Advanced Web Analytics: Harnessing the Predictive Power

By Dr. Ash Pahwa

Lesson 1: Introduction to Analytics

Lesson 1.1: What is Predictive Analytics + CRISP/DM

Lesson 1.2: Data Problems in Predictive Analytics

Outline: Lesson 1.1:
What is Predictive Analytics + CRISP-DM Model

- Definition and Benefits of PA
- Other Technologies
    - Machine Learning, Data Mining, Predictive Analytics
- Classification Algorithm
- Supervised versus Unsupervised Modeling
- Supervised modeling Goals
    - Classification, Estimation
- Supervised Learning Techniques
- Data
- Model Development and Deployment
- CRISP-DM Model
- Summary

What is Predictive Analytics?

- Predictive Analytics is a Business Intelligence Technology that provides
  - Predictive Score for each customer
  - This score represents the probability that they will respond to certain offer

Copyright 2014 - Dr. Ash Pahwa
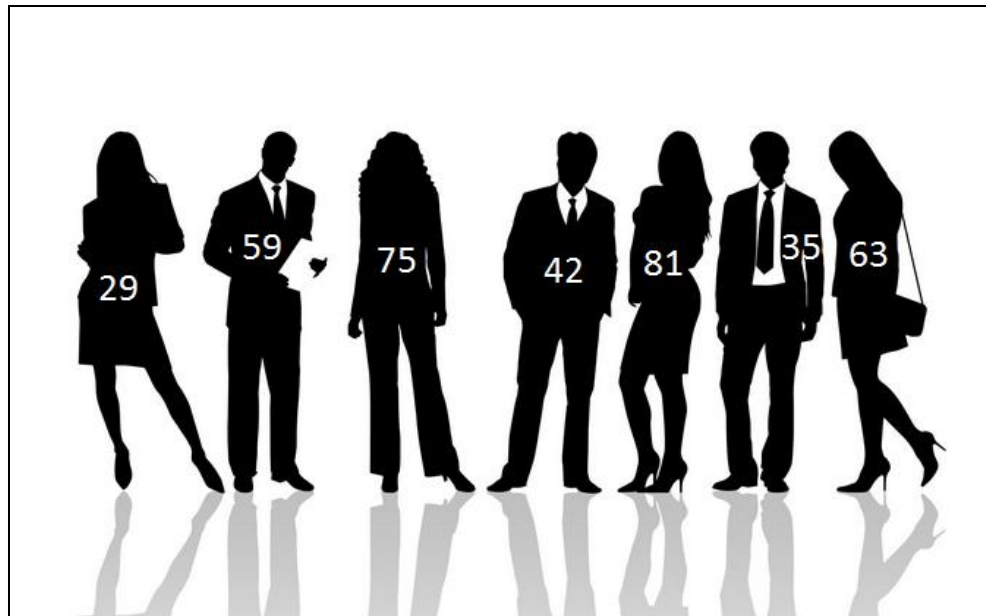
What is Predictive Analytics?

- Predictive Score will predict the customer behavior - chances that that customer will
  - Click on an ad
  - Respond to this offer
  - Cancel their subscription
  - Turn product B for their next purchase
  - Any other behavior relevant to your business
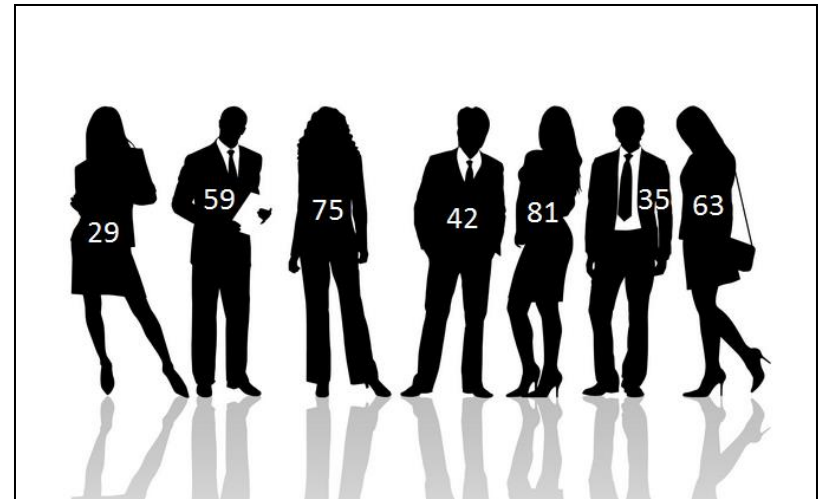
Benefits of Predictive Analytics

- Predictive Analytics turn
    - Uncertainty about the future into
    - Usable probability
- Learning from the data
    - And Using that knowledge to action

Benefits of Predictive Analytics

- Better risk management
  - During the transactions
  - Not after the fact
- Managing complex tradeoffs
- Discovering unexpected patterns
- Better management of uncertainty

Other Technologies

- Machine Learning
  - Machine learns from the corporate data
  - Machines finds something new
- Data Mining
  - Data has many patterns
  - Data mining allows us to see those patterns
- Predictive Analytics
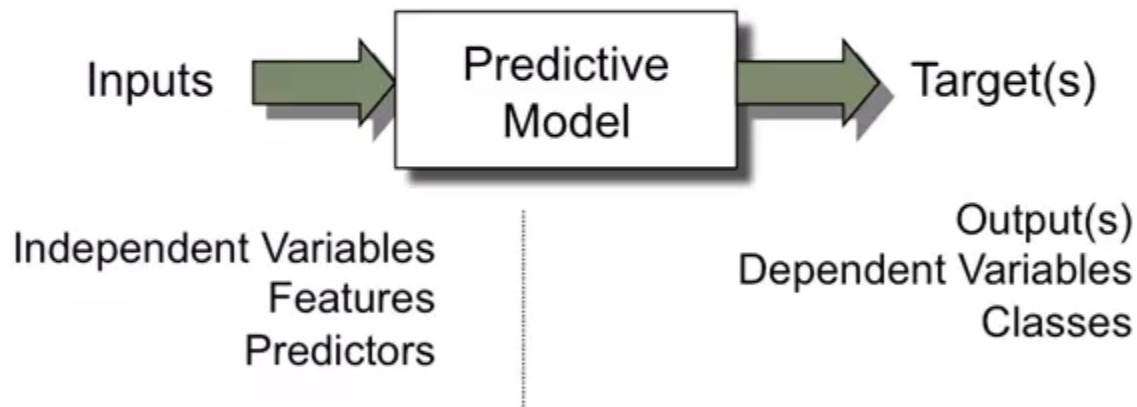 - Giving a predictive score to each customer for a certain offer

What is a Classification Algorithm

- Bunch of input data
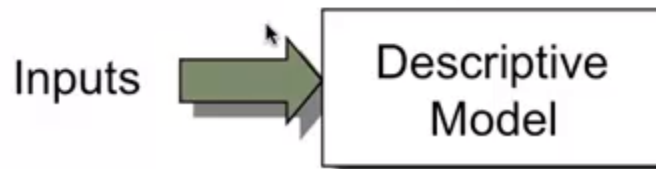- Need to decide in which bin each of the data falls in
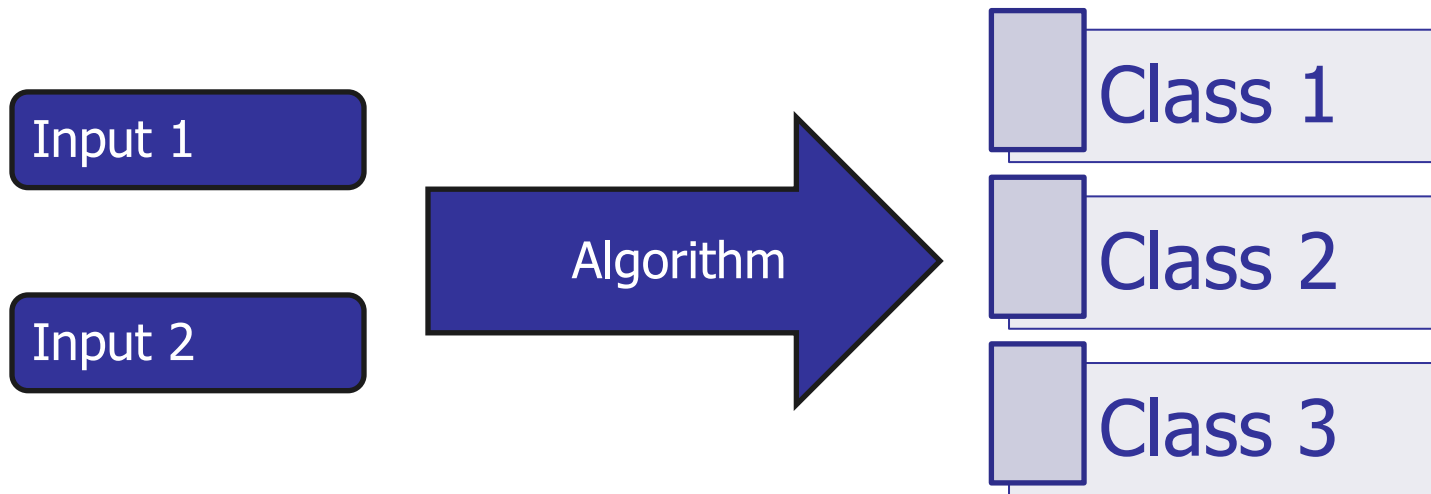
Input 1

Input 2

Algorithm

Class 1

Class 2

Class 3

Predictive Modeling Terms
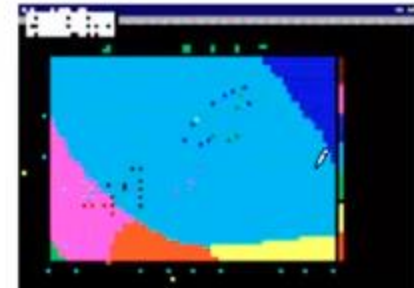
Predictive Analytics Goals: Supervised Learning

- Supervised Learning
  - Known answer exists as a column in the data
  - Two types: estimation and classification

- Classification
  - Yes/No decisions, or multi-value discrete decisions
  - Ex: Is the transaction fraudulent?

- Estimation
  - Continuous value output
  - Ex: What is the price we expect Apple stock to be tomorrow?
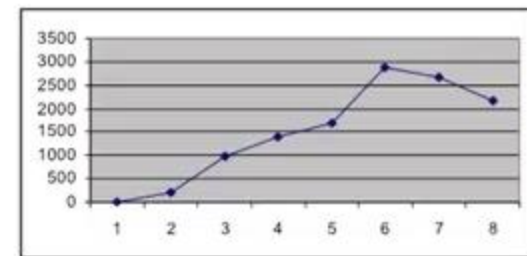
Common Techniques Used in Predictive Modeling:

- Regression - linear, logistic. Regularization, ridge, recursive partitioning
  - The most common statistical technique used in predictive analytics
  - Measures correlations between the variable to be predicted and several other variables.
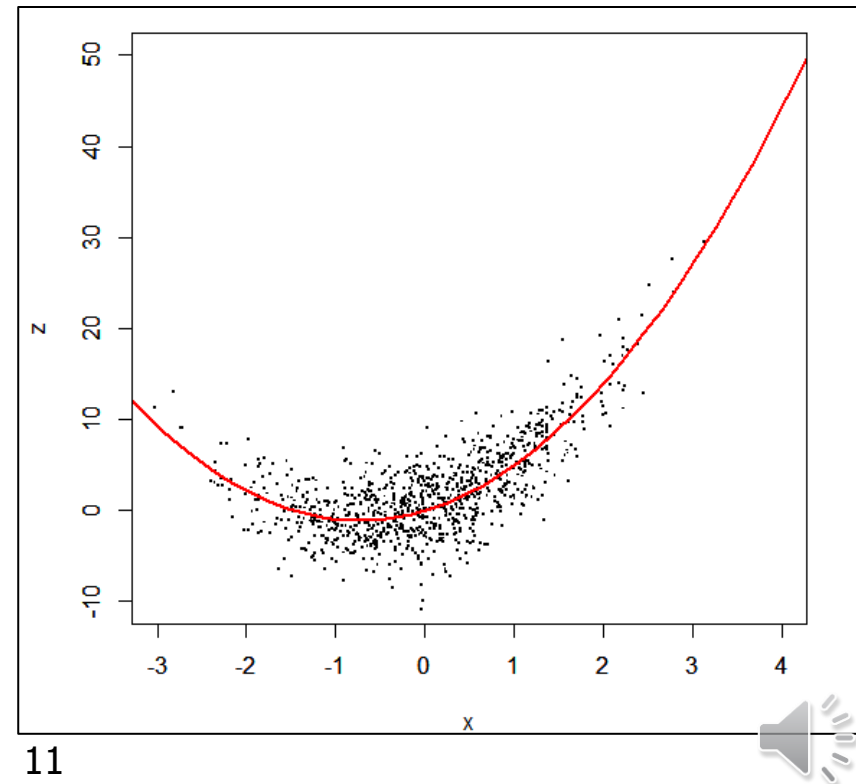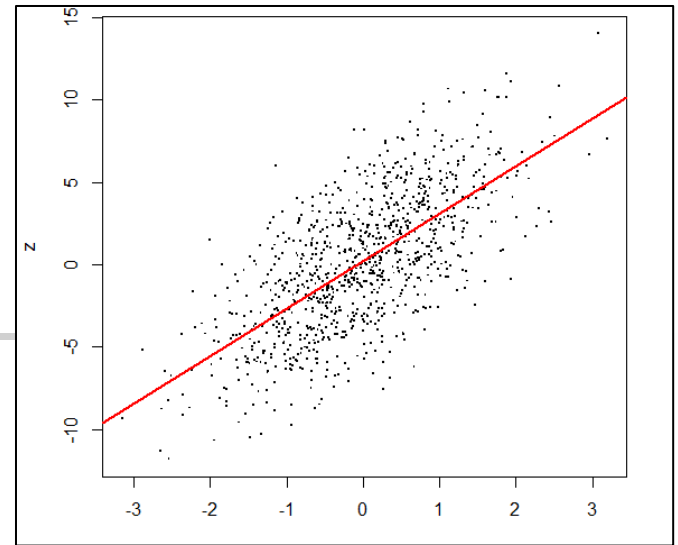
Common Techniques Used in Predictive Modeling:

- Clustering
  - Uses: segmentation of customers, identification of similar products, identification of disease or crime locations, etc.
  - Does not require a variable to be predicted, but several unrelated variables can be used as input.
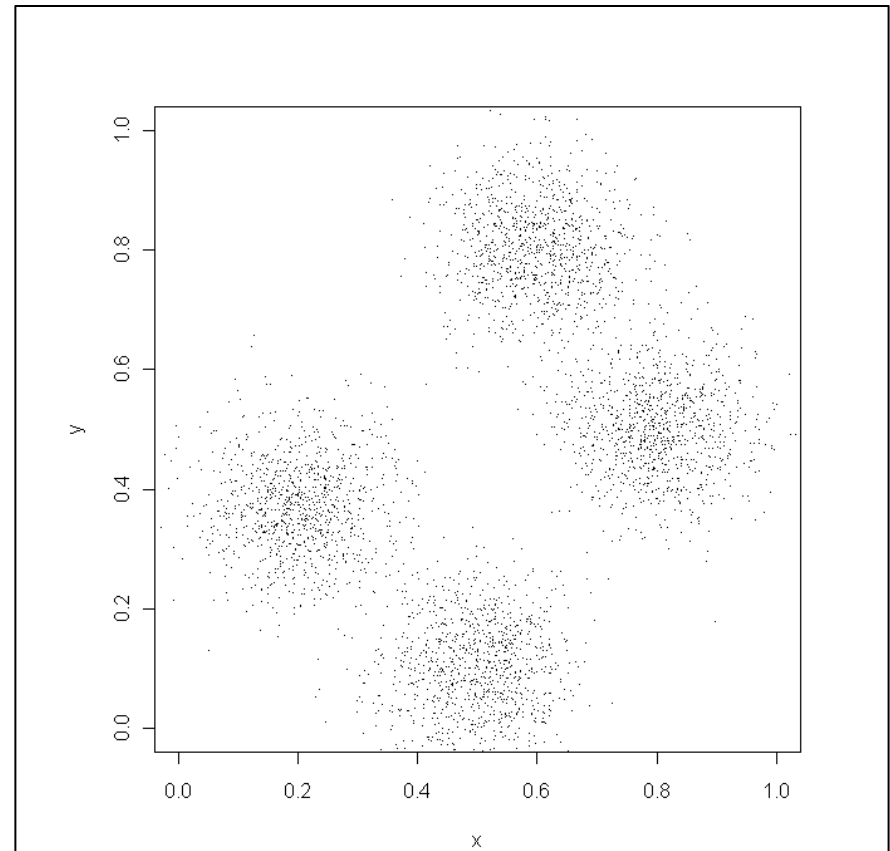
Common Techniques Used in Predictive Modeling:

- ARIMA for time series
  - Uses: forecasting revenue in future quarters
- Simulations
  - Uses: for complex systems, such as weather forecasting, online ad campaign performance
- Naive Bayes
  - Uses: deciding whether an email belongs in a spam folder
- Decision Trees / Random Forest
- Neural Networks

Forecasts from Holt-Winters' additive method
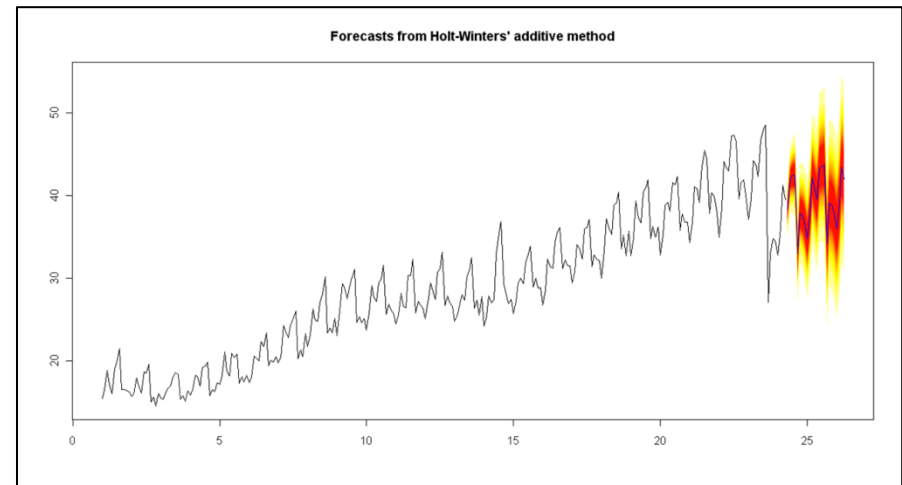
Which Technique is the Best?

- Why do we consider many different techniques?
- Which one is the best?
- No one technique is the best
- All depends upon the data
- Some technique will work better on some data

Hierarchy of Data Analysis

CRM Prediction

PREDICTIVE ANALYTICS

Advanced Analytics

Foundation Analytics
Segmentation: Recency,
Frequency, Monetary

Performance reporting and sales forecasting:
OLAP: Online Analytical Processing

Aggregate Data with Data Warehousing

Data

- Customer Profile
  - Address, age, income gender etc.
- Customer Behavior
  - Purchase history of all customers
  - Personal Income
  - Campaign: customer responded or not
  - How many times that person visited the website
- Data (External)
  - Demographic Data
    - Address (zip code), age, gender, email address etc.

Data

- 90% of a successful outcome is contingent on having good data. What does "good" mean in this context?
  - **Reliable** - the effects can be reproduced. Contrary example: Using sales data from March-April to predict November-December data.
  - **Valid** - the data measures what you want it to measure. Contrary example: Number of page visits shows how popular the page is. How well a job candidate does on a brainteaser indicates whether he will be a good hire.
  - Other data issues are small.

Data Problems

- Missing Data
- Sparse Data
- Inaccurate data
- Data Preparation is needed before building a model

18

Model Development and Deployment

Model Development

| Training Data | Modeling Tools | Predictive Model |

Model Verification

| Test Data | Predictive Model | Verification of Model |

Model Deployment

| Real World Data | Predictive Model | Prediction Response |

Problems in Building Model Memorizing the Training Data

- We cannot memorize the training data
  - Training Data cannot become a look-up table
- There are too many possibilities
  - Every customer is unique
- Instead of memorizing the training data
  - We need to build the generic model
- Find a Model that holds in general
  - Consistent
  - Accurate

Deploy to Take Business Action

Predicted Response → **Business Logic** →

Business Actions

Mail a solicitation

Suggest a cross-sell option

Retain with a promotion

Example-1

- Whether a customer will renew a subscription
- Predictor variables
  - Rural or urban
  - Monthly usage of website
  - New feature exploration of the website
- Result
  - If customer is rural & monthly usage is high
    - Customer will renew
  - If customer is urban & new feature exploration is high
    - Customer will not renew

Example-2

- Customer who buys product A
  - Are likely to buy product B
- Deployment would be
  - Strategy 1
    - Send those customers a coupon of $x to buy B
  - Strategy 2
    - Do not send them a coupon
    - Because they would buy it anyway. Why loose $x of revenues off those customers

Difference between Forecasting and Predictive Analytics

- Forecasting
  - In 2Q this year we sold 120 units
  - In 3Q this year we will sell 140 units
- Predictive Analytics
  - Provides an individual score to every customer which indicates the chances that that customer will behave in a certain way

CRISP-DM Process Model

- www.crisp-dm.org
- CRoss Industry Standard Process for Data Mining

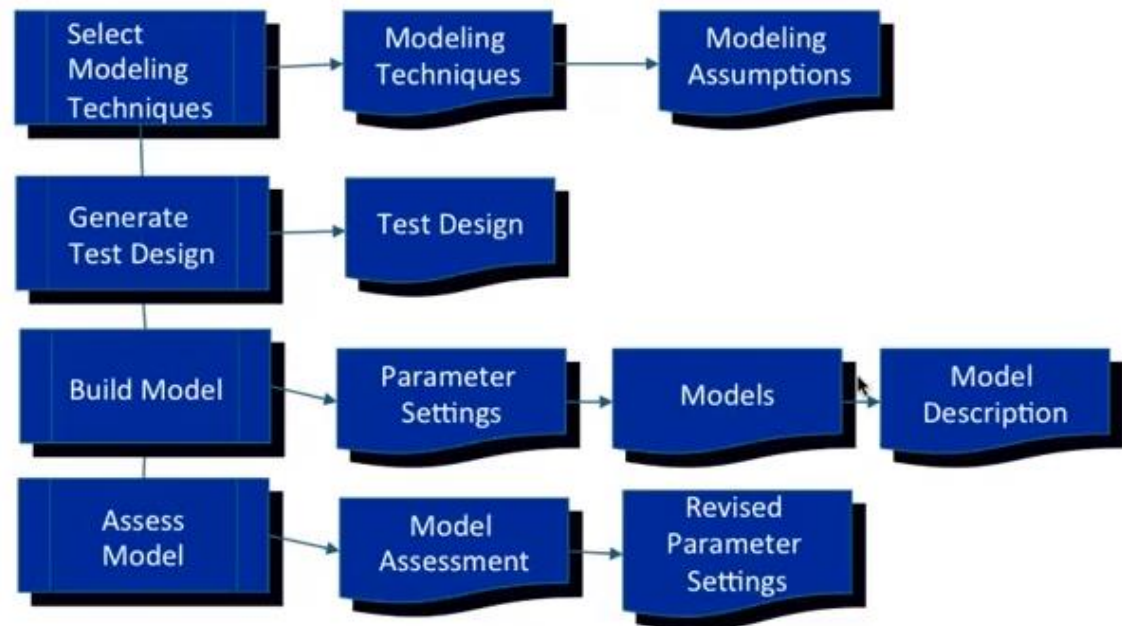The word Data Mining can be interchanged with Predictive Analytics.
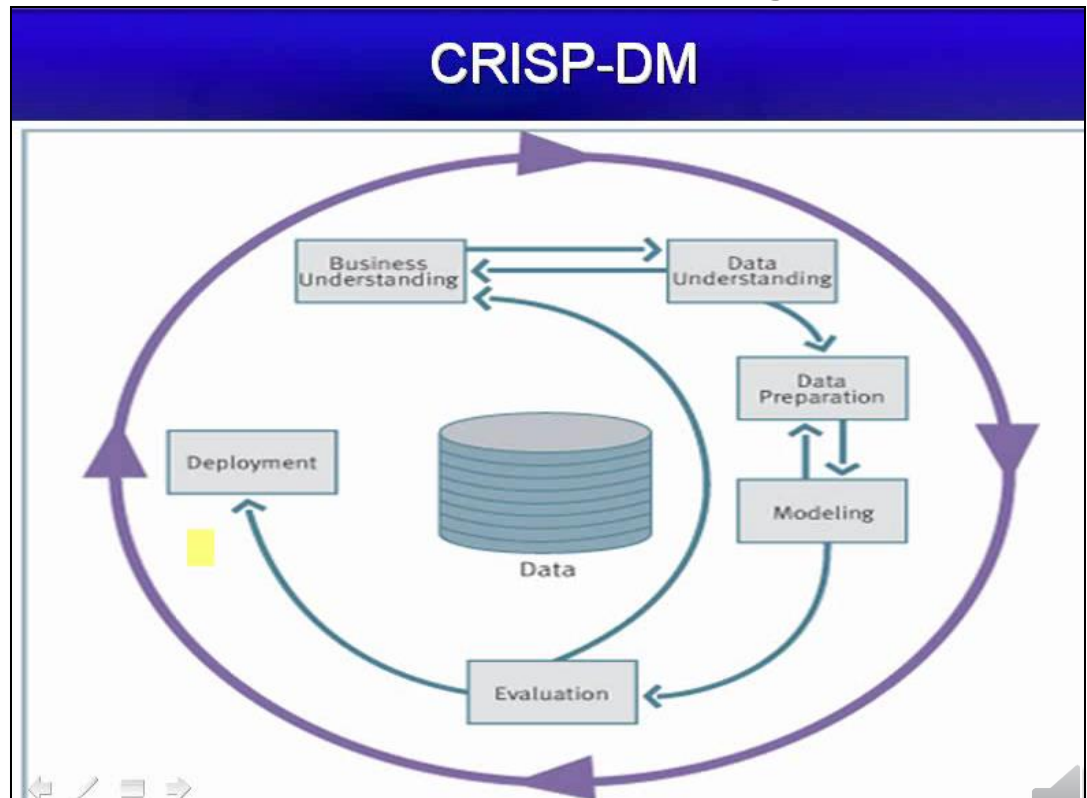
CRISP-DM Process Model

- Step #1
  - Start with business understanding of what you want to do with data mining
- Step #2
  - Interactions between business understanding and data understanding
- Step #3
  - Data preparation
  - Interactions between data preparation and modeling
- Step #4
  - Interactions between model evaluation and business understanding
- Step #5
  - Deployment of Model
- Step #6
  - Results achieved from PA should be compared with the business understanding

CRISP-DM Step 4: Modeling Steps

- Select techniques
  - Supervised / unsupervised
  - Target variable type
- How to select models?
- Model specifications
  - Learning parameters, stop rules
- Model interpretation

Select Modeling Techniques → Modeling Techniques → Modeling Assumptions

Generate Test Design → Test Design

Build Model → Parameter Settings → Models → Model Description

Assess Model → Model Assessment → Revised Parameter Settings

Summary: Lesson 1.1:
What is Predictive Analytics + CRISP-DM Model

- Definition and Benefits of PA
- Other Technologies
  - Machine Learning, Data Mining, Predictive Analytics
- Classification Algorithm
- Supervised versus Unsupervised Modeling
- Supervised modeling Goals
  - Classification, Estimation
- Supervised Learning Techniques
- Data
- Model Development and Deployment
- CRISP-DM Model
- Summary