Data Science Process

Areej Alasiry

Lecture Outline

- Prior Knowledge
- Data Preparation
- Modelling
- Application
- Posterior Knowledge

Data Science Process

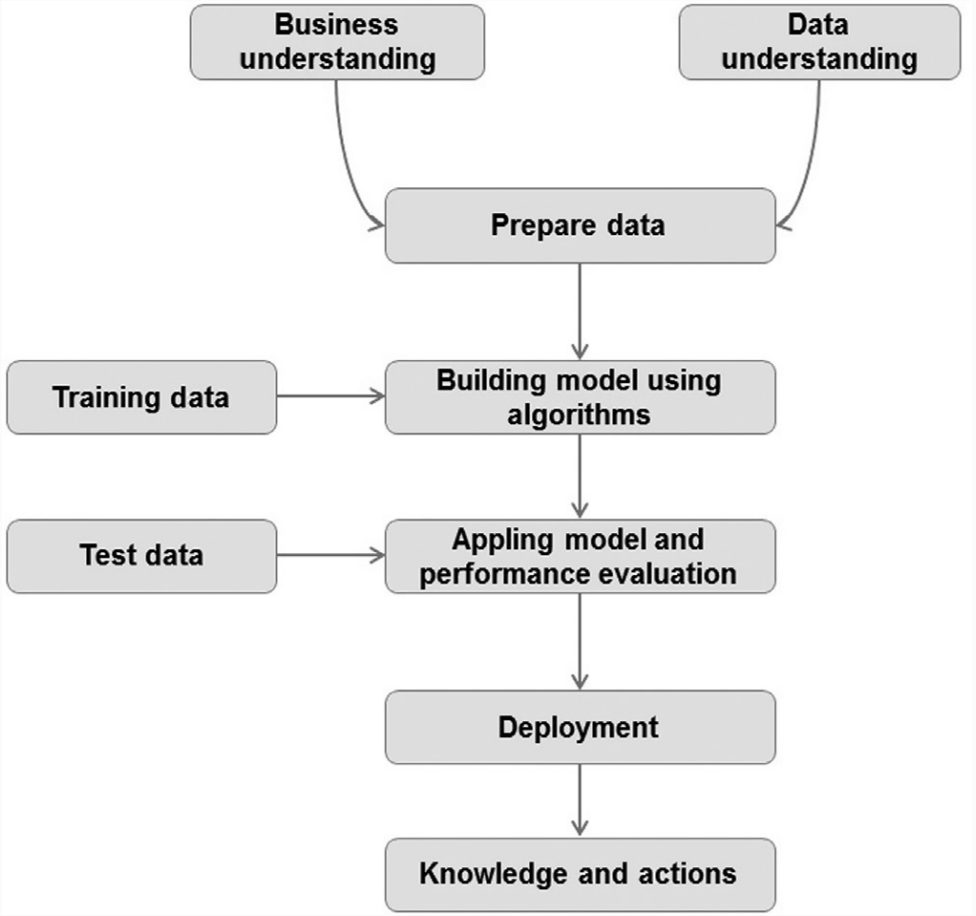
Five Main Stages

- Understand the problem
- Prepare the dataset
- Develop the model
- Apply the model on the dataset
- Deploy and maintain the model

Various Data Science Frameworks

- Cross Industry Standard Process for Data Mining (CRISP-DM)
- SEMMA = Sample, Explore, Modify, Model, and Assess
- DMAIC = Define, Measure, Analyze, Improve, and Control.
- Selection, Preprocessing, Transformation, Data Mining, Interpretation, and Evaluation

Data Science Process



From:
Data Science: Concepts and Practice. Figure 2.2 (2019)

1
Prior Knowledge

Prior Knowledge

- State the data science process objective.
- Understand the subject area
- Understanding how the data is collected, stored, transformed, reported, and used
  - Dataset
  - Data point
  - Attribute
  - Label
  - Identifier
- Causation Vs Correlation

2
Data Preparation

Data Preparation

- Exploratory Data Analysis (EDA)
  - Descriptive Statistics
  - Visual Analysis
- Data Quality
  - Data Cleansing
    - Duplicate Records
    - Outliers
    - Missing Values
  - How to handle missing values? Why are they missing?
    - Substitute with artificial values.
    - Ignored

Data Preparation (Cont.)

- Data Types and Conversion
  - Different data science algorithms impose different restrictions on the attribute data types.
- Transformation
  - Normalisation
- Outliers
- Feature Selection
  - Reducing the number of attributes, without significant loss in the performance of the model

Data Preparation (Cont.)

- Data Sampling
  - The process of selecting a subset of the dataset as a representation of the dataset.
  - Speeds up the model building process
  - Training and Test dataset split.
  - Stratified Sampling
  - Ensemble model.

3
Modelling

Modelling

- Model is “an abstract representation of the data and the relationships in a given dataset”.
- Predictive Models
  - Classification
  - Regression

- Descriptive Models
  - Association Analysis
  - Clustering

Modelling (Cont.)

- Training and Test Dataset
  - Training dataset: used to create the model
  - Test/Validation dataset: used to evaluate the model
  - How to split the dataset?

- Learning Algorithms
  - The problem statement dictate what model to use.

Modelling (Cont.)

- Model Evaluation
  - Overfitting
  - Prediction Error

- Ensemble Modelling:
  - Multiple various models used to predict an outcome.
  - Reduce the generalisation error.

Modelling (Cont.)

- At this stage:
  1. Analyze the business question
  2. Source the data relevant to answer the question
  3. Select a data science technique to answer the question
  4. Pick a data science algorithm and prepare the data to suit the algorithm
  5. Split the data into training and test datasets
  6. Build a generalized model from the training dataset
  7. Validate the model against the test dataset

4
Application

Application

- Deployment of the model.
- It involves the following:
  - Assess the model readiness
  - Technical Integration.
  - Response Time
  - Model Maintenance.
  - Assimilation.

5
Knowledge

Knowledge

- To extract knowledge from data is an art and can be developed with practice
  - Massive Dataset
  - Choosing the right algorithms
  - Reporting results
  - Statistical Analysis

Session Mind Map

References

- Data Science: Concepts and Practice. 2nd Edition by Vijay Kotu and Bala Deshpande, ISBN: 978-0-12-814761-0©2019. [Chapter 2]