Data Mining Overview and Predictive Modeling

DSBA/MBAD 6211 Advanced Business Analytics

Spring 2024

Agenda

Data Mining Process Overview

Predictive Modeling

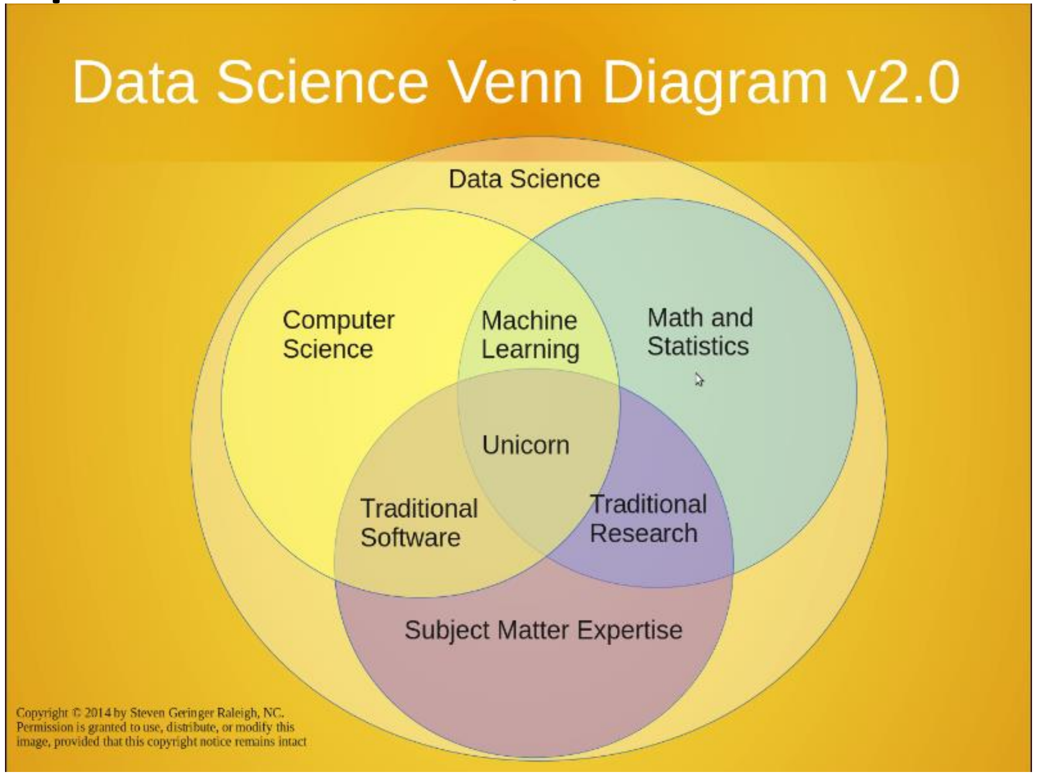
Model Comparison and Evaluation

What Is Data Mining?

- ❖ Data mining is the discovery of ***models*** from data.
  - Data contains value and knowledge.
  - Various types of models
    - Statistical modeling
    - Machine learning
  - Beyond counts, descriptive techniques, and reporting

Origins of Data Mining

- Data mining is a discipline lying at the interface of mathematics/statistics, computer science, and domain expertise
- Challenges
  - Scalability
  - Dimensionality
  - Heterogeneity
  - Ownership and distribution



Adopt from Grossman, R.L., Kamath, C., Kegelmeyer, P., Kumar, V., Namburu, R. Data Mining for Scientific and Engineering Applications

What Is Data Mining?

- **Supervised** learning (with *responses* or *dependent variables*)
  - regression
  - decision trees
  - neural network
- **Unsupervised** learning (without *responses* or *dependent variables)*
  - cluster analysis
  - association rules
- **Reinforcement** learning
  - Figure out What to do to maximize a Reward by itself
  - Does not strictly rely on set of labeled dependent variables
  - Examples
    - Markov decision process
    - Q learning

Data Mining Tasks

- Prediction
  - Predict dependent variables based on independent variables
  - Little focus on mechanism
    - Example: which customers are most likely to purchase?
- Inference
  - Understand the relationship between dependent variables and independent variables
    - Are they associated?
    - What is the relationship?
    - Example: what kind of products customs like to buy together?

Pattern Detection

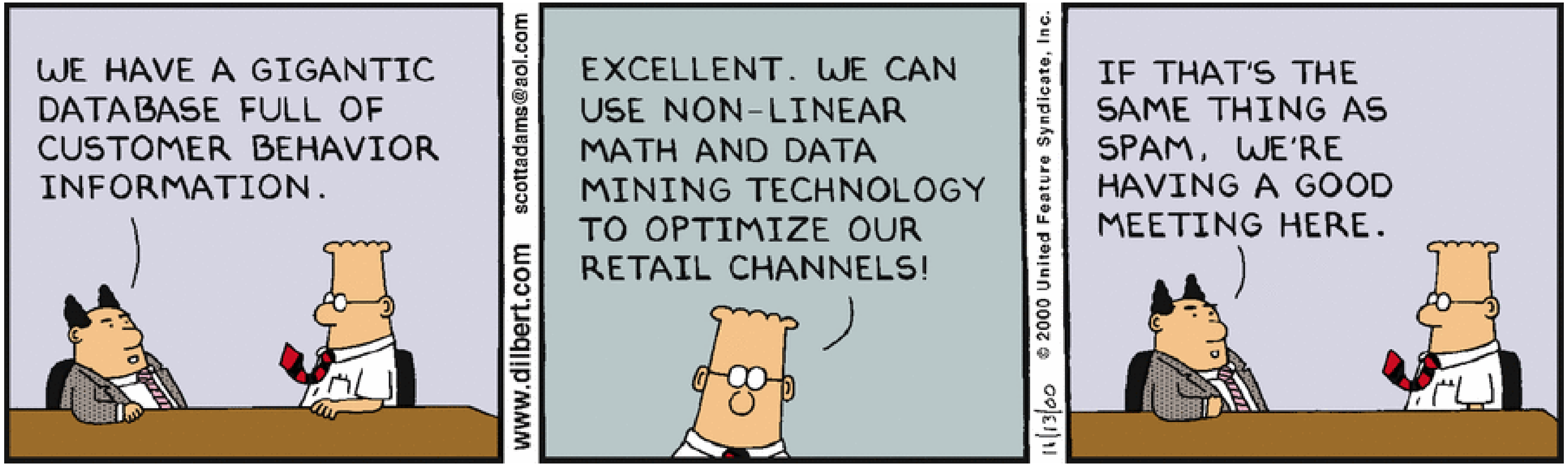
- Patterns may or may not represent any underlying rule.
- Some patterns reflect some underlying reality.
  - The party that holds the White House tends to lose seats in Congress during off-year elections.
- Others do not.
  - When the American League wins the World Series in Major League Baseball, Republicans take the White House.
- Sometimes, it is difficult to tell without analysis.
  - In U.S. presidential contests, the taller candidate usually wins.

What Types of Patterns Are Valuable?

- Evidence
  - Use a statistical criterion to measure the significance of the finding
- Redundancy
  - Similarity to other findings
- Usefulness
  - Meet the goal of the user
- Simplicity
- Generality

"*Data analysis is as much an art as a science*."

What Types of Patterns Are Valuable?



What Types of Patterns Are Valuable?

- Bonferroni's Principle
  - Roughly speaking, a data-mining risk is that you "discover" patterns that are meaningless
  - If you look in more places for interesting patterns than your amount of data will support, you are bound to find crap
  - What your model suggests >>what you should expect

"*Torture the data, and it will confess to anything*." Ronald Coase, Economics Nobel Prize Laureate

Bonferroni’s Principle: An Example

- Terrorist detection
  - Potential rule: two unrelated people who at least twice have stayed at the same hotel on the same day
  - Expected number of “suspicious” pair of people
    - 10^9 people being tracked
    - 1,000 days
    - Each person stays in a hotel 1% of time (1 day out of 100)
    - Hotels hold 100 people (so 10^5 hotels)
    - If everyone behaves randomly
    - Finding: 250,000 “suspicious” pairs

Bonferroni’s Principle: An Example

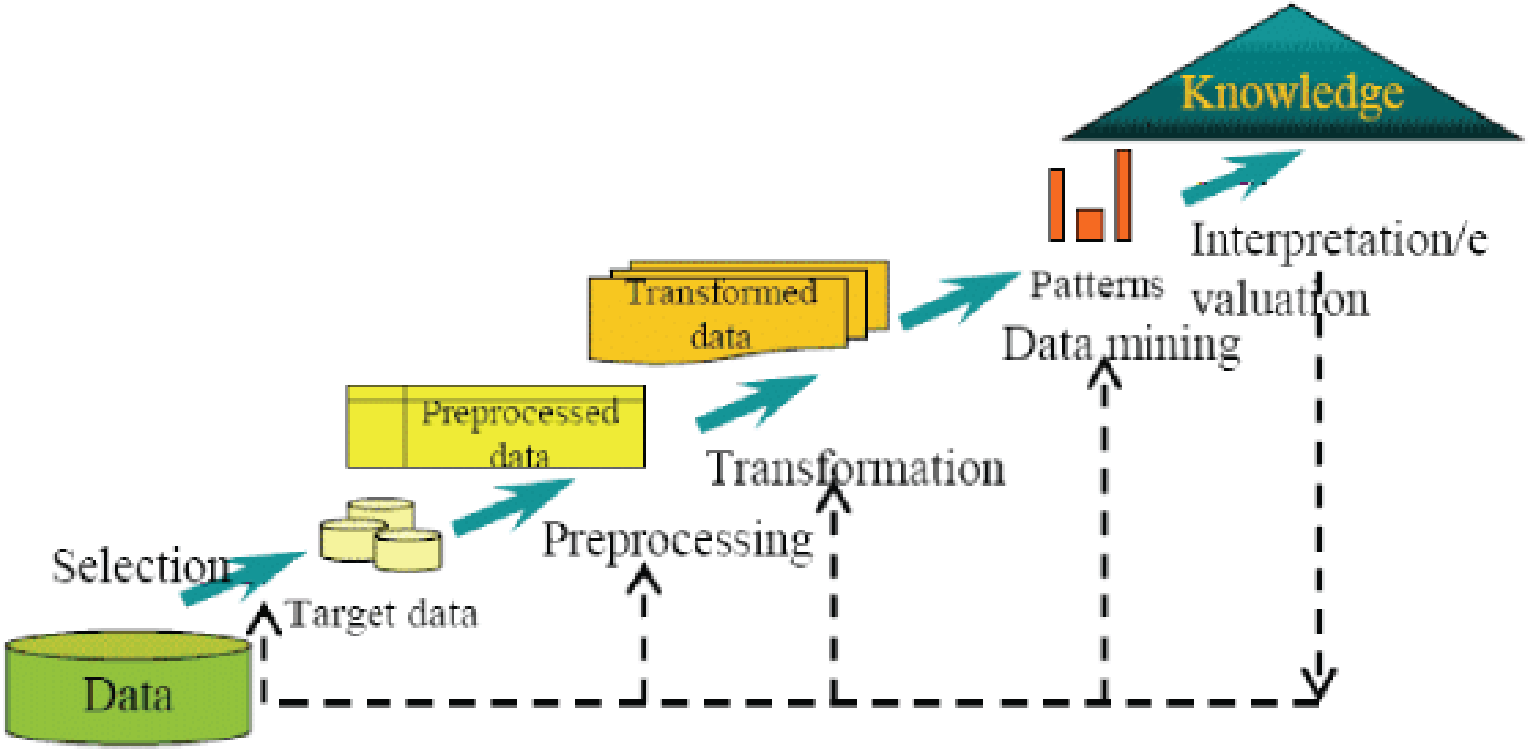
- Your model suggests: 250,000 “suspicious” pairs
- Suppose the reality is that only 10 pairs of evil-doers stayed at the same hotel twice in the past 1000 days
- What your model suggests >>what you should expect
  - Useless model
  - Very expensive to find 10 real cases through 250,000 candidates

Bonferroni’s Principle

- When looking for a property (*e.g., “two people stayed at the same hotel twice”*), make sure that the property does not allow so many possibilities that random data will surely produce facts of interest.

Data Mining Process

- How do we get from Data to “knowledge ”



Adopt from Grossman, R.L., Kamath, C., Kegelmeyer, P., Kumar, V., Namburu, R. Data Mining for Scientific and Engineering Applications

Data Mining Process

1. Define Purpose
2. Obtain Data
3. Explore & Clean Data
4. Determine Date Mining Task (classification, clustering, etc.)
5. Partition the Data (for supervised tasks)
6. Choose Data Mining Methods (regression, neural nets, etc.)
7. Apply Method
8. Evaluation Performance
9. Model Deployment

Case Background

- A financial services company offers a home equity line of credit to its clients. The company has extended several thousand lines of credit in the past, and many of these accepted applicants (approximately 20%) have defaulted on their loans. By using geographic, demographic, and financial variables, the company wants to build a model to predict whether an applicant will default.

1. Define Purpose

- What is the purpose of the project?

The goal of the project is "to predict whether an applicant will default"

Questions to think:

- How will the stakeholder use the results?
- Who will be affected by the results?
- Will the analysis be a one-shot effort or an ongoing procedure?

2. Obtaining Data

- May include sampling from one or more databases
- Sampling is the main technique employed for data selection
  - Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming

Variable Description

Name	Model Role	Measurement Level	Description
BAD	Target	Binary	1: default
CLAGE	Input	Interval	Age of oldest credit line in months
CLNO	Input	Interval	Number of credit lines
DEBTING	Input	Interval	Debt-to-income ratio
DELINQ	Input	Interval	Number of delinquent credit lines
DEROG	Input	Interval	Number of major derogatory reports
JOB	Input	Nominal	Occupational categories
LOAN	Input	Interval	Amount of the loan request
MORTDUE	Input	Interval	Amount due on existing mortgage
NINQ	Input	Interval	Number of recent credit inquiries
REASON	Input	Binary	DebtCon=debt consolidation. HomeImp=home improvement
VALUE	Input	Interval	Value of current property
YOJ	Input	Interval	Years at present job

3. Explore, Clean, and Preprocess

- Data problems
  - Missing values
  - Outliers
  - Errors

Exploring, understanding and visualizing data are perhaps the most important steps in the data mining process

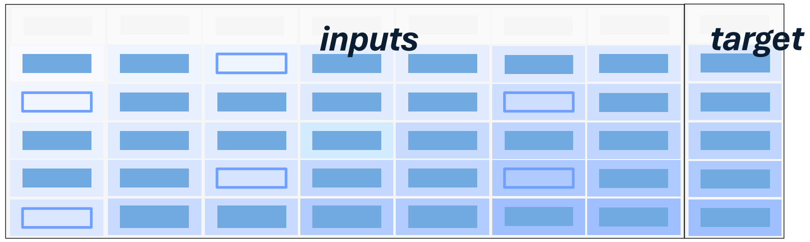
- Understand your data
  - Data type for each variable – correct?
  - Data range – reasonable?
  - Visualization – any model-free patterns?

Manage Missing Values

- Causes
  - Errors
  - Non-applicable measurement
  - Non-disclosed measurement
- Regression & Neural Network models
  - Ignore incomplete observations
- Decision tree
  - Automatically handle missing values with a variety of algorithms

Manage Missing Values

- Problems?
  - A smattering of missing values can cause an enormous loss of data in high dimensions.
  - Assuming that each of the *k* input variables is missing at random with probability α. In this situation, the expected proportion of complete cases is as follows: $(1-\alpha)^k$
    - A 1% probability of missing (α=.01) for 100 inputs leaves only 37% of the data for analysis, 200 leaves 13%, and 400 leaves 2%.



Manage Missing Values

- Synthetic distribution methods
  - Use a one-size-fits-all approach to handle missing values. Any case with a missing impute measurement has the missing value replaced with a fixed number.
- Estimation methods
  - Provide customized imputations for each case with missing values. This is done by viewing the missing value problem as a prediction problem. That is, you can train a model to predict an input’s value from other inputs.

What Are Outliers

❖ Definition: in statistics, an outlier is an observation that is numerically distant from the rest of the data.

❖ A review of outliers is needed to determine if the data point is a result of an error or if it is a special case?

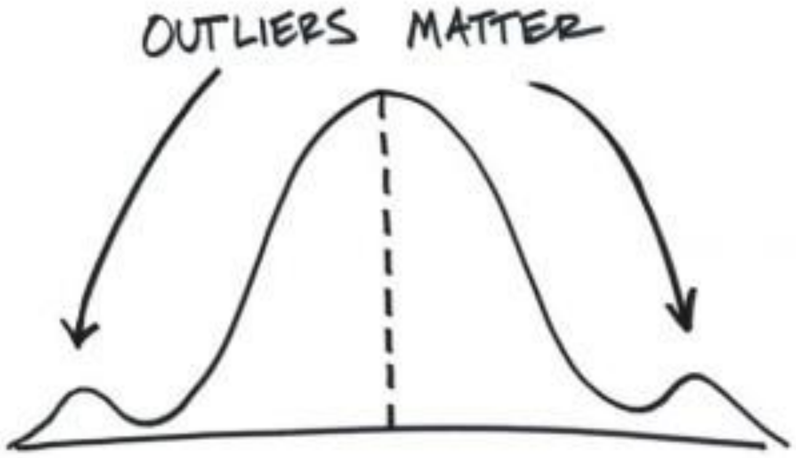
- Bad data : age - 150
- Data variation

❖ Inspection:

- Summary statistics
- Visualization

❖ Ways to handle

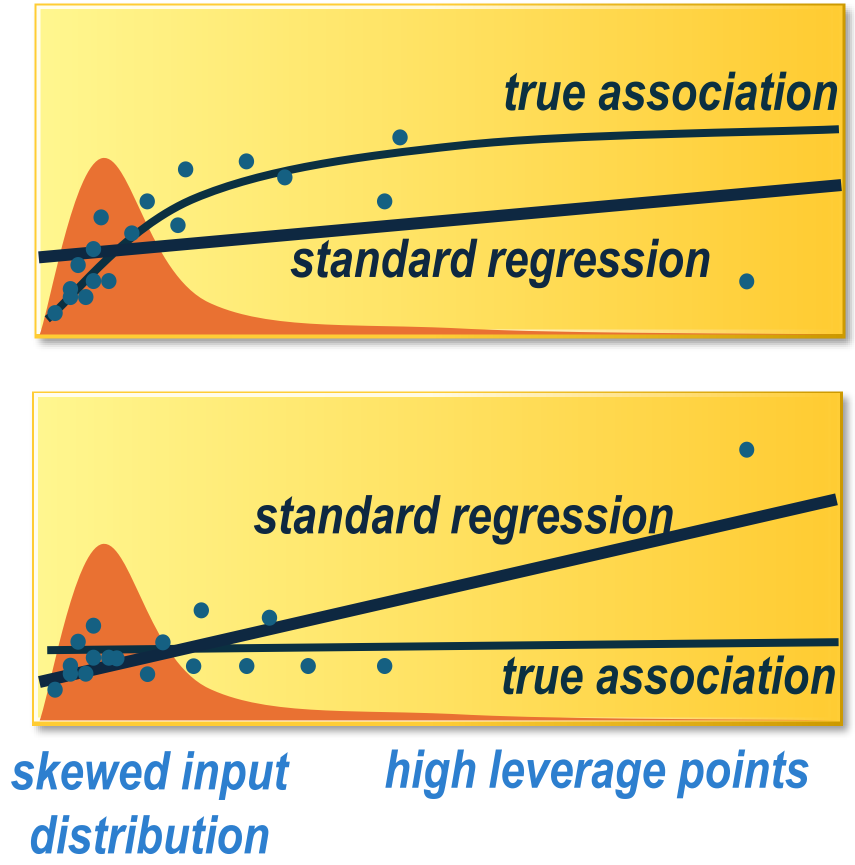
- Error – manual correction
- Small amount-treat as missing values



https://www.alexandergroup.com/insights/the-other-side-of-outliers/

Variable Transformation

- Independent variables with highly skewed distributions
  - A small percentage of the points may have a great deal of influence.
  - Transforming or regularizing variables



Categorical Variables and Dummy Coding

- Categorical variables
  - Variables that can take on one of a limited, number of possible values
  - Examples:
    - Student: Yes/No (binary/dichotomous)
    - Size: Small, Medium, Large, & Extra Large
- Dummy coding
  - Uses only ones and zeros to convey all of the necessary information on categories
  - k categories with k-1 coded variables

Categorical Variables and Dummy Coding

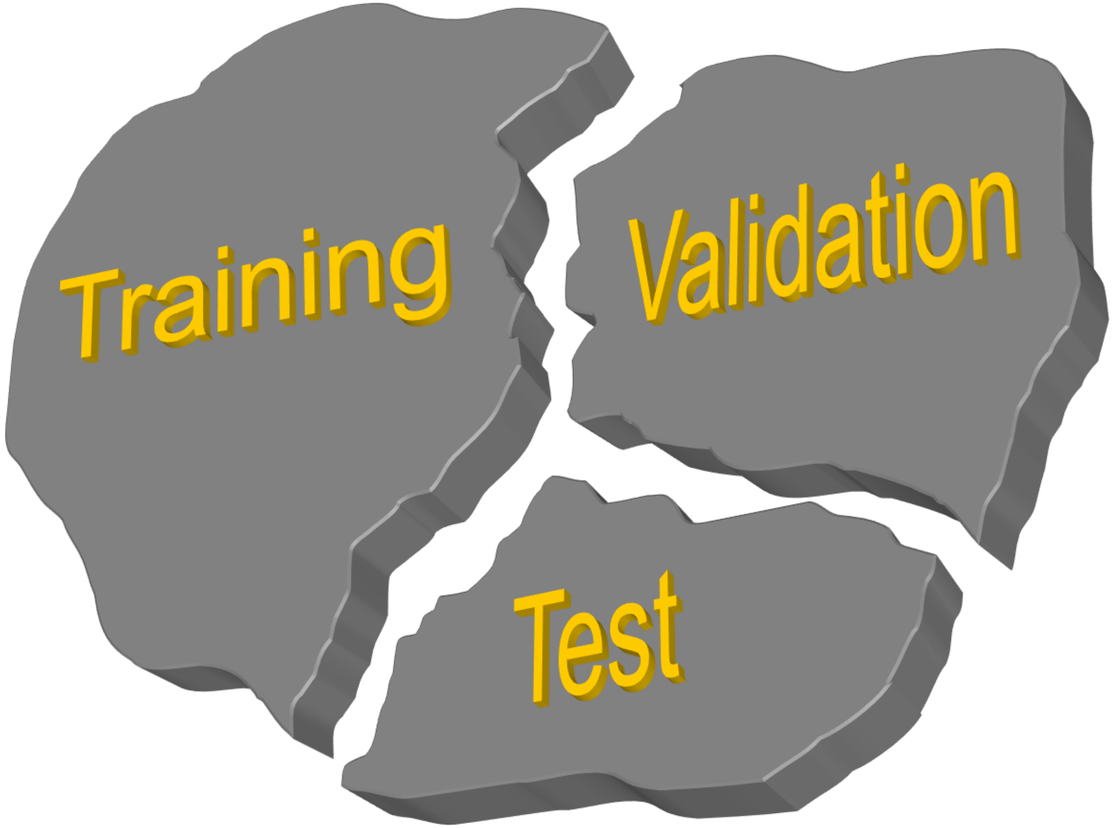
Redundant

Level	D_A	D_B	D_C	D_D	D_E	D_F	D_G	D_H	D_I
A	1	0	0	0	0	0	0	0	0
B	0	1	0	0	0	0	0	0	0
C	0	0	1	0	0	0	0	0	0
D	0	0	0	1	0	0	0	0	0
E	0	0	0	0	1	0	0	0	0
F	0	0	0	0	0	1	0	0	0
G	0	0	0	0	0	0	1	0	0
H	0	0	0	0	0	0	0	1	0
I	0	0	0	0	0	0	0	0	1

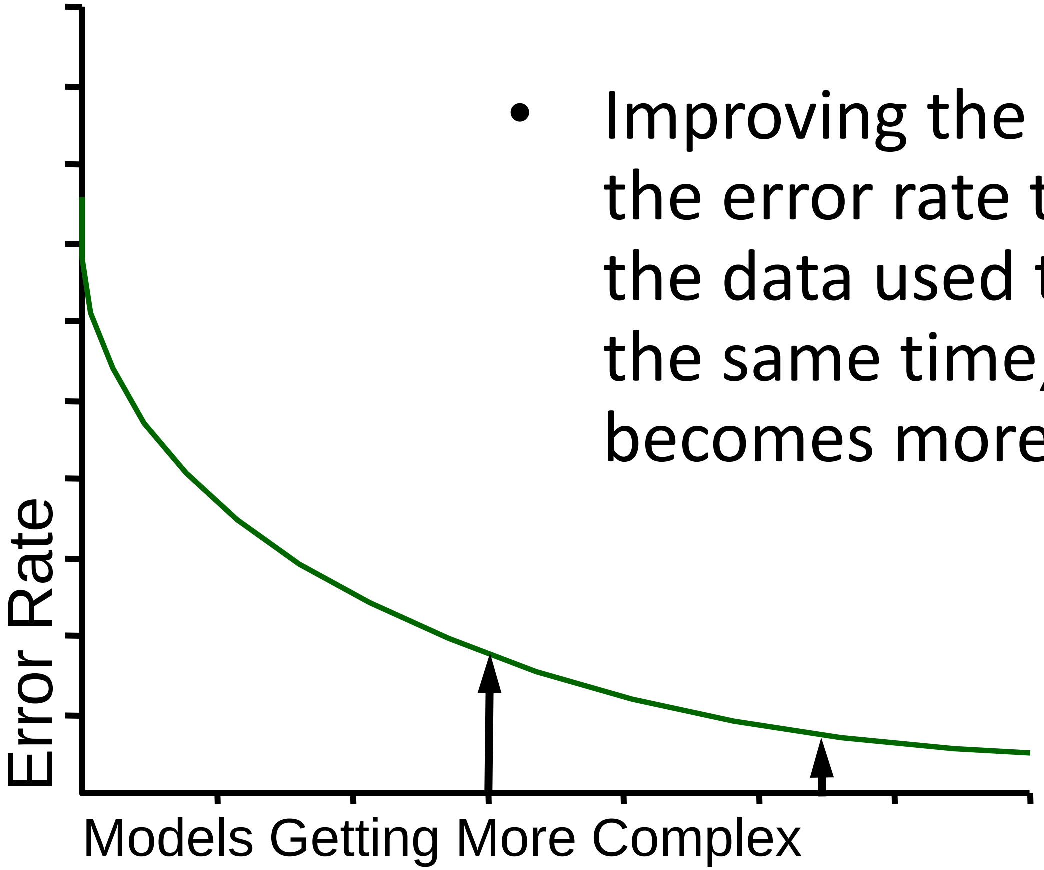
Variable Selection

- Some variables will not enter the model
  - Can not use
    - Legal consideration
    - Privacy/ethical issues
  - Should not use
    - Significant quality issues
    - Constant
    - Conceptually nonrelated
    - Redundant information

5.Data Partition (*for supervised task*)

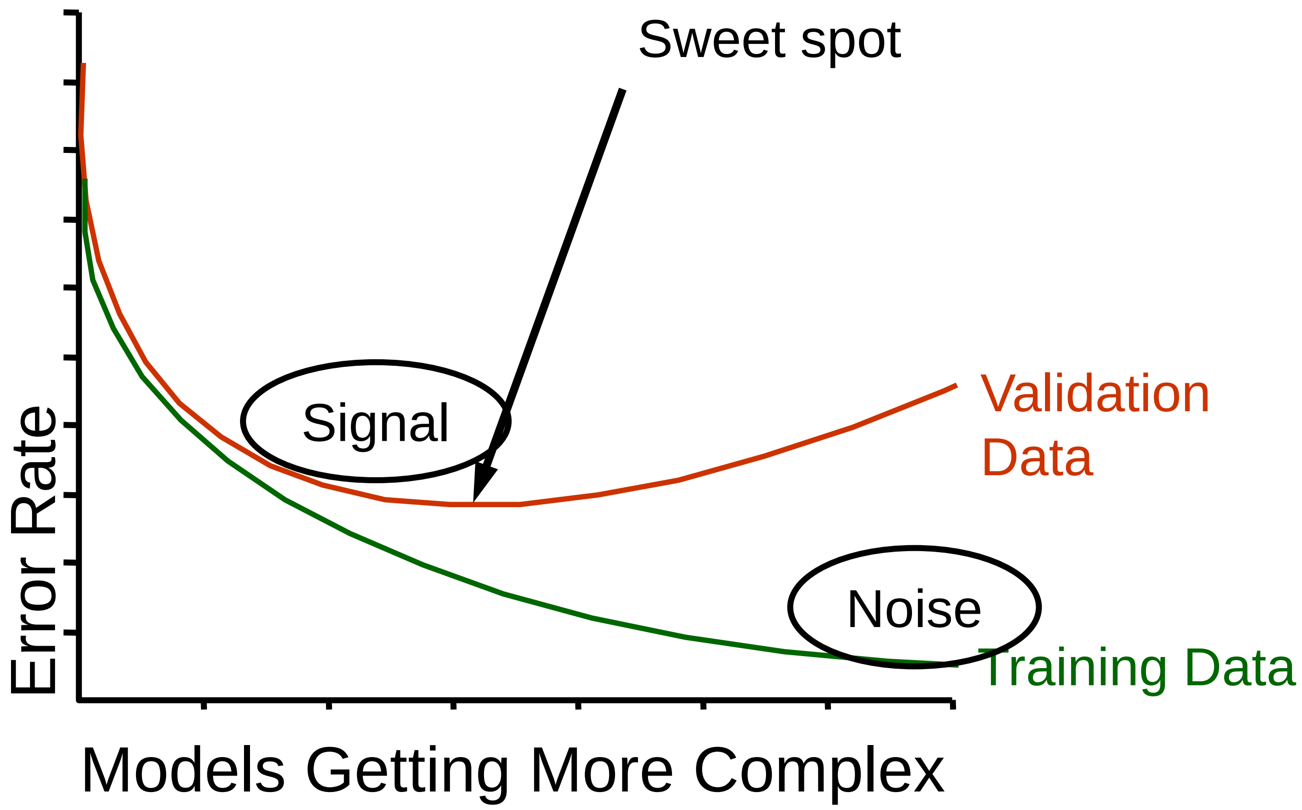


Data Partition

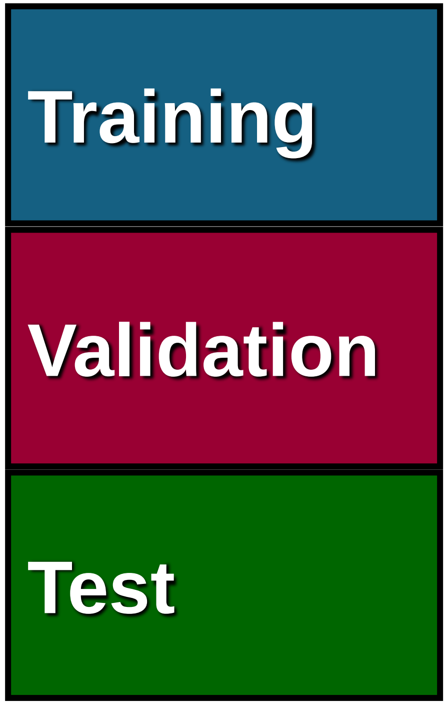


- Improving the model causes the error rate to decline on the data used to build it. At the same time, the model becomes more complex.

Validation Data Prevents Overfitting



Data Partition



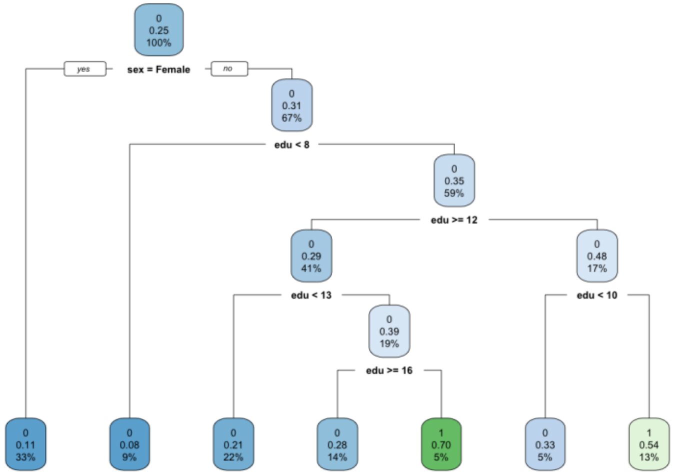
- Mutually exclusive data sets
- Use the ***training set*** to find patterns and create an initial set of candidate models.
- Use the ***validation set*** to select the best model from the candidate set of models.
- Use the ***test set*** to measure performance of the selected model on unseen data.
  - Holdout sample

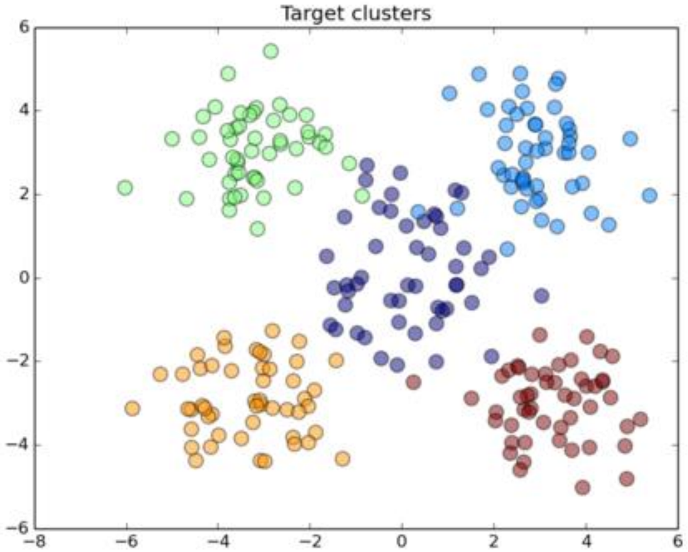
Data Partition

- Trade-off
  - More data is devoted to training results in more stable predictive models, but less stable model assessments (and vice versa).
  - Also, the test partition is used only for calculating fit statistics after the modeling and model selection is complete.

6. Choose Data Mining Methods

- ❖ Model selection referrers both to selecting the “right” model using a method and to selecting between methods.
- ❖ There is no universal best method!





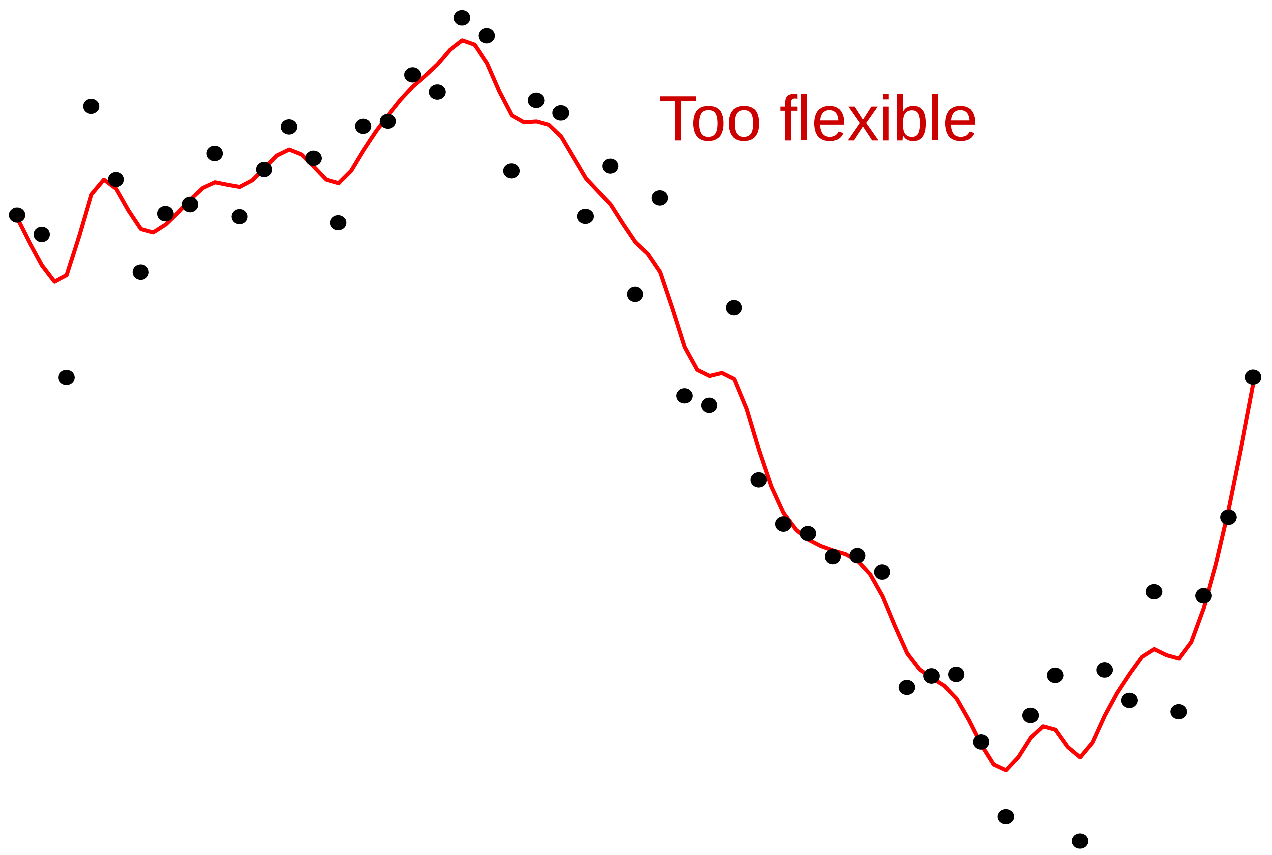
Model Selection

❖ Model selection is dependent on both the data at hand and the data mining goal

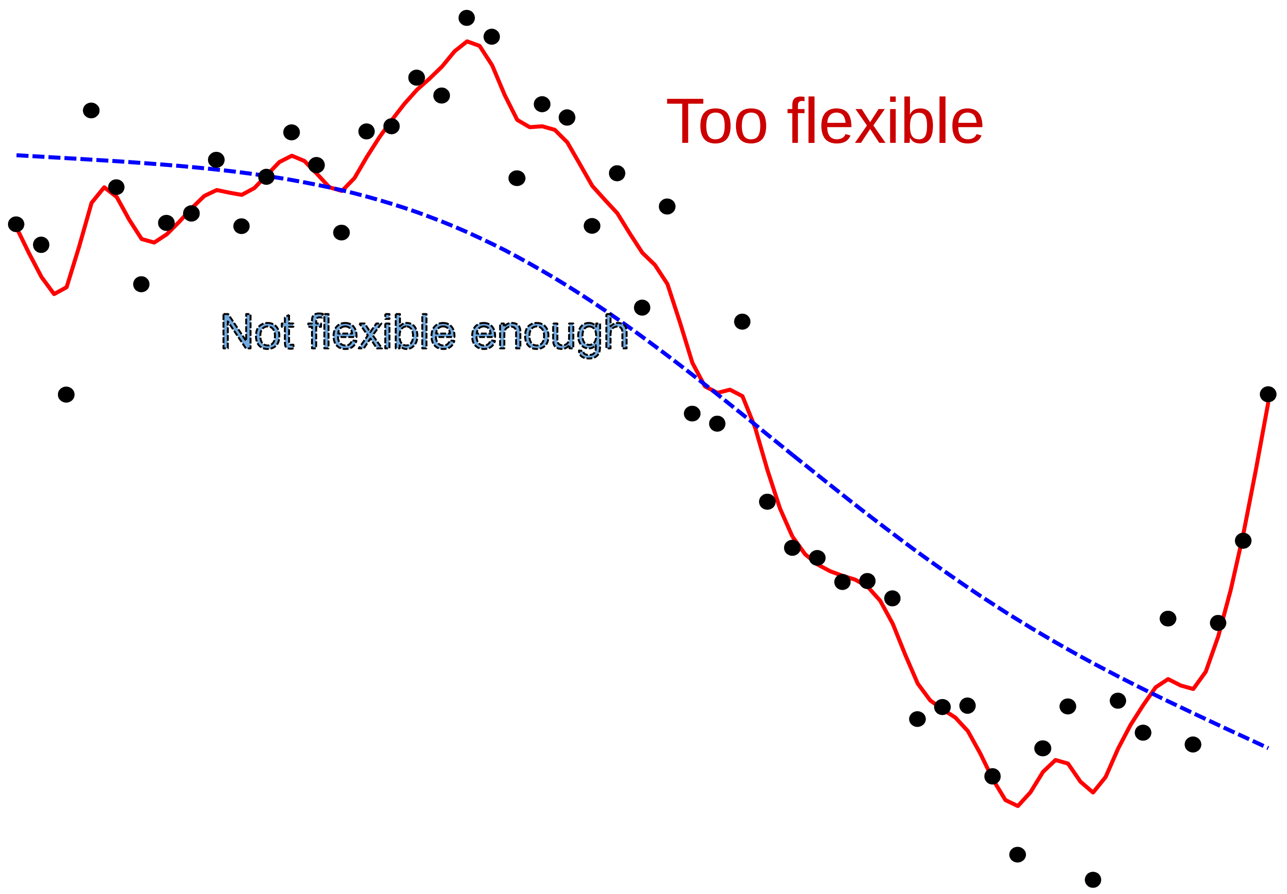
- Key considerations:
  - Accuracy
  - Interpretability
  - Ease of modeling
  - Robustness of model
  - The ease of handling missing values

Model Comparison

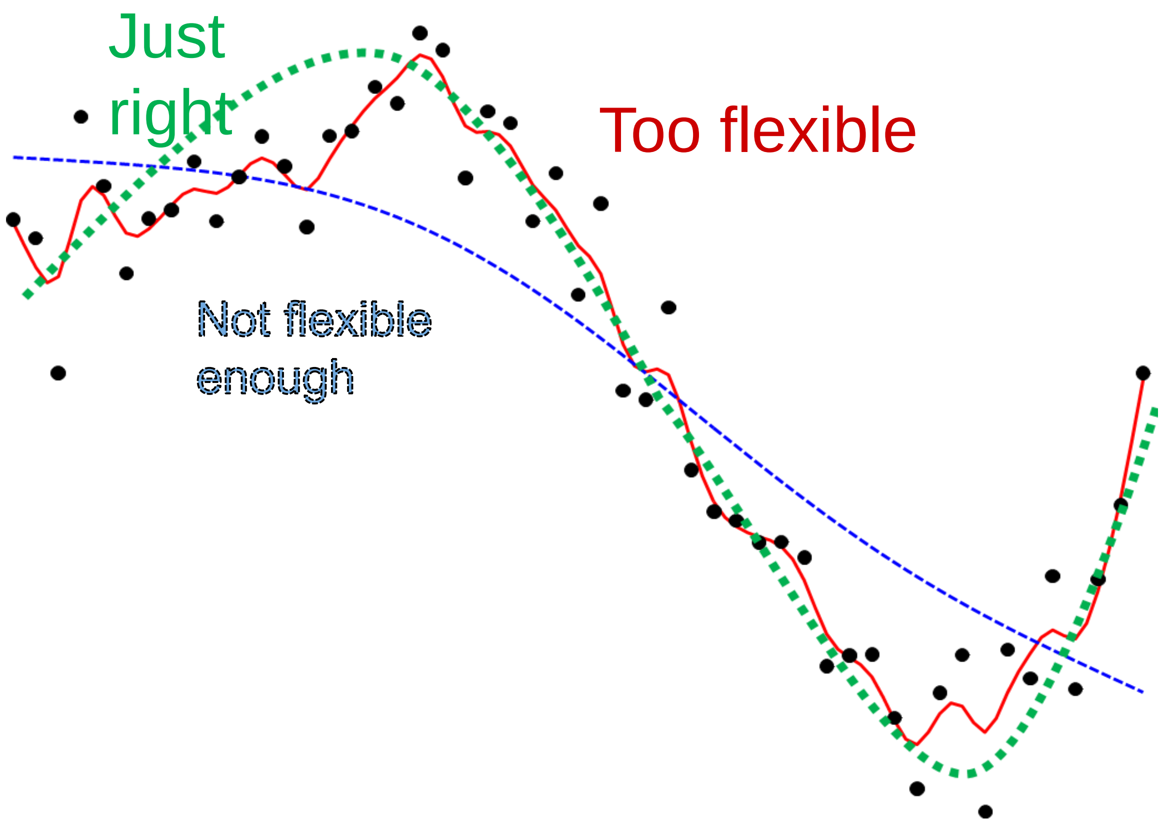
Model Comparison



Model Comparison



Model Comparison



8. Model Evaluation

- The goal is to accurately predicting some outcomes
  - In the context of sales forecasting, where we want to develop a model that best predicts future sales
  - It may be interesting and lead to key insights to think about variables that drive sales, for the purpose of best matching stock with demand we may only want to know future sales with as much accuracy as possible
  - We would like prediction measures that reflect our goal, and give us information about the prediction accuracy

Regression Evaluation

- Average Error: $\frac{1}{n}\sum_{i=1}^{n} e_i$
- MAE (mean absolute error): $\frac{1}{n}\sum_{i=1}^{n}|e_i|$
- MAPE (mean absolute percentage error): 100%* $\frac{1}{n}\sum_{i=1}^{n}|e_i/y_i|$
- RMSE (root-mean-squared-error): $\sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^{\;2}}$
- Total SSE (total sum of squared errors): $\sum_{i=1}^{n} e_i^2$

Model Evaluation— Categorical Outcomes

		Observed		
		True	**False**	
Predicted	**True**	**True positive**	**False positive** (Type I error)	Precision = True positive/Predicted positive
	False	**False negative** (Type II error)	**True negative**	Negative predictive value = True negative/Predicted negative
		Sensitivity = True positive/Actual positive	Specificity = True negative/Actual negative	Accuracy=No. of correct decisions/All cases

Lift and Gain Charts

- ❖ Charts to evaluate performance of classification models
  - To compare predictive model to random events (i.e., no model)
  - Lift = $\frac{Predicted\ Model}{Random\ Selection}$
  - Gain: percentage of responses

Lift and Gain Charts: Example

- An email campaign with 20% average response rate

Total customer contacted	Responses
1000	200

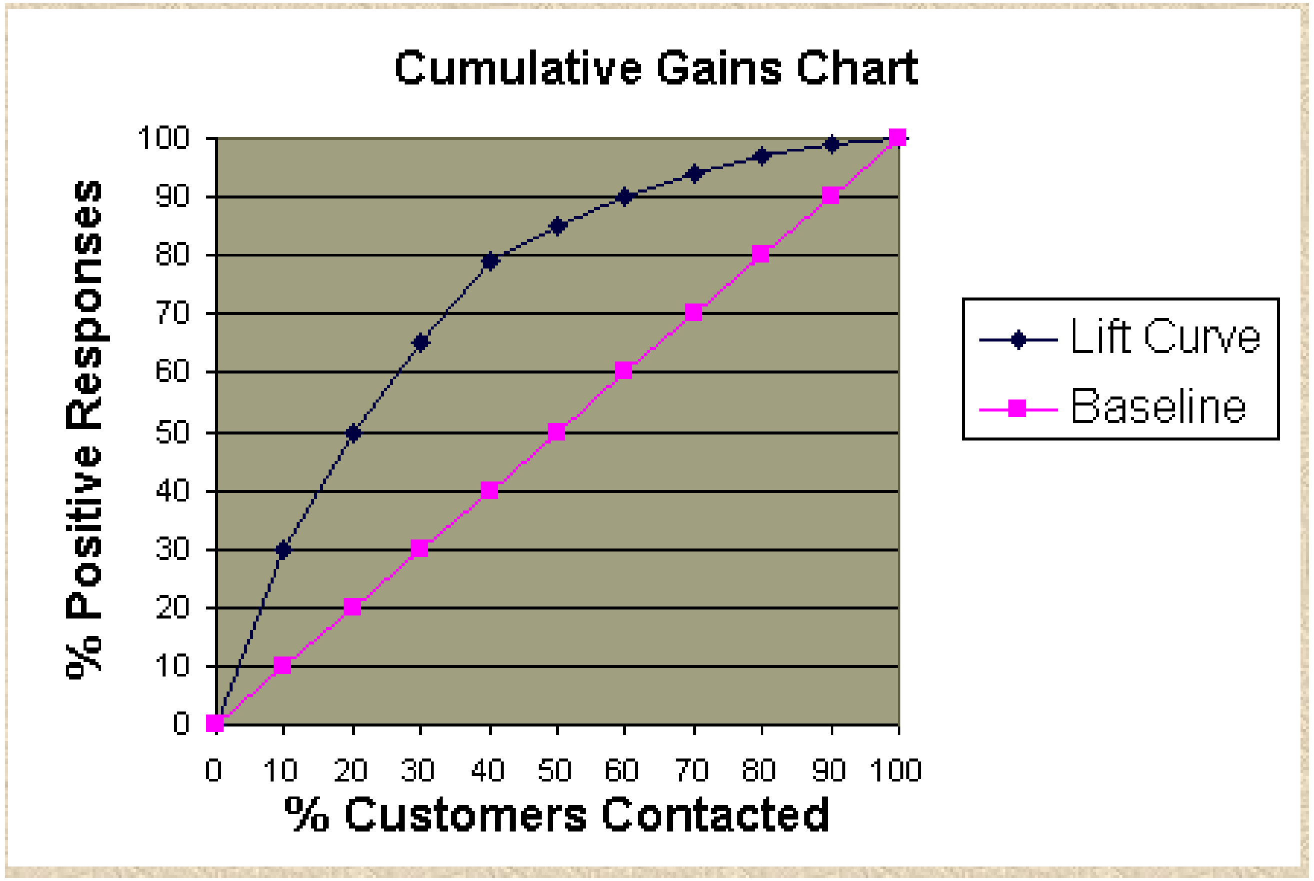
- Based on a predictive modeling, rank customers from most likely to response to least likely to response

Total customer contacted	Responses
100	60
200	100
300	130
400	158
500	170
600	180
700	188
800	194
900	198
1000	200

Lift and Gain Charts: Example

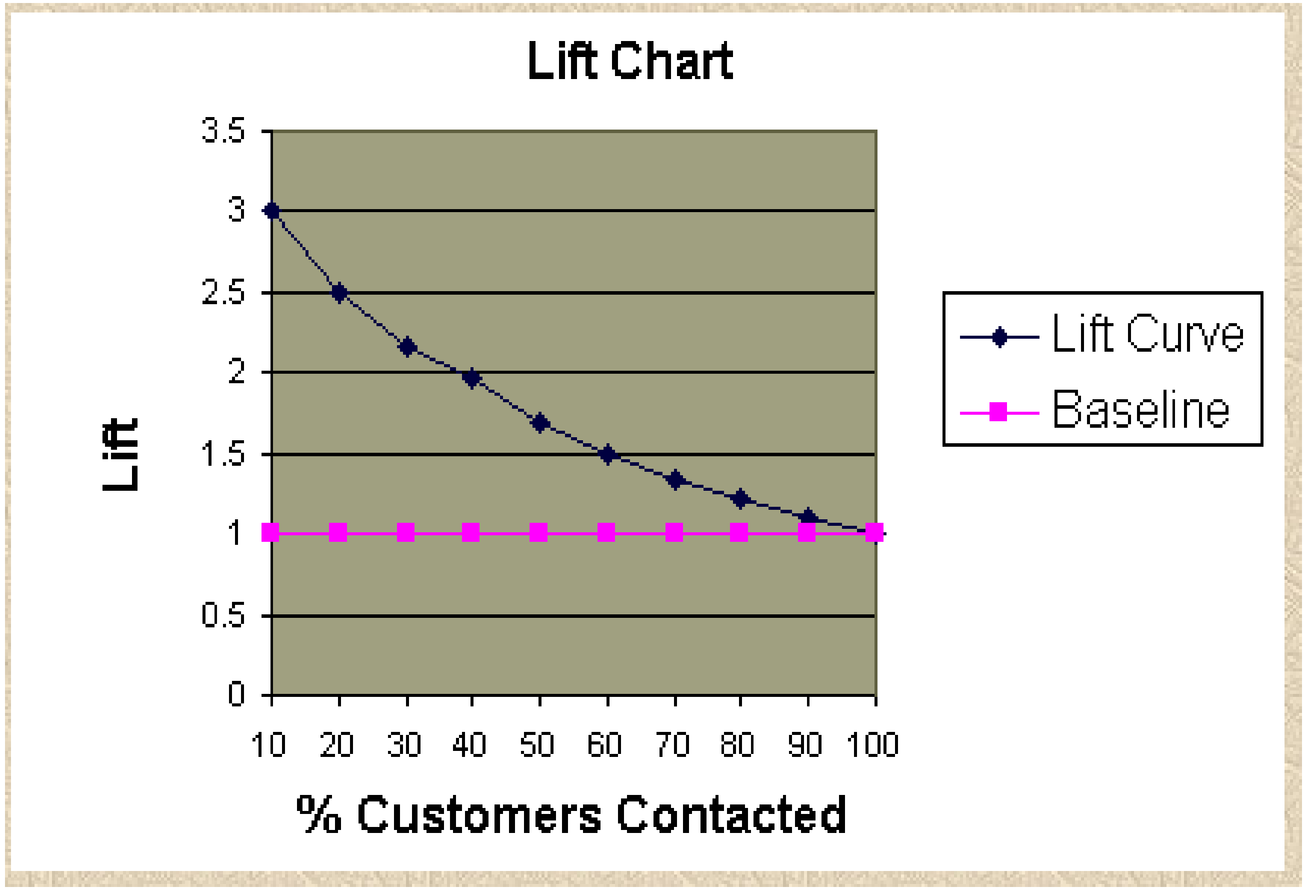
Total customer contacted	Model Responses	Random Selection	Cumulative Lift	Gain
100	60	20	=60/20 = 3	=60/200 =
200	100	40	= 100/40 = 2.5	= 100/200 = 50%
300	130	60		
400	158	80		
500	170	100		
600	180	120		
700	188	140		
800	194	160		
900	198	180		
1000	200	200		

Gain Chart



- X-axis: percentage of customers contacted
- Y-axis: percentage of responses
- Baseline: random selection
- Lift curve: predicted model

Lift Chart



- First 10% customers contacted:

Lift= $\frac{Predicted\ model}{Random\ selection}$

= $\frac{60\%}{20\%}$

=3