


Data Analysis – Data Mining

November 2022



Outline

- **Introduction**
 - **Business understanding**
 - **Data understanding**
 - **Data preparation**
 - **Modelling**
 - **Evaluation**
 - **Deployment**
- 



Introduction





Introduction

CertDA Learning Outcomes

- Know how to use commercial awareness to articulate business questions
- identify and manipulate relevant data and deeply analyse it by applying appropriate techniques
- know how findings from analysis can and should be visualised and communicated, enabling relevant stakeholders to make sound business decisions
- learn and understand ethical security issues around data analytics
- be introduced to popular statistical and programming tools such as SQL, R and Python, as well as an introduction to artificial intelligence and machine learning.



Introduction

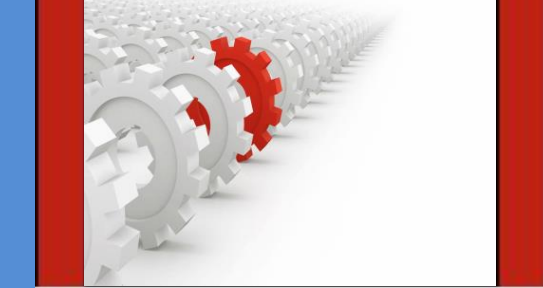
What is Data Mining?

- Data mining is the process of identifying relationships, trends and patterns in large sets of data, effectively turning raw data into useful information.
- Data mining approaches involve various methods such as statistics, machine learning, and database systems.
- The information obtained through the data mining process can then be further processed and used to support decision-making.

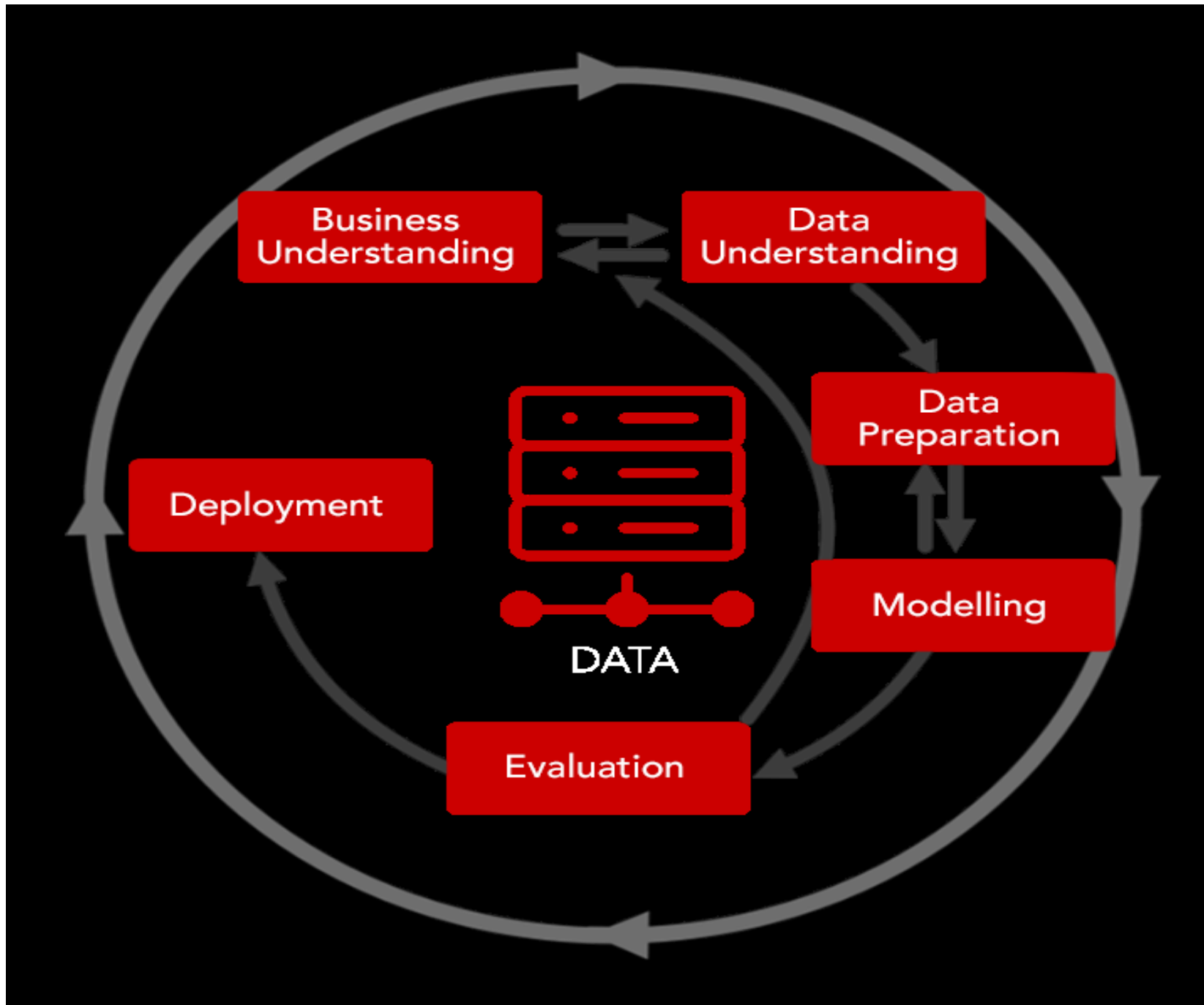


The CRISP-DM Framework

- CRISP-DM is a cross-industry process for data mining and is a process model designed to facilitate a structured approach to data mining.
- It was first conceived in 1996, and in 1997 it became an official European Union project under the ESPRIT funding initiative.
- The project was spear-headed by some companies: Integral Solutions Ltd (ISL), Teradata, Daimler AG, NCR; its methodology being published as a data mining guide in 1999.



- Recent research indicates that CRISP-DM is the most widely used data-mining process, because of its various advantages which solved the existing problems in the data mining industries.
- The apparent success and wide use of the CRISP-DM is that it is industry, tool, and application neutral.





- The process model is composed of six distinct but connected phases which represent the ideal sequence of activities involved in the data mining process.
- In practice some of these activities may be performed in a different order.
- Some of the paths between activities are two-way, indicating that it will frequently be necessary to return to earlier steps depending on the outcome of a particular activity.



Business understanding

- Business understanding is the essential and mandatory first phase in any data mining or data analytics project.
- It involves identifying and describing the fundamental aims of the project from a business perspective.
- This may involve solving a key business problem or exploring a particular business opportunity.



Such problems might be:

- Establishing whether the business has been performing or under-performing and in which areas
- Monitoring and controlling performance against targets or budgets
- Identifying areas where efficiency and effectiveness in business processes can be improved
- Understanding customer behaviour to identify trends, patterns and relationships
- Predicting sales volumes at given prices
- Detecting and preventing fraud more easily
- Using scarce resources most profitably
- Optimising sales or profits.



- Having identified the aims of the project to address the business problem or opportunity, the next step is to establish a set of project objectives and requirements.
- These are then used to inform the development of a project plan. The plan will detail the steps to be performed over the course of the rest of the project and should cover the following:
 1. Deciding which data needs to be selected from internal or external sources
 2. Acquiring suitable data



3. Determining criteria to determine whether or not the project will have been a success.
4. Developing an understanding of the acquired data
5. Cleaning and preparing the data for modelling
6. Selecting suitable tools and techniques for modelling
7. Creating appropriate models from the data
8. Evaluating the created models
9. Visualising the information obtained from the data
10. Implementing a solution or proposal that achieves the original business objective.





Data understanding

- The second phase of the CRISP-DM process involves obtaining and exploring the data identified as part of the previous phase and has three separate steps, each resulting in the production of a report.

Data
Acquisition

Data
Description

Data
Exploration



Data Acquisition

- This step involves retrieving the data from their respective sources and the production of a data acquisition report that lists the sources of data, along with their provenance, the tools or techniques used to acquire them.
- It should also document any issues which arose during the acquisition along with the relevant solutions.
- This report will facilitate the replication of the data acquisition process if the project is repeated in the future.



Data Description

- For quantitative data, this should include descriptive statistics such as minimum and maximum values as well as their mean and median and other statistical measures.
- For qualitative data, the summary data should include the number of distinct values, known as the cardinality of data, and how many instances of each value exists.



Data Description

- For quantitative data, this should include descriptive statistics such as minimum and maximum values as well as their mean and median and other statistical measures.
- For qualitative data, the summary data should include the number of distinct values, known as the cardinality of data, and how many instances of each value exists.



- The first step is to describe the raw data.
- For instance, if analysing a purchases ledger, you would at this stage produce counts of the number of transactions for each department and cost centre, the minimum, mean and maximum for amounts, etc.
- Relationships between variables are examined in the data exploration phase (eg. by calculating correlation).
- For both types of data, the report should also detail the number of missing or invalid values in each of the attributes.



- If there are multiple sources of data, the report should state on which common attributes these sources will be joined.
- Finally, the report should include a statement as to whether the data acquired is complete and satisfies the requirements outlined during the business understanding phase.



Data Exploration

- This step builds on the data description and involves using statistical and visualisation techniques to develop a deeper understanding of the data and their suitability for the analysis.
- These may include:
 - Performing basic aggregations
 - Studying the distribution of data; either through producing descriptive statistics such as means, medians and standard deviations or by plotting histograms



- Examining the relationships between pairs of attributes; eg. by correlation for numeric data using regression analysis or chi-square testing.
- Exploring the distribution and relationships in significant subsets of the data
- These exploratory data analysis techniques can help provide an indication on the likely outcome of the analysis and may uncover patterns in the data that may be worth subjecting to further examination.
- The results of the exploratory data analysis should be presented as part of a **data exploration report** that should also detail any initial findings.



Data Preparation





- As with the data exploration phase, the data preparation phase is composed of multiple steps and is about ensuring that the correct data is used in the correct form in order for the data analytics model to work effectively.

**Data
Selection**

**Data
Cleaning**

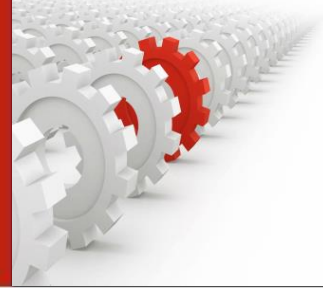
**Data
Integration**

**Feature
Engineering**



Data Selection

- The first step in data preparation is to determine the data that will be used in the analysis.
- This decision will be informed by the reports produced in the data understanding phase but may also be based on the relevance of particular datasets or attributes to the objectives of the data mining project, as well as the capabilities of the tools and systems used to build analytical models.
- There are two distinct types of data selection, both of which may be used as part of this step.



Feature selection is the process of eliminating features or variables which exhibit little predictive value or those that are highly correlated with others and retaining those that are the most relevant to the process of building analytical models such as:



- Multiple linear regression, where the correlation between multiple independent variables and the dependent variable is used to model the relationship between them.
- Decision trees, simulating human approaches to solving problems by dividing the set of predictors into smaller and smaller subsets and associating an outcome with each one.
- Neural networks, a naïve simulation of multiple interconnected brain cells that can be configured to learn and recognise patterns.



- Sampling may be needed if the amount of data exceeds the capabilities of the tools or systems used to build the model.
- This normally involves retaining a random selection of rows as a predetermined percentage of the total number of rows.
- Often, surprisingly small samples can give reasonably reliable information about the wider population of data, such as obtained from voter exit polls in local and national elections.
- Any decisions taken during this step should be documented, along with a description of the reasons for eliminating non-significant variables or selecting samples of data from a wider population of such data



Data Cleaning

- Data cleaning is the process of ensuring the data can be used effectively in the analytical model.
- The next step is to process missing and erroneous data identified during the data understanding or collection phase.
- Erroneous data, values outside of reasonably expected ranges, are generally set as missing.



- Missing values in each feature are then replaced either using simple rules of thumb, such as setting them to be equal to the mean or median of data in the feature or by building models that represent the patterns of missing data and using those models to "predict" the missing values.
- Other data cleaning tasks include transforming dates into a common format and removing non-alphanumeric characters from text.
- The activities undertaken, and decisions made during this step should be documented in a data cleaning report.



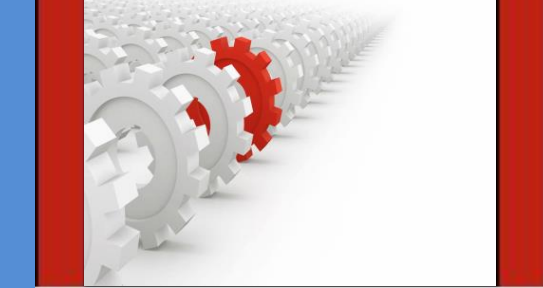
Data Integration

- Data mining algorithms expect a single source of data to be organised into rows and columns.
- If multiple sources of data are to be used in the analysis, it is necessary to combine them.
- This involves using common features in each data set to join the datasets together.
- For example, a dataset of customer details may be combined with records of their purchases.
- The resulting joined data set will have one row for each purchase containing attributes of the purchase combined with attributes related to the customer.



Feature Engineering

- This optional step involves the creation or inclusion of new variables or derived attributes into the existing variables or features originally included to improve the model's capability.
- This step is frequently performed when the data analyst feels that the derived attribute or new feature or variable is likely to make a positive contribution to the modelling process and where it involves a complex relationship that the model is unlikely to infer by itself.



An example of a derived feature might be adding such attributes such as the amount a customer spends on different products in a given time period, how soon they pay and how often they return goods to more reliably assess the profitability of that customer, rather than just measure the gross profit generated by the customer based on sales values.



Modelling





- This key part of the data mining process involves creating generalised, concise representations of the data. These are frequently mathematical in nature and are used later to generate predictions from new, previously unseen data.

Determine the modelling techniques to be used

- The first step in creating models is to choose the modelling techniques which are the most appropriate, given both the nature of the analysis and of the data used. Many modelling methods make assumptions about the nature of data. For examples, some methods can perform well in the presence of missing data whereas others will fail to produce a valid model.



Design a Testing Strategy

- Before proceeding to build a data analytics model, you will need to determine how you are going to assess the quality of predictive ability of the model.
- This is done using data specially held aside for this purpose, in other words, how well the model will perform on data it hasn't yet seen.
- This involves using a subset of data kept aside for this purpose and using it to evaluate how far off the model's predictions of the dependent variable are from the actual values in the data.



Evaluation





At this stage in the project, you need to verify and document that the results you have obtained from modelling have the veracity (are reliable enough) for you to prove or reject your hypotheses in the business understanding stage.

EXAMPLE

- If you have performed a multiple regression analysis on predicting sales based on weather patterns, are you sure that the results you have obtained are statistically significant enough for you to implement the solution, or have you checked that there are no other intermediate variables linked to the X, Y variables in your relationship which are a more direct causal link?



- Before proceeding to final deployment of the model, it is important to thoroughly evaluate it and review the steps executed to create it; to be certain the model properly achieves the business objectives.
- A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.



- At this stage, you will determine if it is feasible to move on to the final phase deployment, or whether it is preferable to return to and refine some of the earlier steps.
- The outcome of this phase should be a document providing an overview of the evaluation and details of the final decision together with a supporting rationale for proceeding



Deployment



- During this final phase, the outcome of the evaluation will be used to establish a timetable and strategy for the deployment of the data mining models, detailing the required steps and how they should be implemented.
- Data mining projects are rarely "set it and forget it" in nature. At this time, you will need to develop a comprehensive plan for the monitoring of the deployed models as well as their future maintenance.
- This should take the form of a detailed document.



- Once the project has been completed there should be a final written report, re-stating and re-affirming the project objectives, identifying the deliverables, providing a summary of the results and identifying any problems encountered and how they were dealt with.
- Depending on the requirements, the deployment phase can be as simple as generating a report and presenting it to the sponsors or as complex as implementing a repeatable data mining process across the enterprise. In many cases, it is the customer, not the data analyst, who carries out the deployment steps.



- However, even if the analyst does carry out the deployment, it is important for the customer to clearly understand which actions need to be carried out in order to actually make use of the created models.
- This is where data visualisation is most important as the data analyst hands over the findings from the modelling to the sponsor or the end user and these should be presented and communicated in a form which is easily understood.



PIVOT TABLE



The background of the slide features a large, faint, 3D-rendered gear that spans across the middle and right sections. The gear is light gray and has a complex, multi-toothed design. The overall aesthetic is clean and professional, with a focus on mechanical imagery.




MICROSOFT POWER BI






Questions

- What happened ?
 - What is happening ?
 - Why did it happen ?
 - What will happen?
 - What do I want to happen ?
- 



Power BI vs data warehouse

- Extract
 - Transform
 - Load
- 



What you can get

- Introduction of Power BI & Concepts
 - Power BI connection with different sources and its implications
 - Use of Power Query for Data transformation
 - Creating Models in Power BI
 - Data Analysis expression
 - Reports & Visualisation
- 

Clipboard: Paste, Cut, Copy, Format painter

Data: Get data, Excel workbook, Power BI datasets, SQL Server, Enter data, Dataverse, Recent sources, Transform data, Refresh

Insert: New visual, Text box, More visuals

Calculations: New measure, Quick measure


Sensitivity: Sensitivity

Share: Publish


Manage and connect to recent sources.

Add data to your report


Once loaded, your data will appear in the Fields pane.




Import data from Excel



Import data from SQL Server



Paste data into a blank table

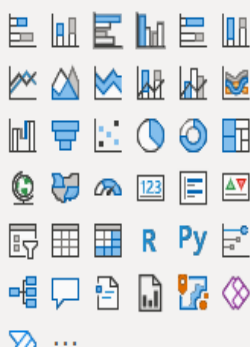


Try a sample dataset

[Get data from another source →](#)

Visualizations
Fields

Build visual



Filters

Values

Add data fields here

Drill through

Cross-report Off

Keep all filters On

Add drill-through fields here

You haven't loaded any data yet. [Get data](#)

