Institut Informatika & Bisnis DARMAJAYA
Yayasan Alfian Husin

SAINS DATA
Institut Informatika & Bisnis Darmajaya

Kampus Merdeka INDONESIA JAYA

MERDEKA BELAJAR

DATA SCIENCE DARMAJAYA
"YOUR BEST FUTURE IN DATA"

MEETING: [3]

BUSINESS PROBLEMS AND DATA SCIENCE SOLUTIONS

BY: HENDRA KURNIAWAN

[12-09-2023] | [KODE MK: BDG23407 | MATAKULIAH: DATA SCIENCE AND BUSINESS ANALYTICS | SKS: 2/2 | VERSI: 01

Business Problems and Data Science Solutions

1. From Business Problems to Data Mining Tasks
2. Supervised Versus Unsupervised Methods
3. Data Mining and Its Result
4. The Data Mining Process
5. Implication for Managing The Data Science Team
6. Other Analytics Techniques and Technologies

Fundamental Concept

- An important principle of data science is that data mining is a process with fairly well understood stages.

- Some involve the application of information technology, such as the automated discovery and evaluation of patterns from data, while others mostly require an analyst's creativity, business knowledge, and common sense

Fundamental Concept

- Since the data mining process breaks up the overall task of finding patterns from data into a set of well-defined subtasks, it is also useful for structuring discussions about data science.

- This chapter introduces the data mining process, but first we provide additional context by discussing common types of data mining tasks.

From Business Problems to Data Mining Tasks

- Each data-driven business decision-making problem is unique, comprising its own combination of goals, desires, constraints, and even personalities.
- In collaboration with business stakeholders, data scientists decompose a business problem into subtasks.
- The solutions to the subtasks can then be composed to solve the overall problem.
- Some of these subtasks are unique to the particular business problem, but others are common data mining tasks.

From Business Problems to Data Mining Tasks

For example, our telecommunications churn problem is unique to MegaTelCo:

There are specifics of the problem that are different from churn problems of any other telecommunications firm. However, a subtask that will likely be part of the solution to any churn problem is to estimate from historical data the probability of a customer terminating her contract shortly after it has expired.

From Business Problems to Data Mining Tasks

- Despite the large number of specific data mining algorithms developed over the years, there are only a handful of fundamentally different types of tasks these algorithms address.
- “an individual” will refer to an entity about which we have data, such as a customer or a consumer, or it could be an inanimate entity such as a business.

From Business Problems to Data Mining Tasks

- Often we want to find correlation between a particular variable describing an individual and other variables. For example, in historical data we may know which customers left the company after their contracts expired.
- We may want to find out which other variables correlate with a customer leaving in the near future.
- Finding such correlations are the most basic examples of classification and regression tasks.

Classification

- Classification and class probability estimation attempt to predict, for each individual in a population, which of a (small) set of classes this individual belongs to.
- An example classification question would be: “Among all the customers of MegaTelCo, which are likely to respond to a given offer?”
    - Two classes could be called will respond and will not respond.

Regression

- Regression (“value estimation”) attempts to estimate or predict, for each individual, the numerical value of some variable for that individual.
- An example regression question would be: “How much will a given customer use the service?”
  - *classification* predicts whether something will happen
  - *regression* predicts how much something will happen

Similarity Matching

- Similarity matching attempts to identify similar individuals based on data known about them. Similarity matching can be used directly to find similar entities
- IBM is interested in finding companies similar to their best business customers, in order to focus their sales force on the best opportunities.
- They use similarity matching based on “firmographic” data describing characteristics of the companies.
- Similarity matching is the basis for one of the most popular methods for making product recommendations (finding people who are similar to you in terms of the products they have liked or have purchased).

Clustering

- Clustering attempts to group individuals in a population together by their similarity, but not driven by any specific purpose
- An example clustering question would be: “Do our customers form natural groups or segments?”
- Clustering also is used as input to decision-making processes focusing on questions such as: What products should we offer or develop? How should our customer care teams (or sales teams) be structured?

Co-Occurance Grouping

- Co-occurrence grouping (also known as frequent itemset mining, association rule discovery, and market-basket analysis) attempts to find associations between enti- ties based on transactions involving them.
- An example clustering question would be: “What items are commonly purchased together?”
- For example, analyzing purchase records from a supermarket.
- recommendation systems

Profiling

- Profiling attempts to characterize the typical behavior of an individual, group, or population.
- An example clustering question would be: “: “What is the typical cell phone usage of this customer segment?”
- Profiling is often used establish behavioral norms for anomaly detection applications.
  - Fraud detection and monitoring for intrusions to computer system.

Link Prediction

- Link prediction attempts to predict connections between data items, usually by suggesting that a link should exist, and possibly also estimating the strength of the link.
- Link prediction is common in social networking systems: “Since you and Ka- ren share 10 friends, maybe you’d like to be Karen’s friend?”

Data Reduction

- Data reduction attempts to take a large set of data and replace it with a smaller set of data that contains much of the important information in the larger set.
- For example, a massive dataset on con- sumer movie-viewing preferences may be reduced to a much smaller dataset revealing the customer taste preferences.

Causal Modeling

- Causal modeling attempts to help us understand what events or actions actually influence others.
- For example, consider that we use predictive modeling to target advertisements to consumers, and we observe that indeed the targeted consumers purchase at a higher rate subsequent to having been targeted. Was this because the advertisements influenced the consumers to purchase? Or did the predictive models simply do a good job of identifying those consumers who would have purchased anyway?

Supervised vs Unsupervised Methods

- Unsupervised: no specific purpose or target specified.
  - Do our customers naturally fall into different group?
- Supervised: specific target defined.
  - “Can we find groups of customers who have particularly high likelihoods of canceling their service soon after their contracts expire?”

Supervised vs Unsupervised Methods

- So, if the target can be provided, the problem can be phrased as supervised one.
- Supervised task require different techniques than unsupervised tasks do, and the results often are much more useful.

Supervised vs Unsupervised Methods

- Supervised data mining: there must be data on the target.
- An example:

  Customer ID . . . . . Loyal

  1345 . . . . . Yes

  1234 . . . . . No
- Acquiring data on the target is a key data science investment.

Supervised vs Unsupervised Methods

- Supervised data mining: there must be data on the target.
- The value for the target vaiable for an individual is often called the individual's label. So customer 1345's label is YES.

Supervised vs Unsupervised Methods

- Two main subclasses pf supervised data mining:
  - Classification
    - Will this customer purchase service S1 if given incentive I?”
    - “Which service package (S1, S2, or none) will a customer likely purchase if given incen- tive I?”
    - What is the probability that a customer will continue to subscribe to the service?” class probability estimation.
  - Regression
    - “How much will this customer use the service”

Supervised vs Unsupervised Methods

- A vital part in the early stages of the data mining process
  - To decide whether the line of attacj will be supervised or unsupervised.
  - If supervised, to produce a precise definition of a target variable, this variable must be specific quantity that will be the focus of the data mining.

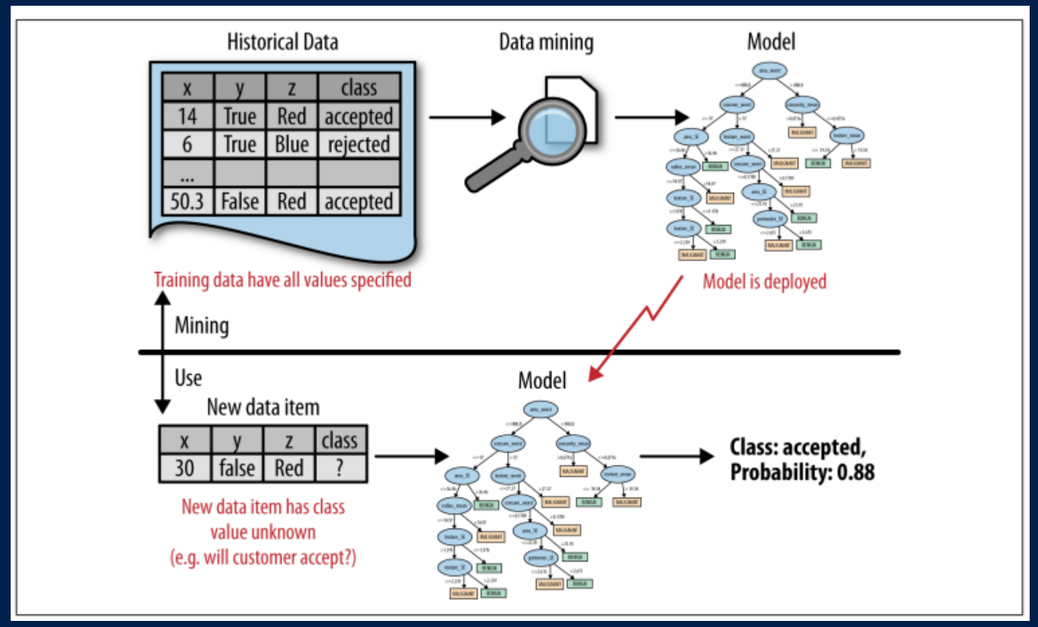
Data Mining and Its Results

- distinction pertaining to mining data:
    - mining the data to find patterns and build models
    - using the results of data mining

Figure 1. Data mining versus the use of data mining results.

Data Mining and Its Results

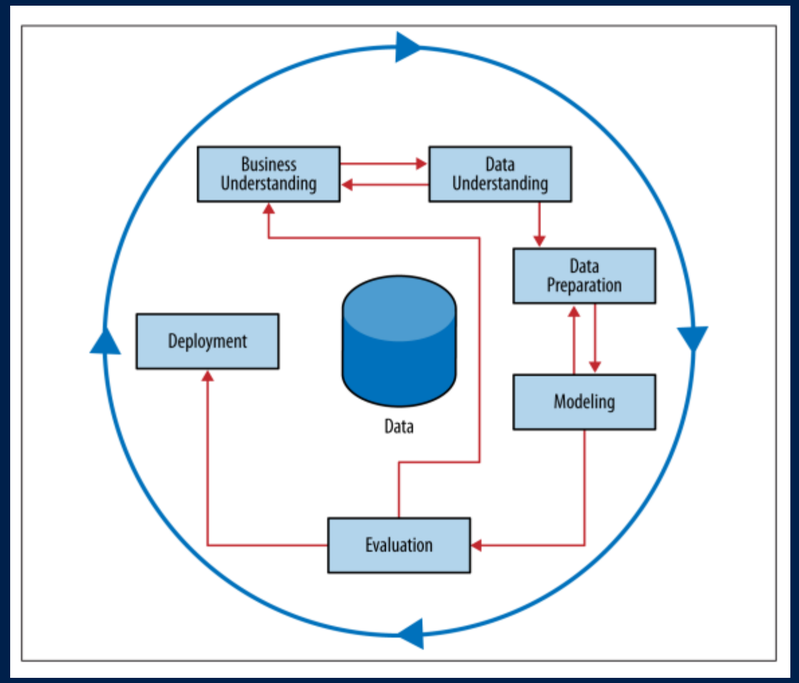
Figure 1. Data mining versus the use of data mining results.



The upper half of the figure illustrates the mining of historical data to produce a model. Importantly, the his- torical data have the target ("class") value specified. The bottom half shows the result of the data mining in use, where the model is applied to new data for which we do not know the class value. The model predicts both the class value and the probability that the class variable will take on that value.

The Data Mining Process

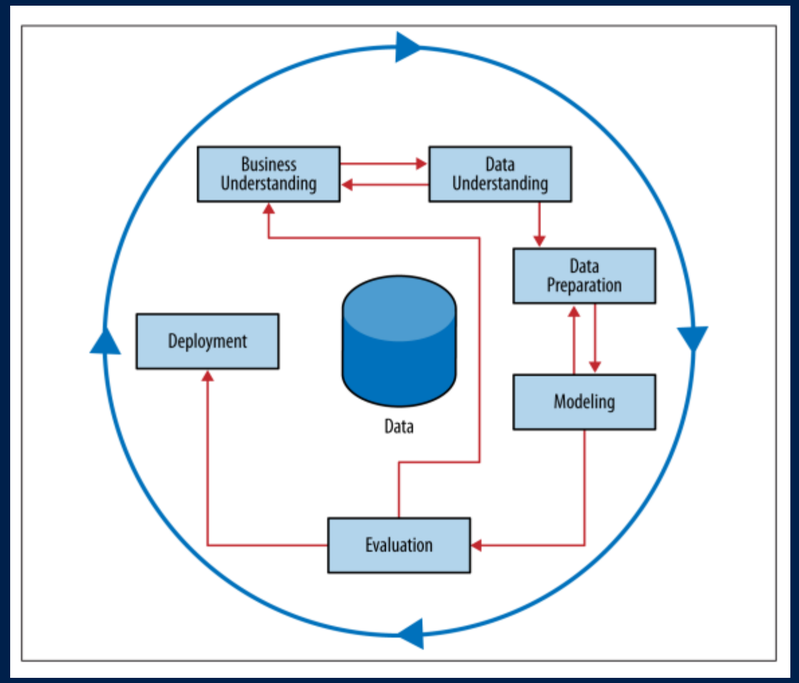
Figure 2. The Cross Industry Standard Process (CRISP) for Data mining process



This process diagram makes explicit the fact that iteration is the rule rather than the exception. Going through the process once without having solved the problem is, generally speaking, not a failure. Often the entire process is an exploration of the data, and after the first iteration the data science team knows much more. The next iteration can be much more well-informed. Let's now discuss the steps in detail.

The Data Mining Process

Figure 2. The Cross Industry Standard Process (CRISP) for Data mining process



This process diagram makes explicit the fact that iteration is the rule rather than the exception. Going through the process once without having solved the problem is, generally speaking, not a failure. Often the entire process is an exploration of the data, and after the first iteration the data science team knows much more. The next iteration can be much more well-informed. Let's now discuss the steps in detail.

Business Understanding

- It is vital to understand the problem to be solved.
- The Business Understanding stage represents a part of the craft where the analysts' creativity plays a large role.
- The design team should think carefully about the use scenario.

Data Understanding

- The data comprise the available raw material from which the solution will be built.
- Estimating the costs and benefit of each data source and deciding whether further investment is merited.
- Ex: Credit card fraud & Medicare fraud.

Data Preparation

- Often proceeds along with with data understanding.
- Converting data to tabular format.
- Removing or inferring missing values.
- Converting data to different types.
- Leaks

  A leak is a situation where a variable collected in historical data gives information on the target variable—information that appears in historical data but is not actually available when the decision has to be made. As an example, when predicting whether at a particular point in time a website visitor would end her session or continue surfing to another page, the variable "total number of webpages visited in the session" is predictive.

Modelling

- Output of modeling is some of model or pattern capturing regularities in the data.

Evaluation

- assess the data mining results rigorously and to gain confidence that they are valid and reliable before moving on.
- Includes both quantitative and qualitative assessment. Stakeholders should check and see whether the model is going to do more good than harm.
- Data science team must consider the comprehensibility of the model to the stakeholders.

Deployment

- Put into real use in order to realize some return on investment.
- The clearest cases of deployment involve implementing a predictive model in some information system or business process.
- Churn example: a model for predicting the likelihood of churn could be integrated with the business process for churn management. for example, by sending special offers to customers who are predicted to be particularly at risk.

Implications for Managing the Data Science Team

- It is tempting – but usually a mistake – to view the data mining process as a software development cycle.
- Software skills versus analytics skills.

Implications for Managing the Data Science Team

- It is tempting – but usually a mistake – to view the data mining process as a software development cycle.
- Software skills versus analytics skills.

Statistics

Two different uses in business analytics.

- First, it is used as a catchall term for the computation of particular numeric values of interest from data
- denote the field of study that goes by that name
  - Help us understand different data distribution
  - Help us understand how to use data to test hypothesis and to estimate the uncertainty of conclusions.
  - hypothesis testing can help determine whether an observed pattern is likely to be a valid, general regularity as opposed to a chance occurrence in some particular dataset

Other Analytics Techniques and Technologies

- Present six group of related group techniques.
- Comparisons and contrasts with data mining.
- Data mining => automated search for knowledge, patterns, or regularities from data.
- Business analyst => to recognize what sort of of analytic technique is appropriate for addressing a particular problem.

Data Querying

- A query is a specific request for a subset of data or for statistics about data, formulated in a technical language and posed to a database system.
- Differ fundamentally from data mining in that there is no discovery of patterns and models.
- Ex:

```
SELECT * FROM CUSTOMERS WHERE AGE > 45 and SEX='M' and DOMICILE = 'NE'
```

Data Warehousing

- Data warehouses collect and coalesce data from across an enterprise, often from multiple transaction-processing systems, each with its own database.
- For example, if a data warehouse integrates records from sales and billing as well as from human resources, it can be used to find characteristic patterns of effective salespeople.

Regression Analysis

This will involve estimating or predicting values for cases that are not in the analyzed data set.

Machine Learning and Data Mining

- The collection of methods for extracting (predictive) models from data, now known as machine learning methods, were developed in several fields contemporaneously, most notably Machine Learning, Applied Statistics, and Pattern Recognition.
- Machine Learning as a field of study arose as a subfield of Artificial Intelligence, which was concerned with methods for improving the knowledge or performance of an intelligent agent over time, in response to the agent's experience in the world.

Institut Informatika & Bisnis DARMAJAYA
Yayasan Alfian Husin

SAINS DATA
Institut Informatika & Bisnis Darmajaya

Kampus Merdeka INDONESIA JAYA

MERDEKA BELAJAR

THANK YOU!!

DATA SCIENCE DARMAJAYA “YOUR BEST FUTURE IN DATA”