THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Business Analytics (MM 5425)

L2. Data Mining Basics

Dr. Yue (Katherine) FENG

Agenda

I. What is business analytics?

II. Comparison of data analytical methods

III. Data mining terminologies

What is Business Analytics?

Non-Trivial/Meaningful Patterns

> Beers and diapers are often purchased together by customers.

More Data-Driven Cases - Amazon

Anticipatory Shipping: ship a package before you order it !

- Cut delivery time
- Discourage consumers from visiting physical stores

Recommendation system (collaborative filtering)

- Scale to massive data sets and produce high quality recommendations in real time

An Example: Customer Retention

> **Which customers to target with a special offer, prior to contract expiration?**

Traditional Solutions

- Offer incentives to every customer before contract expires.
- Contact each customer to probe propensity to terminate contract.
- …

BA Solutions

> **Decision tree**

- If Education = low and Gender = male, then customer will be likely to churn.

> **Logistic regression**

- Calculate the probability of churning given the features of a customer.

> **Nearest neighbor**

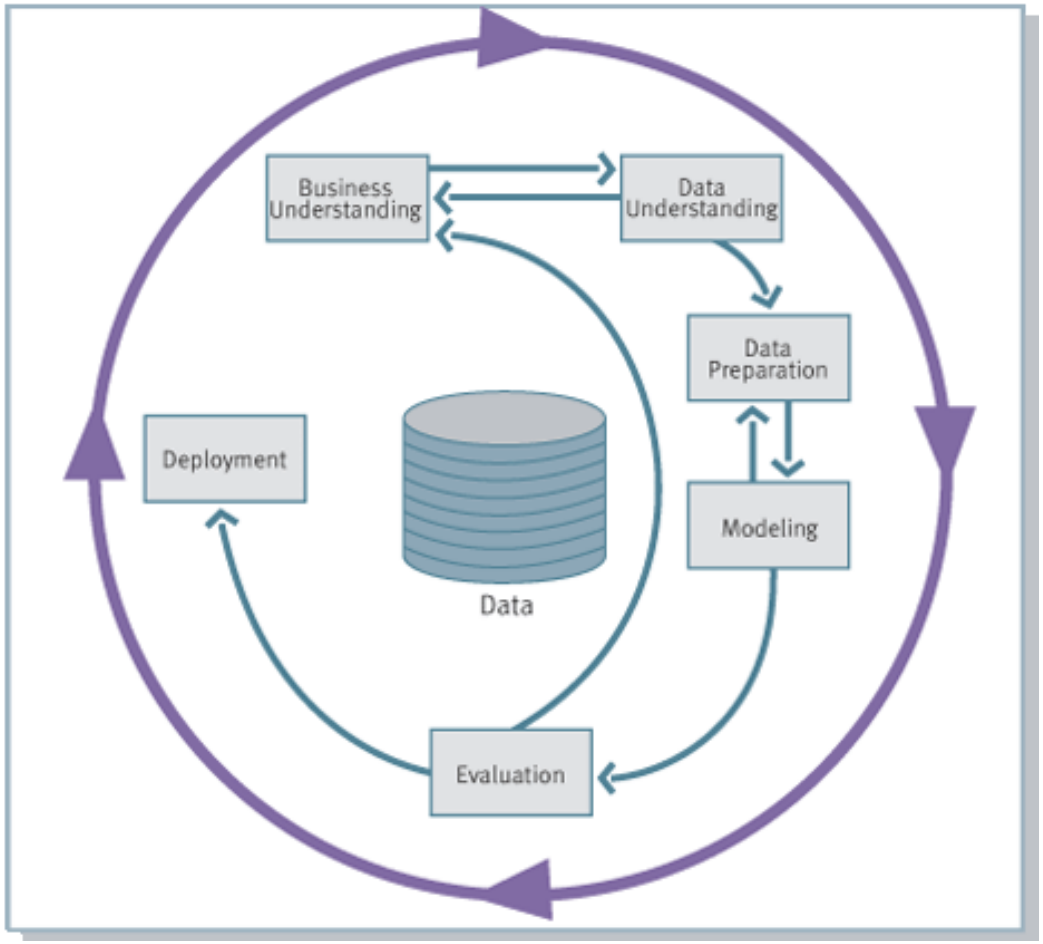
- Calculate how similar a customer is to existing churning customers.

What is Business Analytics (BA)?

> Business Analytics (BA) is the practice and art of bringing ***quantitative data*** to bear on ***decision-making***.

- It is an end-to-end ***process*** from business understanding to production deployment.
- It includes a range of data analysis methods, more than counting, rule-based checking, and basic statistical summary.

How to Run a BA Project? A Process View

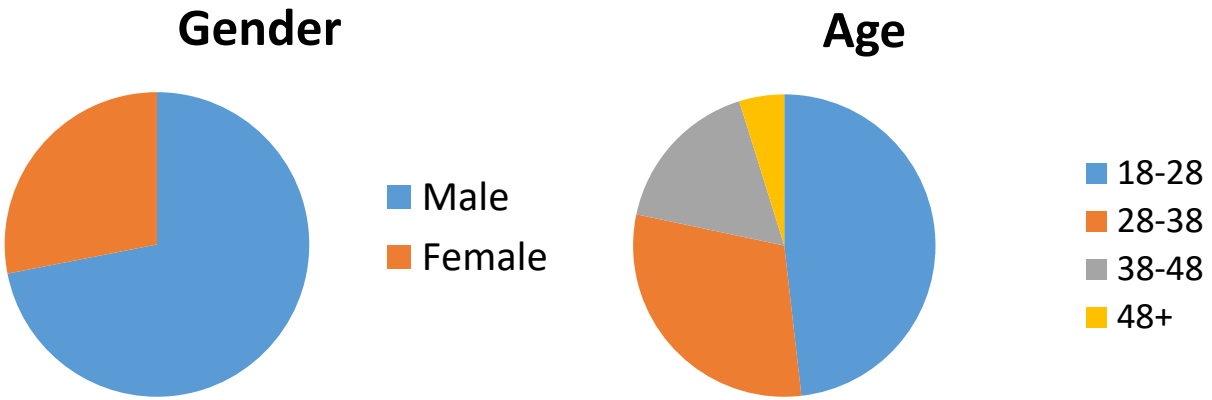


Comparison of Data Analytical Methods

Data Description

> Typically focuses on **current facts** (vs. future outcomes)

> Ad hoc **queries and reports**, not modeling (vs. statistical and machine learning techniques)



Statistical Analysis

> Typically based on ***hypothesis testing*** and estimation of parameters

> Concerned more about **causality** than correlation

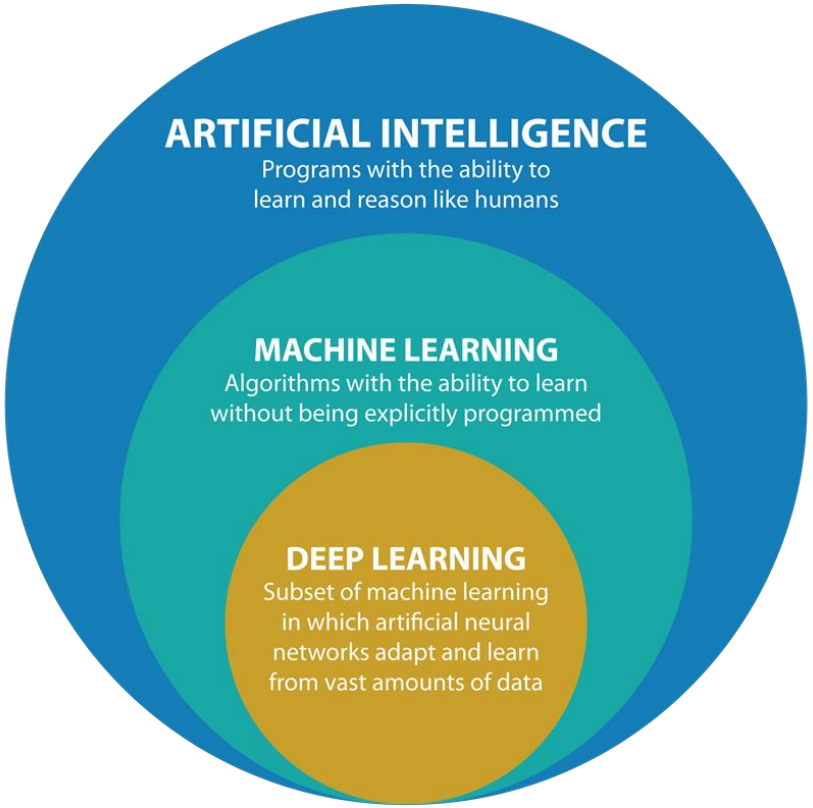
> Target to identify the effects and explain the **underlying mechanisms**

Data Mining/Machine Learning

> Learn interested **patterns from existing data** by statistical modeling and apply the model for future prediction

> Concerned more about **correlation** than causality

> Target to **predict** future outcomes and improve business decisions/efficiency



Exercises: Which method will you apply to solve the below problems?

1. Who exactly are my most spending customers?
2. Is there really a difference in spending between these customers and the average customers?
3. Can I characterize these customers and separate them from other customers?
4. Can I predict whether a new customer will be profitable?

Objectives Achieved by Data Analytical Methods

- **Describe** customers' decisions
- **Explain** factors affecting customers' decisions
- **Predict** the business outcomes

Data Mining Terminologies

Data

> Example (Instance)
- A fact typically includes a set of attributes (fields, variables, features)

> A data set
- A set of examples
- A sample/subset of the universe

Variables

Name	Balance	Age	Default
Mike	123,000	50	No
Mary	51,100	40	Yes
Bill	68,000	55	No
Jim	74,000	46	Yes
Dave	23,000	44	No
Anne	100,000	50	Yes

One example/instance

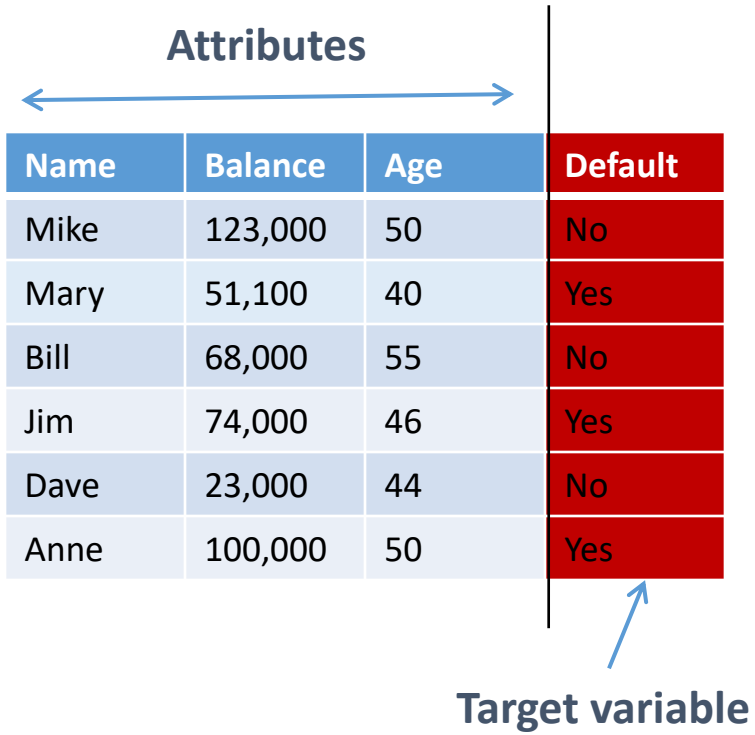
Variables

> **Target Variable**: A special variable that is the interest/target of the task.

<u>Equivalent statistics terminology</u>

> **Attributes/Features**: independent variables

> Target: dependent variable



Name	Balance	Age	Default
Mike	123,000	50	No
Mary	51,100	40	Yes
Bill	68,000	55	No
Jim	74,000	46	Yes
Dave	23,000	44	No
Anne	100,000	50	Yes

Model

> A Model (also referred to Theory)

- A summarization of relationships in the data. A description of the data in concise form. A general representation of reality created for a specific purpose.

Examples:

1. IF Balance >= 50K and Age > 45;
   Then Default = ‘no’; Else Default = ‘yes’
2. Default amount = 0.001*Income+2*Age

Supervised vs. Unsupervised Learning

> **Supervised learning:** captures relationships between a set of features and a pre-defined, known **target outcome**.

> **Unsupervised learning:** finds relationships in the data without reference to independent/dependent variables.

Key difference: is there a specific, objective *target* that we are trying to predict?

Supervised vs. Unsupervised: Customer Segmentation Example

> Do my customers fall into different groups with respect to their behavior of default?

[Supervised: use the segmentation to take action based on predicted likelihood of default]

Name	Balance	Age	Default
Mike	123,000	50	No
Mary	51,100	40	Yes
Bill	68,000	55	No
Jim	74,000	46	Yes
Dave	23,000	44	No
Anne	100,000	50	Yes

Induces a pattern from examples

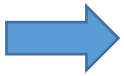
IF Balance >= 50K and Age > 45
Then Default = 'no'
Else Default = 'yes'

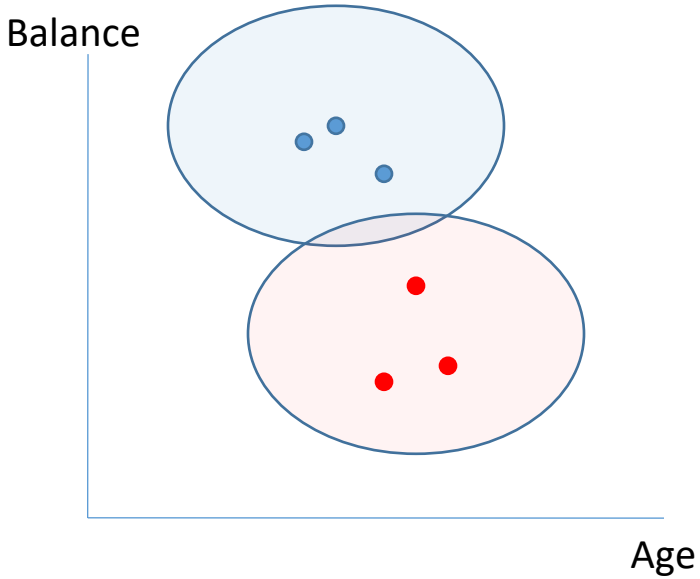
Supervised vs. Unsupervised: Customer Segmentation Example

> Do my customers naturally fall into different groups?

[Unsupervised: no objective target stated]

Name	Balance	Age
Mike	123,000	50
Mary	51,100	40
Bill	68,000	55
Jim	74,000	46
Dave	23,000	44
Anne	100,000	50





Classification vs. Regression

> Both are **supervised learning**

> The difference is the type of ***target variable***:

- classification → categorical target variable
- regression → numerical (continuous) target variable

Examples: Classification vs. Regression

1. Is this customer “loyal” or “likely to terminate contract”?
2. How much a customer is going to spend?
3. Is a credit card use “legitimate” or “fraudulent”?
4. Classifying credit risk as “high” versus “low”.
5. What is the credit score of a customer?

Classification vs. Regression

> **An important note**

- Most classifications are based on estimated class **probabilities**, which is numeric/continuous though.

> Example: the target for our churn example is “customer leaves before the contract expiration”. We may want to model the probability that a customer leaves. This is still considered classification modeling rather than regression, because the **underlying target variable is categorical**. Sometimes this is called “class probability estimation”.

Philosophy of Predictive Modeling

Name	Balance	Age	Default
Mike	123,000	50	No
Mary	51,100	40	Yes
Bill	68,000	55	No
Jim	74,000	46	Yes
Dave	23,000	44	No
Anne	100,000	50	Yes
…	…	…	…
…	…	…	…
Henry	61,100	35	???
Amy	68,000	52	???
Allen	22,000	21	???
Tom	123,000	60	???
Jane	100,000	47	???

Data universe

Data you have (a subset/sample from the universe of the data)

Build model

Model

Apply to new data

Use the subset of data you have to find the pattern of the universe.

Model Training vs. Testing

> Question:

- After an induction algorithm learns a model, can we have an estimate on how well our model will perform with new data?

> Solution: separate data that you have into two parts

- **Training** data to develop the model
- **Testing** data to evaluate performance of learned model on “new” data

Data Splitting for Training and Testing

Name	Balance	Age	Default
Mike	123,000	50	No
Mary	51,100	40	Yes
Bill	68,000	55	No
Jim	74,000	46	Yes
Dave	23,000	44	No
Anne	100,000	50	Yes

Training data

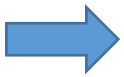
Name	Balance	Age	Default
Mike	123,000	50	No
Mary	51,100	40	Yes
Bill	68,000	55	No
Jim	74,000	46	Yes

Testing data

Name	Balance	Age	Default
Dave	23,000	44	No
Anne	100,000	50	Yes

Model Training

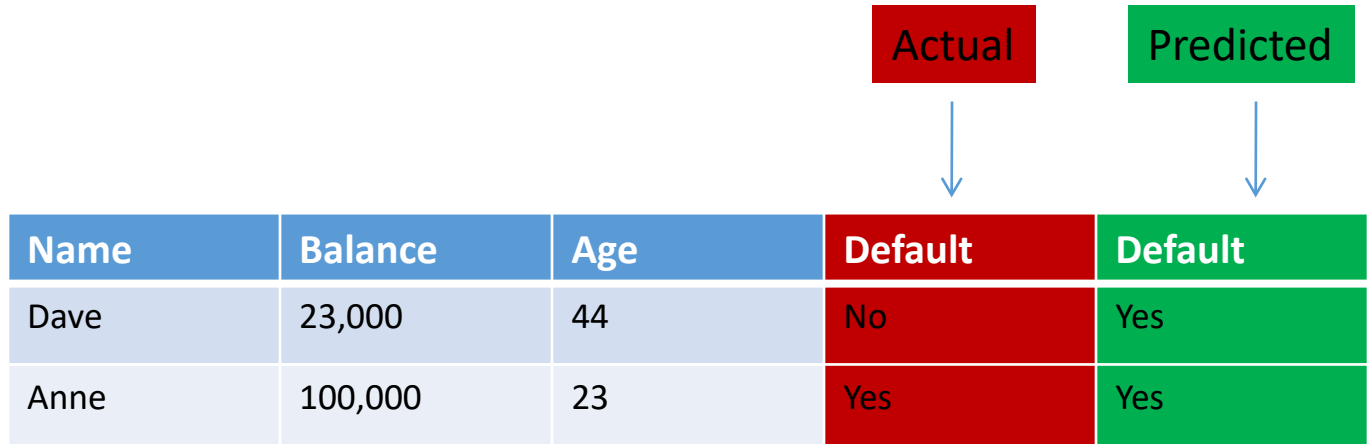
Name	Balance	Age	Default
Mike	123,000	50	No
Mary	51,100	40	Yes
Bill	68,000	55	No
Jim	74,000	46	Yes



Induces a pattern from examples

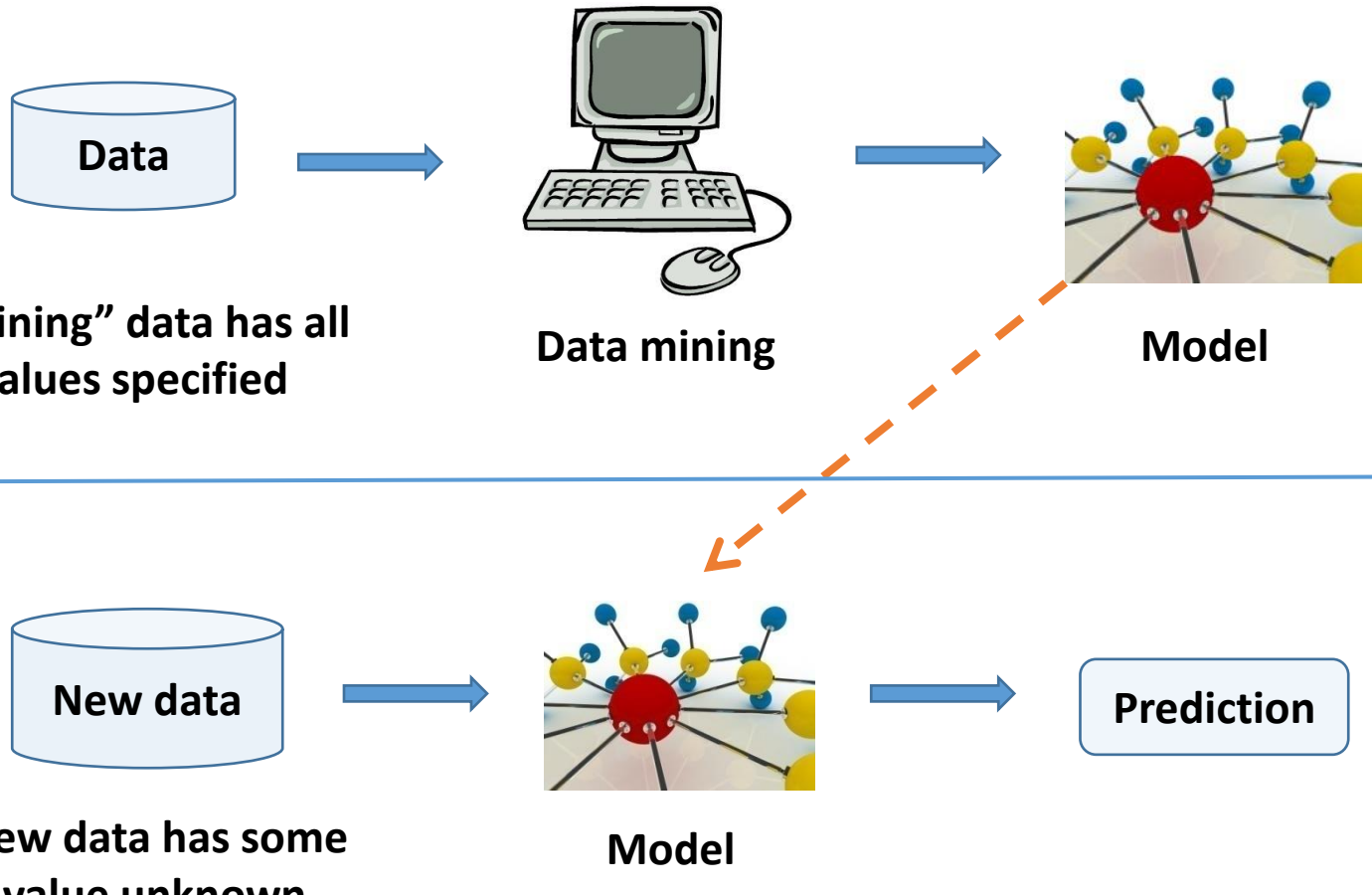
IF Balance >= 50K and Age > 45
Then Default = ‘no’
Else Default = ‘yes’

Model Testing



Name	Balance	Age	Default	Default
Dave	23,000	44	No	Yes
Anne	100,000	23	Yes	Yes

Modeling Phase vs. Use Phase



Basic Principles for Business Analytics

> Human decisions and behaviors are not random.

> Analyze patterns and relationships based on previous behaviors (historical data).

> Rely on these patterns and relationships to improve the decision-making process and business outcomes.

Thank You !

IDEAS Innovation-driven Education and Scholarship | Faculty of Business 工商管理學院

Discover · Design · Deliver

Department of MANAGEMENT & MARKETING 管理及市場學系

Opening Minds • Shaping the Future
啟迪思維 • 成就未來