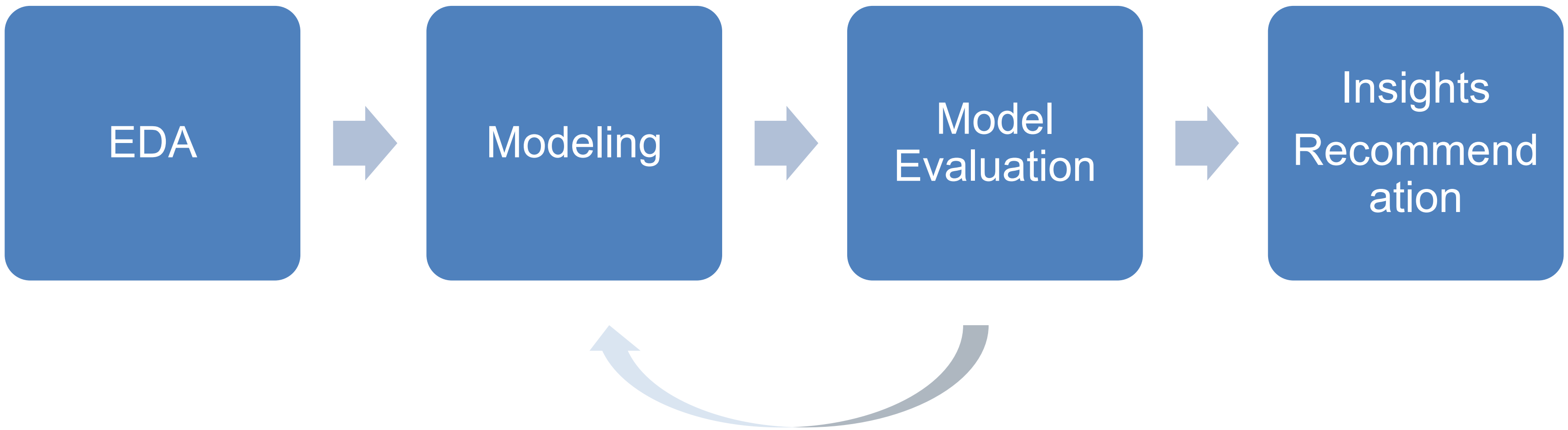
Machine Learning Project Procedure



Exploratory Data Analysis

- Data Visualization
- Descriptive Statistics
- Data Processing
  - Treat outliers
  - Treat missing values
  - Re-Categorize/Regroup values
    - E.g. Airlines: Korean Carriers vs. Foreign Carriers
    - E.g. Age: less than 50, 50 or more
    - E.g. Trip Purpose: Business, Leisure

- EDA (especially data processing) may determine overall results and quality of the project.

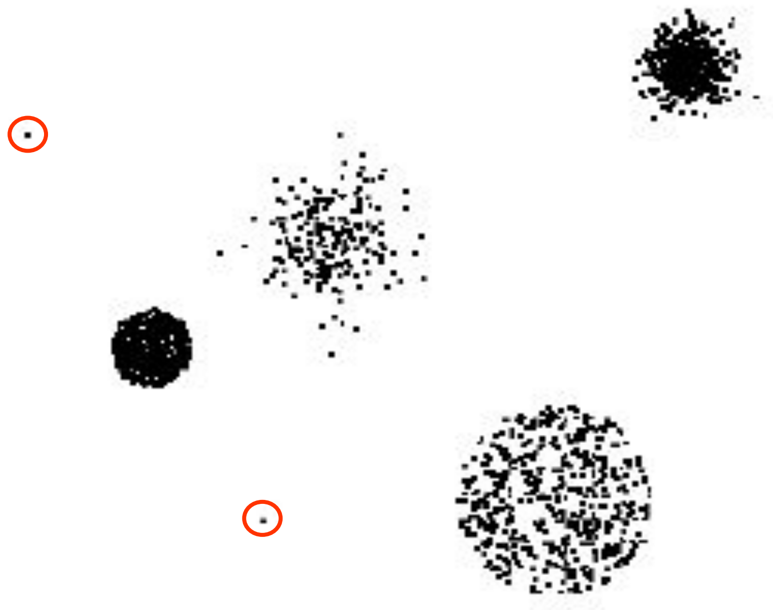
Data Quality

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

- Examples of data quality problems:
  - Noise and outliers
  - missing values
  - duplicate data

Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



Missing Values

- Reasons for missing values
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)

- Handling missing values
  - Eliminate Data Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)

Modeling

- Description Methods
  - Find human-interpretable patterns that describe the data
  - Vs. prediction methods
    - **Purpose of this study is not prediction.**

- Classification Problem
  - Dependent (y) variable is categorical.
    - Airport choice
    - Airline choice
  - Vs. regression problem

Methodologies

- Discrete Choice Models: Traditional approach (Econometrics)
  - Logit (Logistic regression)
  - Multinomial Logit
- Data Mining Models (You can explore other models as well.)
  - **Decision Tree**
  - Neural Networks*
  - Support Vector Machines*

Variable Selection Procedure

- Iterative Approach
  - <u>Stepwise</u>
  - <u>Forward selection</u>
  - Backward selection

one independent variable at a time is added or deleted based on selected measures (p-value, *F* statistic, R^2, …)

- Best-subset approach

Different subsets of the independent variables are evaluated

Variable Selection Procedure

- General tips on initial model building
  - Visualization and simple exploratory data analysis help a lot to identify key independent variables.
  - Correlation analysis and ANOVA can be used to identify initial set of independent variables.
    - Numerical dependent variable: High correlations between dependent and independent variables
    - Classification problem: ANOVA test

Model Evaluation for Statistical Models

- AIC, BIC, adjusted R^2
  - Can be used to measure training and test errors
    - often used for model selection on training data sets.
  - Usually works for statistical models
    - Regression
    - Logit models (logistic regression)
    - Other variations of linear models such as discrete choice models (multinomial logit, nested logit, mixed logit)

AIC, BIC, Adjusted R^2

- AIC (Akaike Information Criterion)

$$\mathrm{AIC} = -2\log L + 2 \cdot d$$

  - d: # of parameters, L: likelihood
    - The AIC criterion is defined for a large class of models fit by maximum likelihood.

- BIC (Bayesian information criterion)

$$\mathrm{BIC} = \frac{1}{n}\left(\mathrm{RSS} + \log(n) d\hat{\sigma}^2\right)$$

  - RSS: residual sum of squares
    - BIC will tend to take on a small value for a model with a low test error, and so generally we select the model that has the lowest BIC value

AIC, BIC, Adjusted R^2

- adjusted R^2

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}.$$

  - TSS: total sum of squares
  - Unlike the R^2 statistic, the adjusted R^2 statistic pays a price for the inclusion of unnecessary variables in the model.

Model Evaluation for Classification

- Confusion Matrix:
  - Can be used to measure training and test accuracy
  - Usually requires hold-out (test) dataset

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

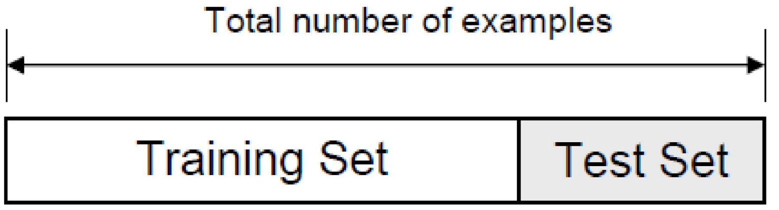
d: TN (true negative)

Validation

- Holdout method
- Cross validation
  - Random subsampling
  - K-fold cross validation

Holdout Method

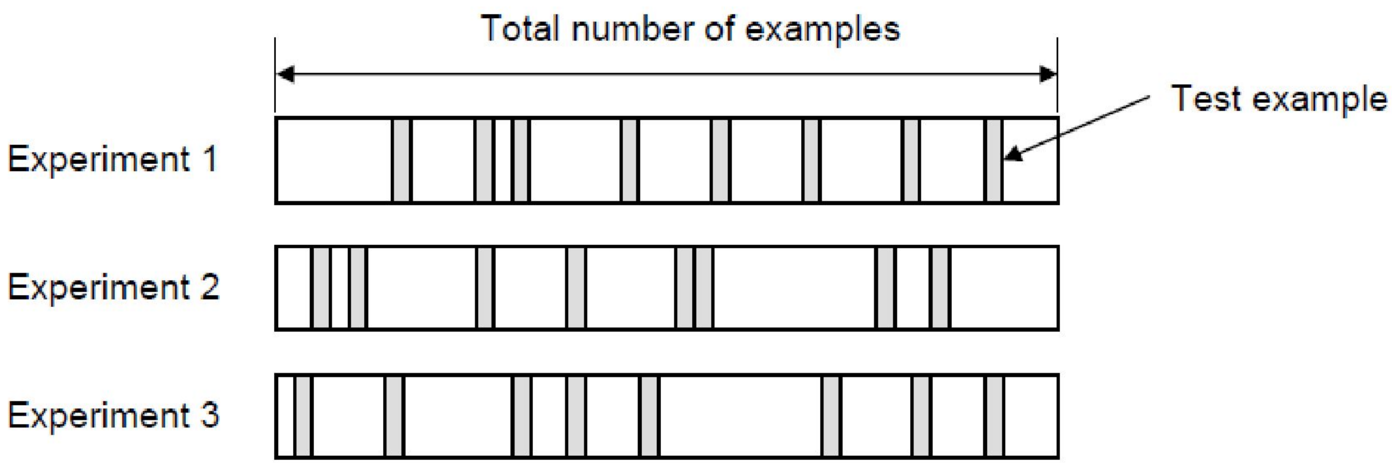
- Split dataset into two groups
  - Training set
  - Test set



- Drawback
  - In problems where we have a sparse dataset we may not be able to afford the “luxury” of setting aside a portion of the dataset for testing
  - Since it is a single train-and-test experiment, the holdout estimate of error rate will be misleading if we happen to get an “unfortunate” split

Random Subsampling

- Random Subsampling performs K data splits of the entire dataset
  - Each data split randomly selects a (fixed) number of examples without replacement

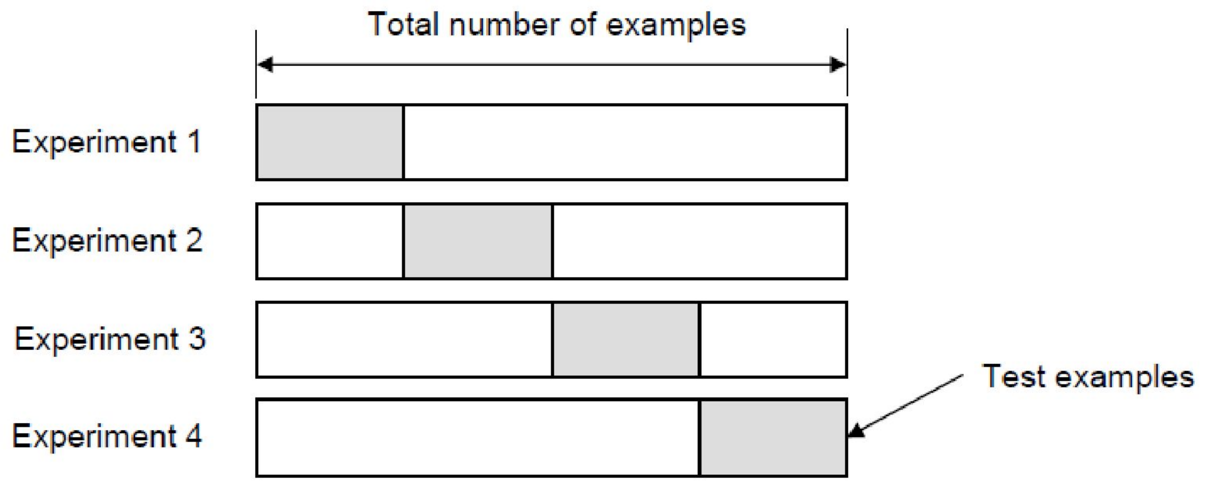


  - Error estimate E: average of the separate errors E_i

$$E = \frac{1}{K}\sum_{i=1}^{K} E_i$$

K-fold Cross Validation

- Create a K-fold partition of the dataset
  - For each of K experiments, use K-1 folds for training and a different fold for testing



  - all the examples in the dataset are eventually used for both training and testings
  - Error estimate E: average of the separate errors E_i

$$E = \frac{1}{K}\sum_{i=1}^{K} E_i$$

How many folds are needed?

- With a large number of folds
  - (+) The bias of the true error rate estimator will be small (the estimator will be very accurate)
  - (-) The variance of the true error rate estimator will be large
  - (-) The computational time will be very large as well (many experiments)
- The choice of the number of folds depends on the size of the dataset
  - For large dataset, smaller K may be enough.
- A common choice for K-Fold Cross Validation is K=10