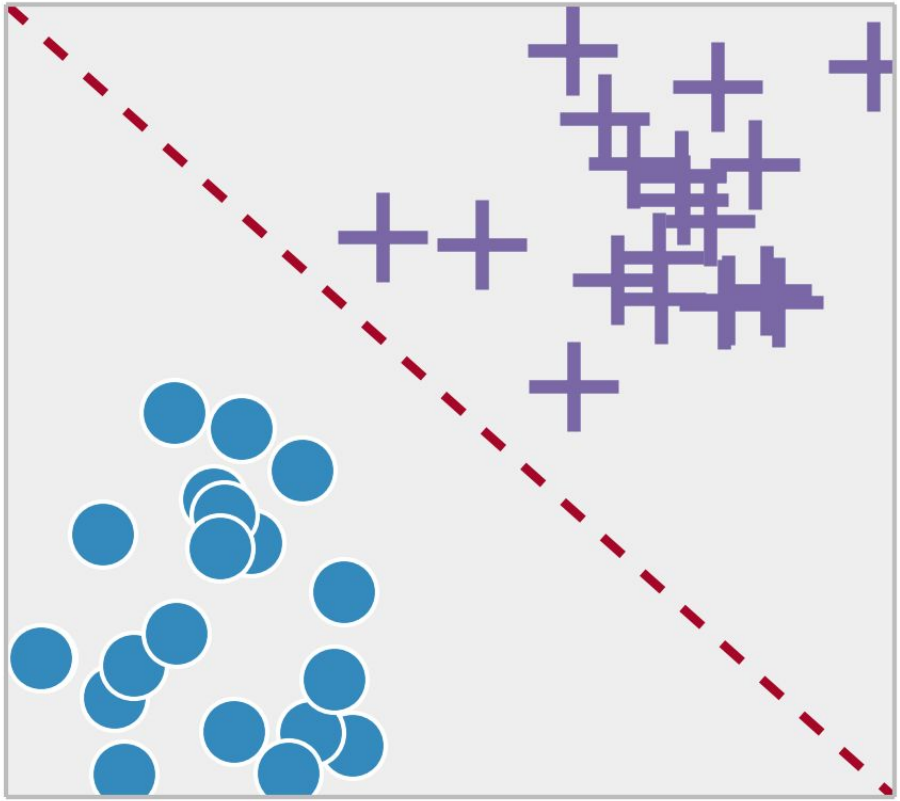
Classification is the task of assigning objects to one of several predefined categories.

- 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

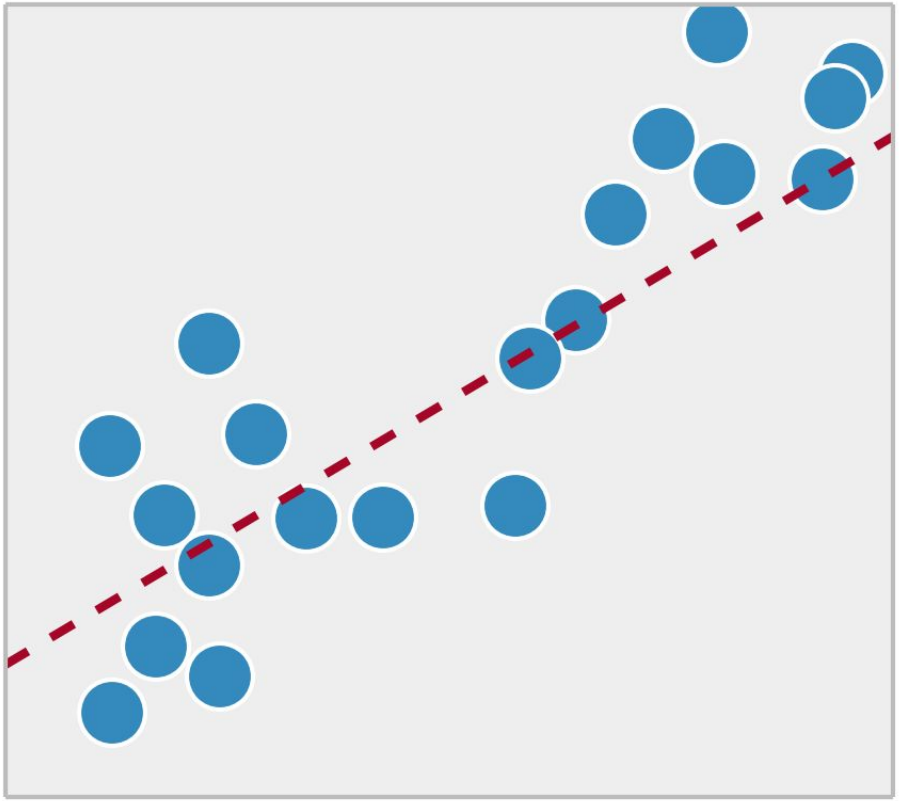
- A classification model can serve as an explanatory tool to distinguish between objects of different classes -- descriptive analytics
- It can also be used to predict the class label of unknown records -- predictive analytics
- Classification techniques are most suited for predicting or describing data 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



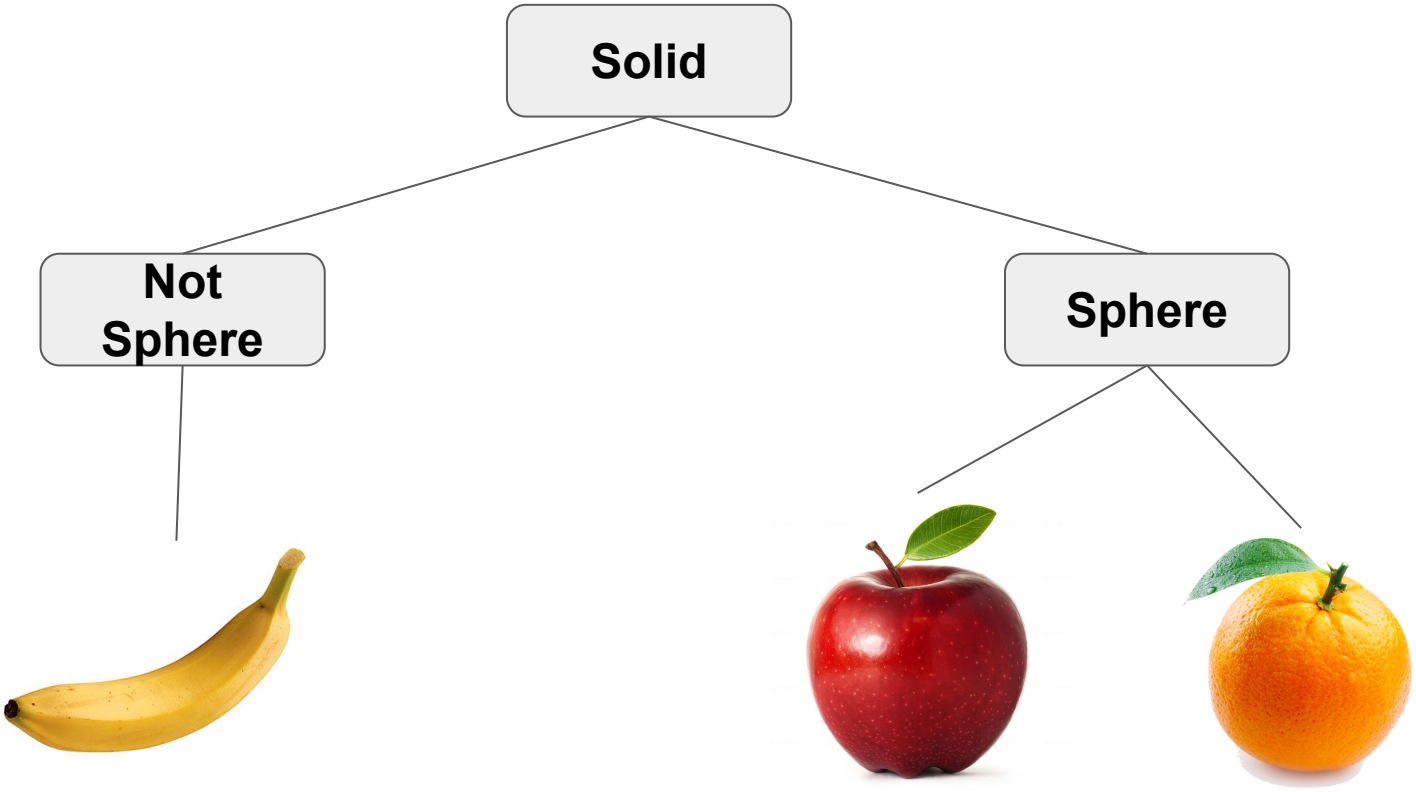
Output: Discrete (labels); Decision boundary
Evaluation: Accuracy;

Regression

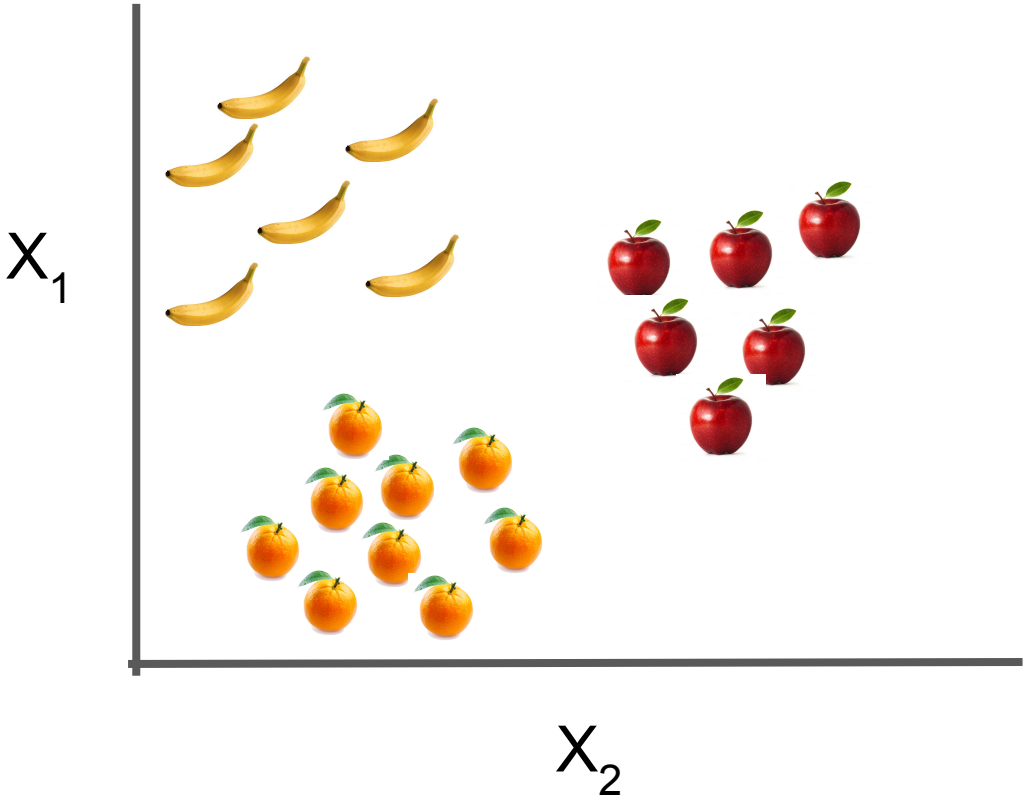


Output: Continuous (number); best fit line
Evaluation: Sum of Errors; R^2

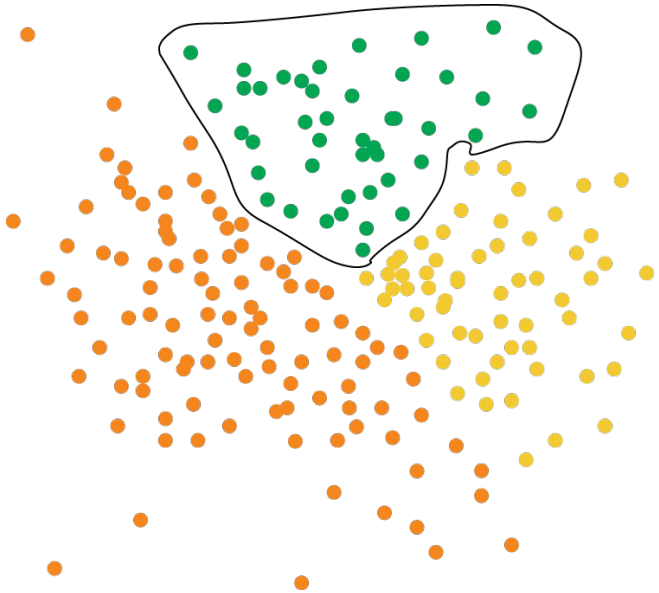
Classification Example

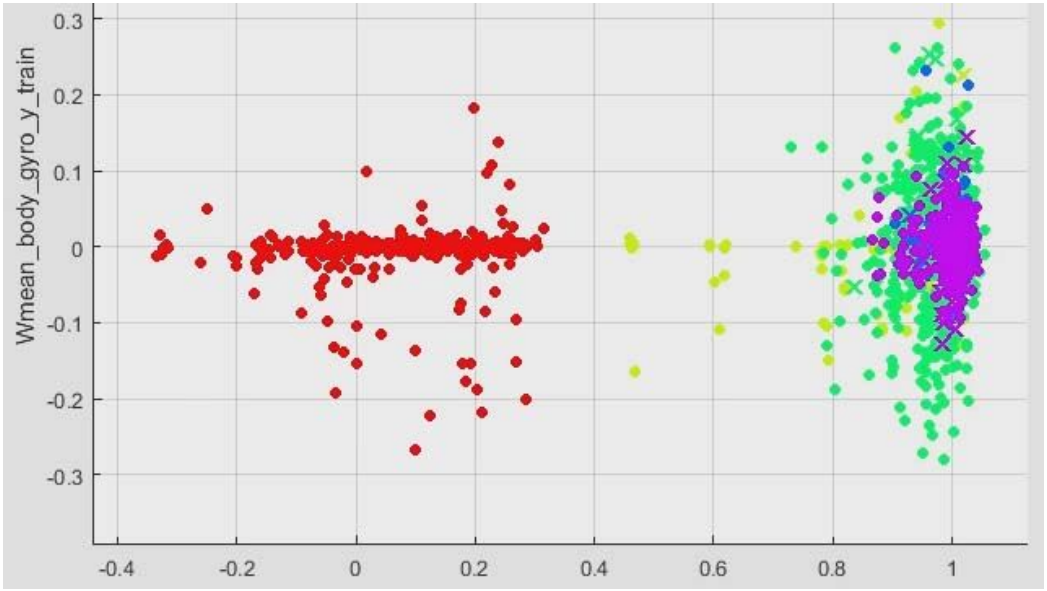


Statistical Classification



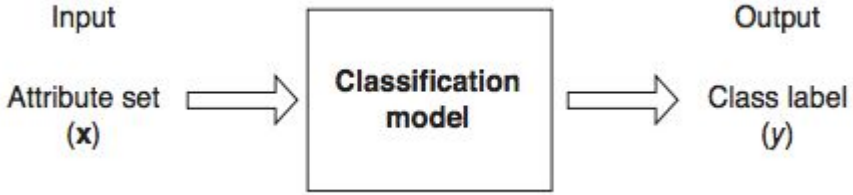
Statistical Classification Examples



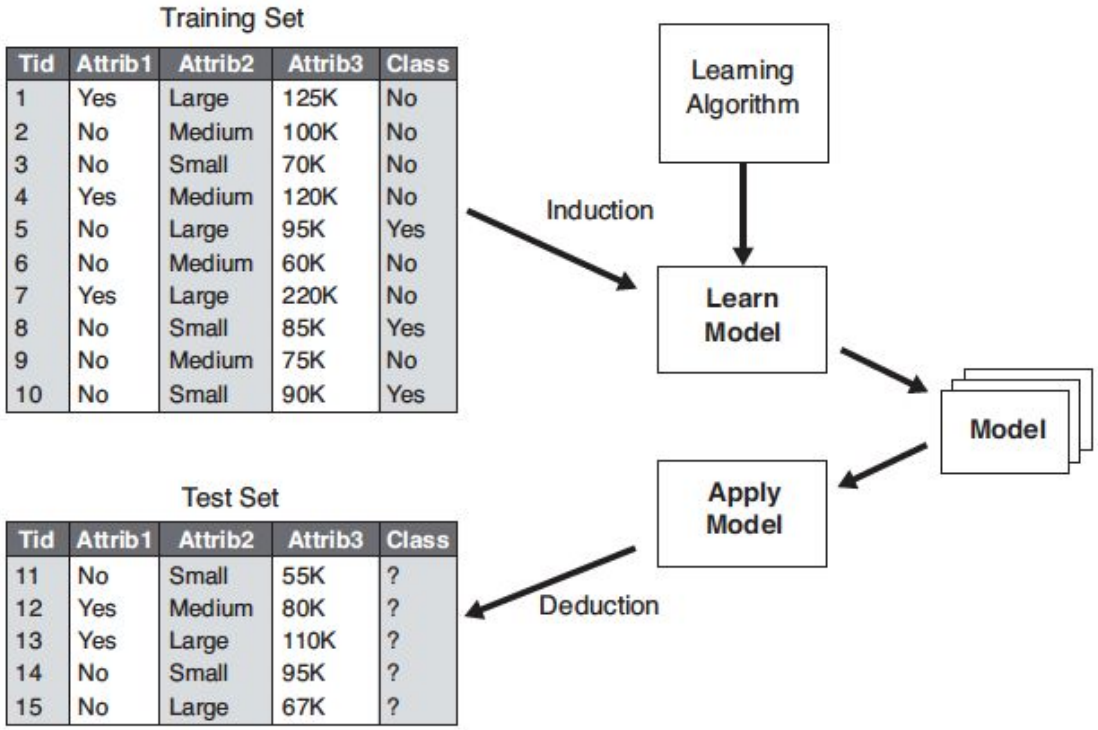


Modeling Process

- 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.

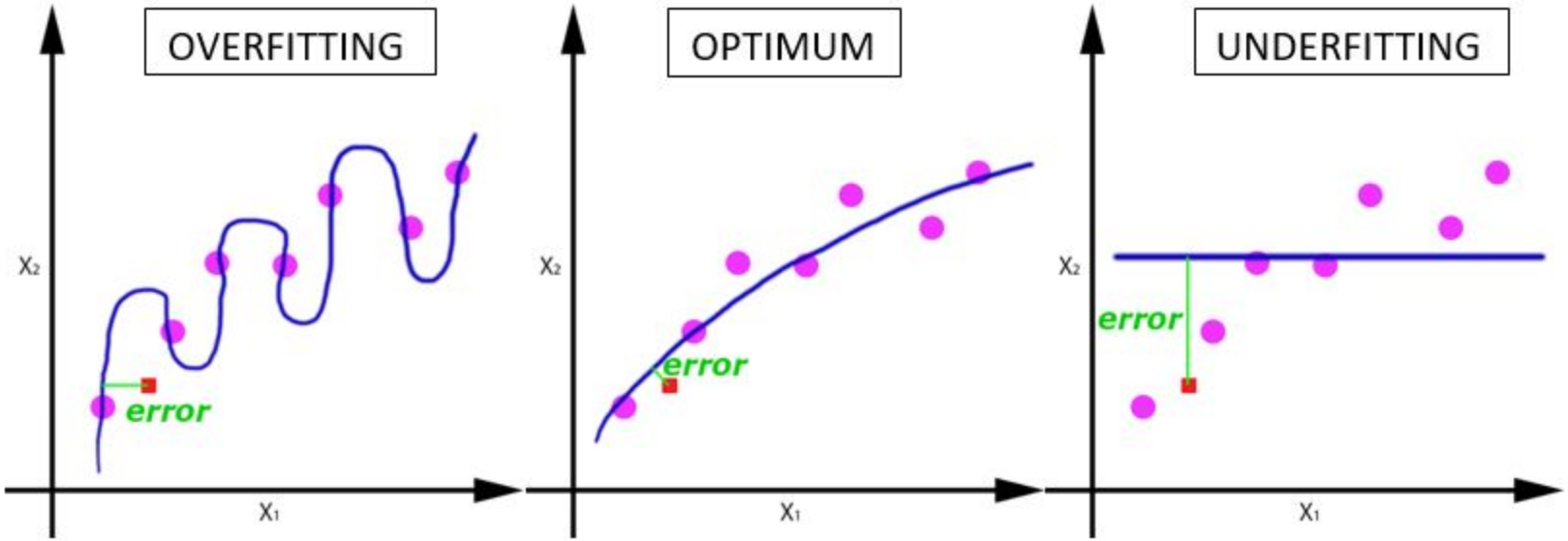


Recap of Modeling Process



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



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:

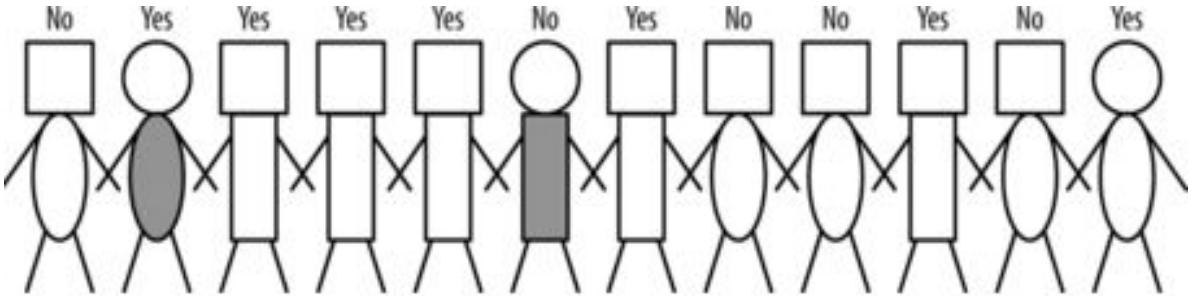
$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

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

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

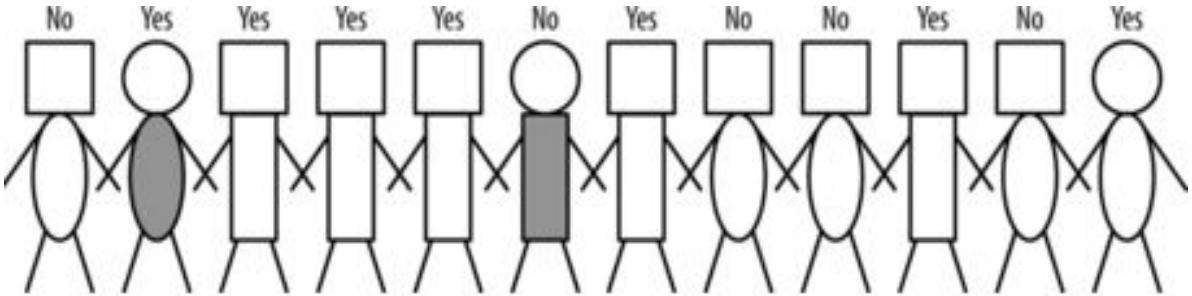
Classification Problem

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



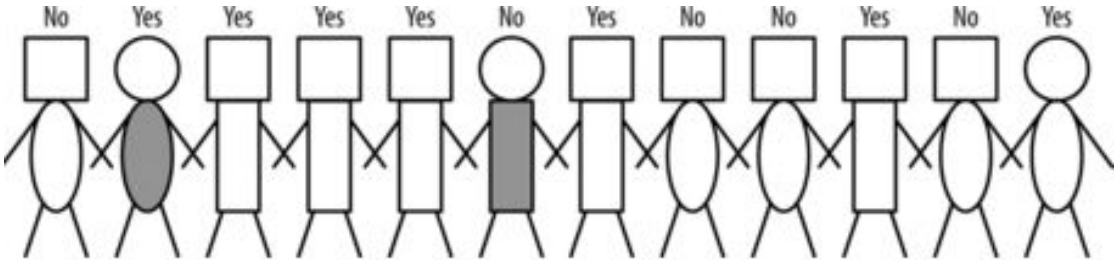
Classification Problem

- 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

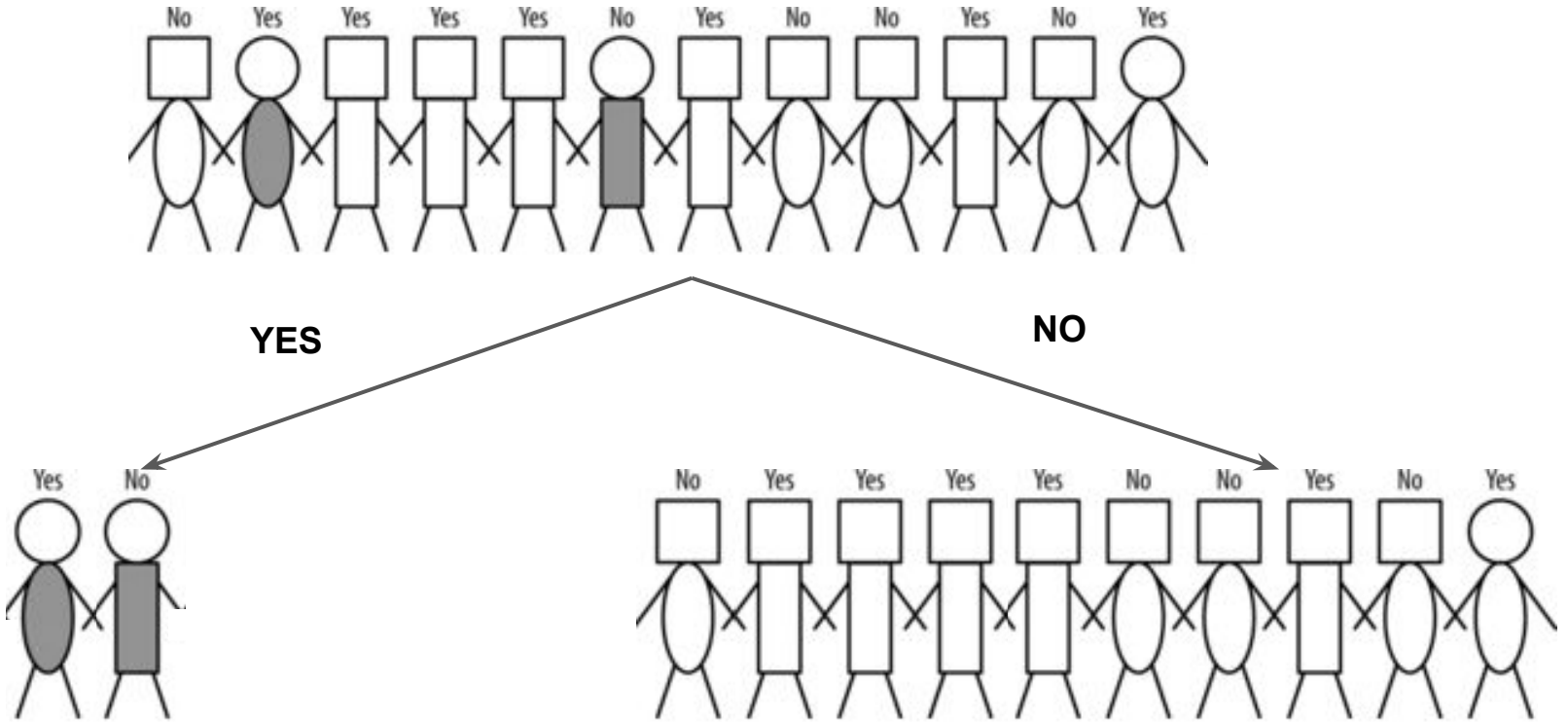


Classification Problem

- 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.

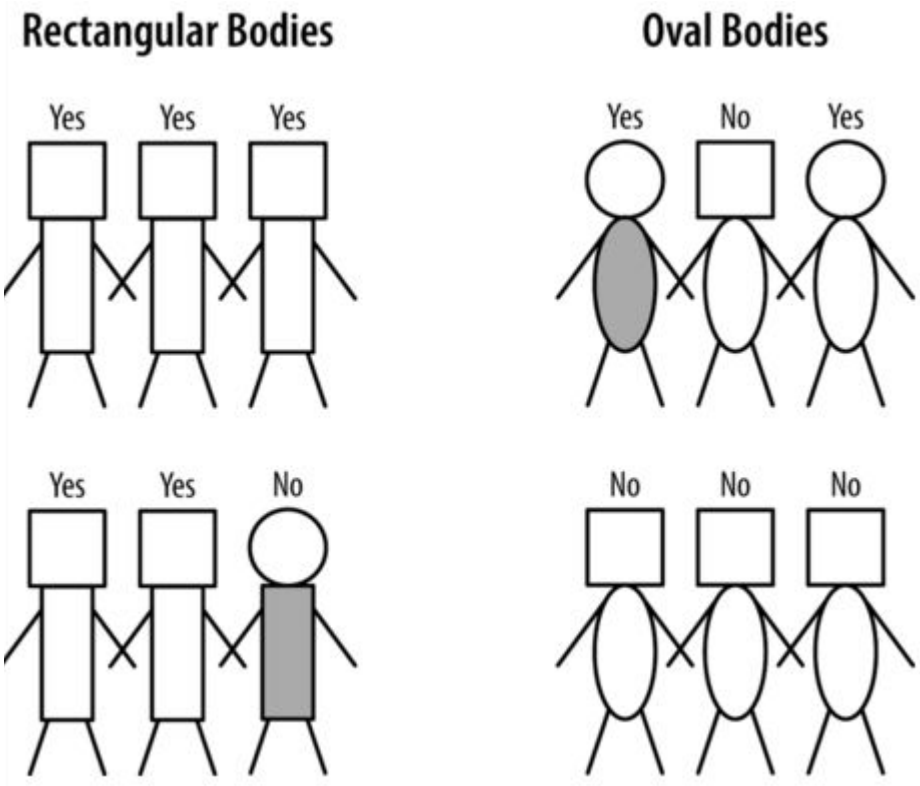


body-color = gray

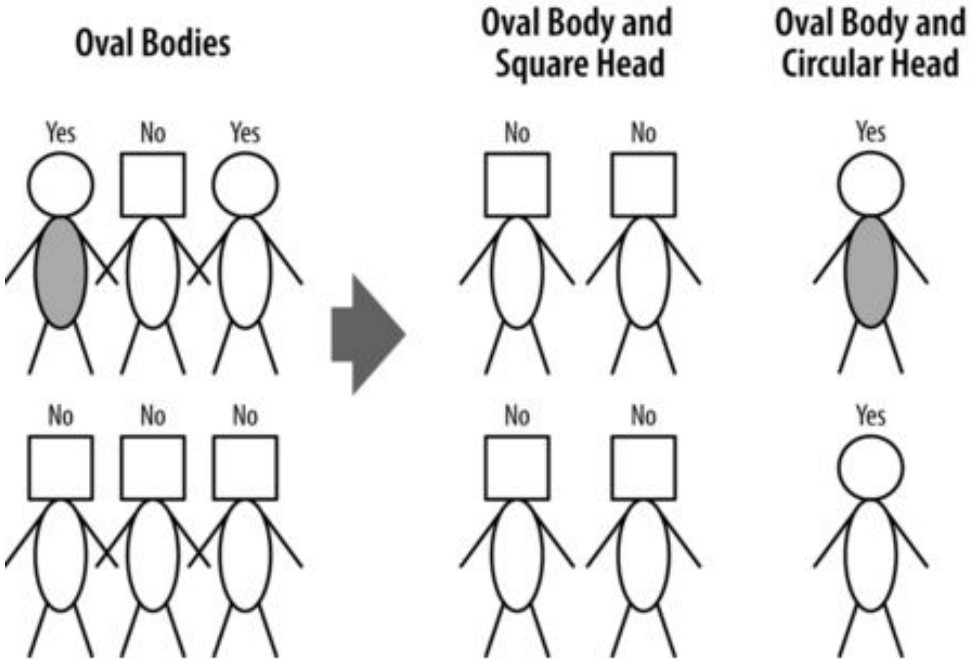


Are these groups pure?

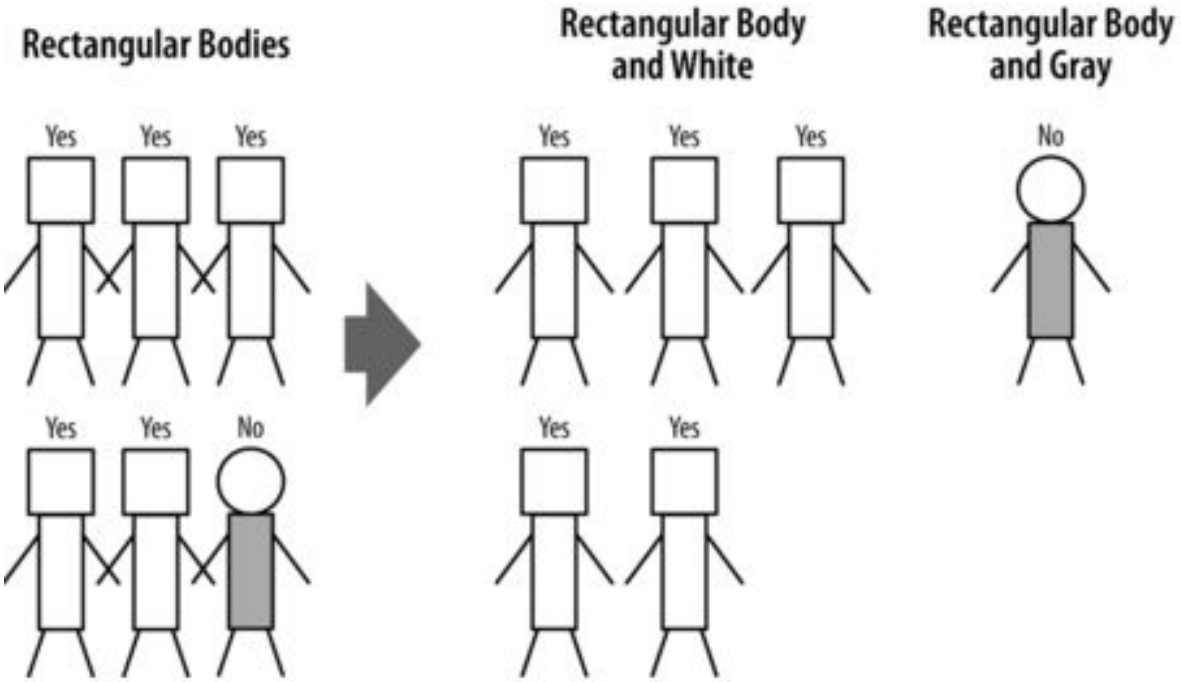
First partitioning: Body shape



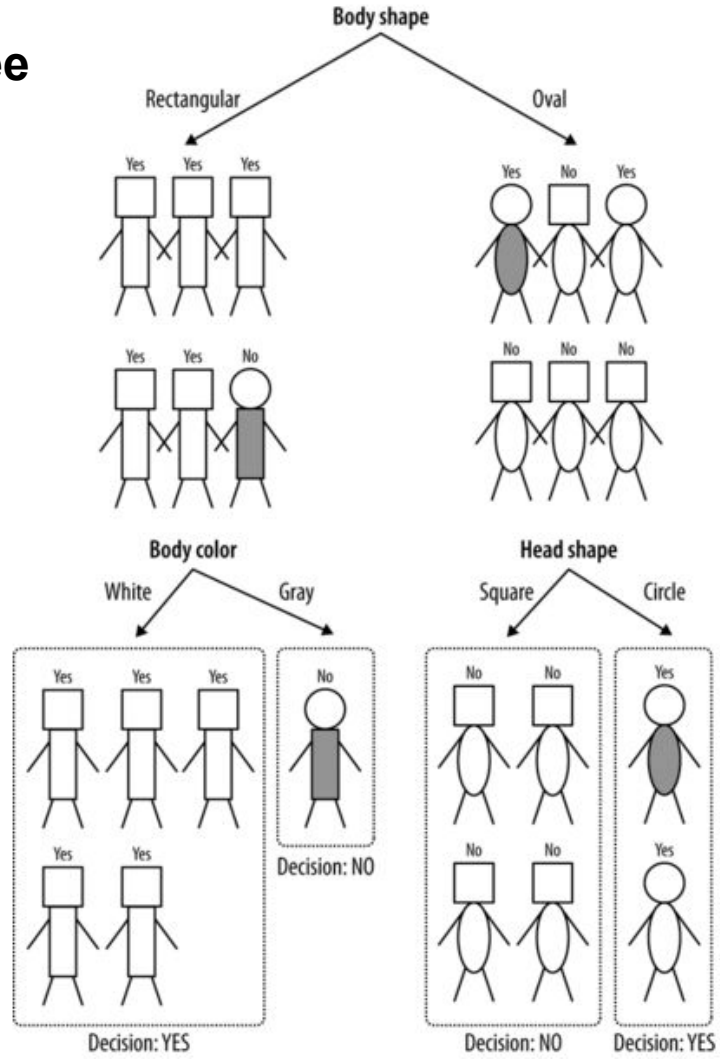
Second partitioning: Oval body people subgrouped by head type



Third partitioning: Rectangular body people subgrouped by body color

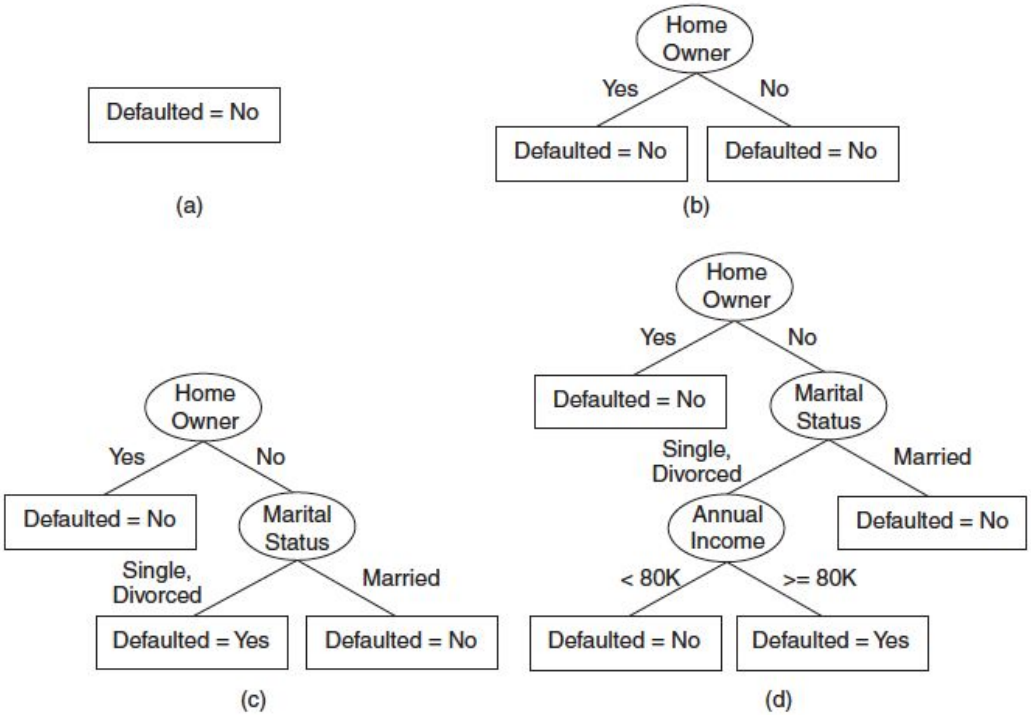


The classification tree resulting from the splits



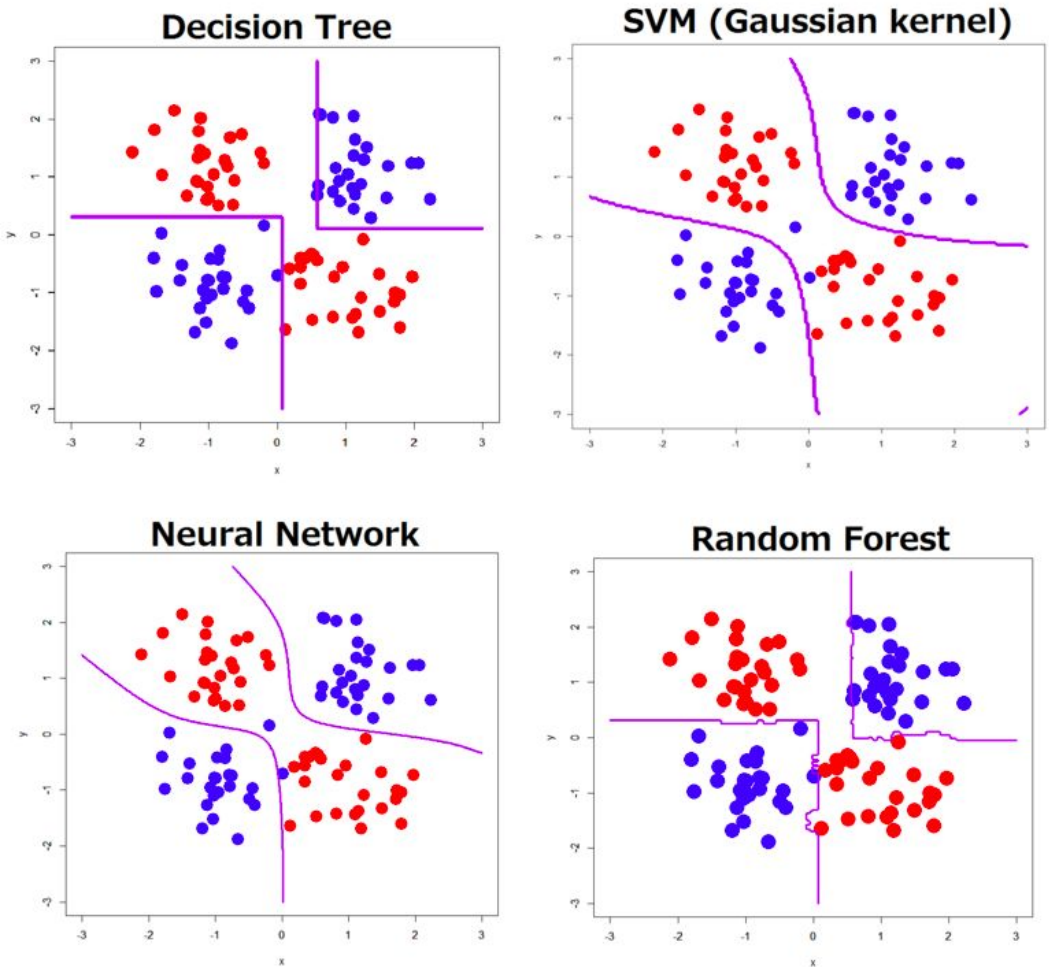
Classification Problem

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



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.


Decision Tree
SVM (Gaussian kernel)
Neural Network
Random Forest

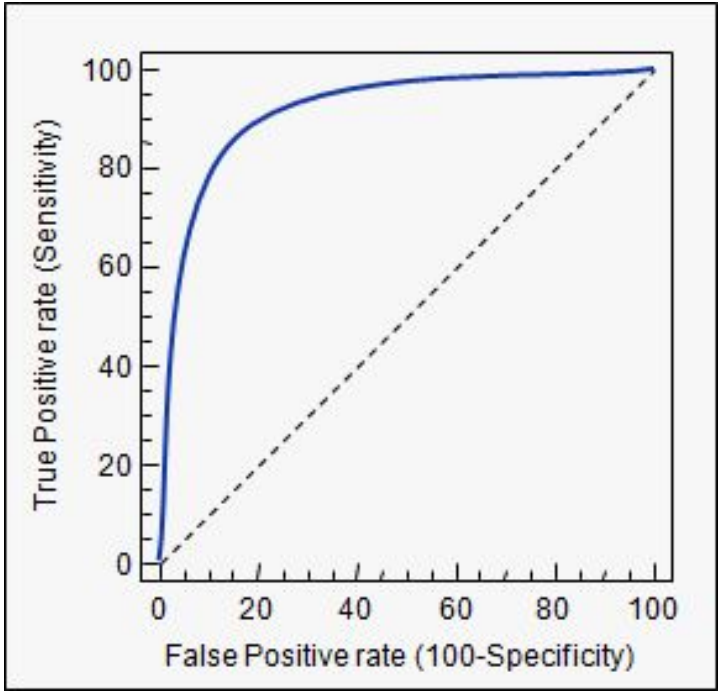
Performance of a logistic regression (Confusion Matrix): It is nothing but a tabular representation of Actual vs Predicted values. This helps us to find the accuracy of the model and avoid overfitting. This is how it looks like:

You can calculate the accuracy of your model with:

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

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.



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)

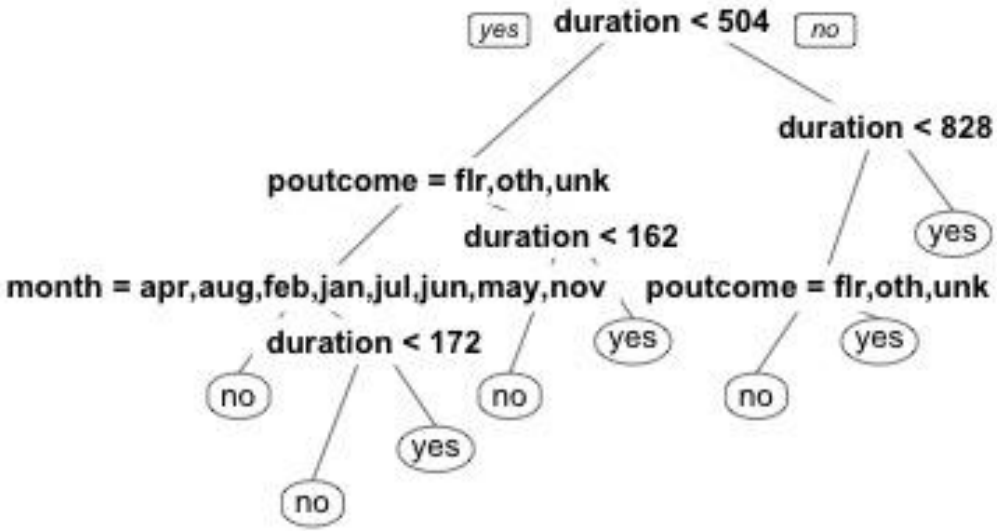
15 - previous: number of contacts performed before this campaign and for this client (numeric)

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?



Homework 2 - Part 1: Linear Modeling

Campbell's and Progresso are two popular brands of canned soups that are available in the U.S. You are provided data for sales of Progresso soup in the U.S. The data are derived from approximately 2000 supermarkets across the country and span 6 years (2001-06) and includes Income binary data (“Low_income” vs “High_income”), competitors’ price (Campbell “Price.Campbell” and private lables “Price.PL”)

1. **Business Question - Is soup consumption increasing during the Winter months?**
    a. Create a dummy variable for “Winter” months defined as Oct, Nov, Dec, Jan & Feb. Use the “Month” variable to create this.
    b. Compute the “Market Share” for Progresso (as percentage of total sales) in the Winter vs. non-Winter months using the variable created in (a).
2. **Business Question - Can we predict sales so we can work with our supply chain vendors and inventory management?**
    a. Develop a linear regression model to predict Progresso sales. Explain the results of the regression model (model strength, variable importance, relationship between the predictor and dependent variables)
    b. Predict Progresso Sales for stores listed in the “Progresso_Soup_Prediction.csv ”

Modeling data: https://raw.githubusercontent.com/jcbonilla/BusinessAnalytics/master/BAData/Progresso_Soup.csv

Prediction data: https://raw.githubusercontent.com/jcbonilla/BusinessAnalytics/master/BAData/Progresso_Soup_Prediction.csv

Homework - Part 2: Classification Interpretation

Business Question - What is the strategy this bank should follow to increase deposits subscription?

The data is related with 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, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

- Run the Classification_Simple_Example code
- Interpret the classification three
- Write a recommendation statement as per the business question above

Data catalog: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Code: https://github.com/jcbonilla/BusinessAnalytics/blob/master/04%20-%20Predictive%20Models/Classification_Simple_Example.R

Homework 2 - Part 3: Proof Something

Journalist and certainly politicians continue to point at the severity of terrorism. Using the dataset from data.world on U.S. terrorism cases since 2001 validate or disproof the statement that terrorism is an **increasing** thread in the USA
Source: https://data.world/carlvlewis/terrorism-cases-2001-2016

To submit homework 2
Submit **PDF** file with your visualizations, analysis, and response. (you do not have to submit R code)