



Business Intelligence and Analytics

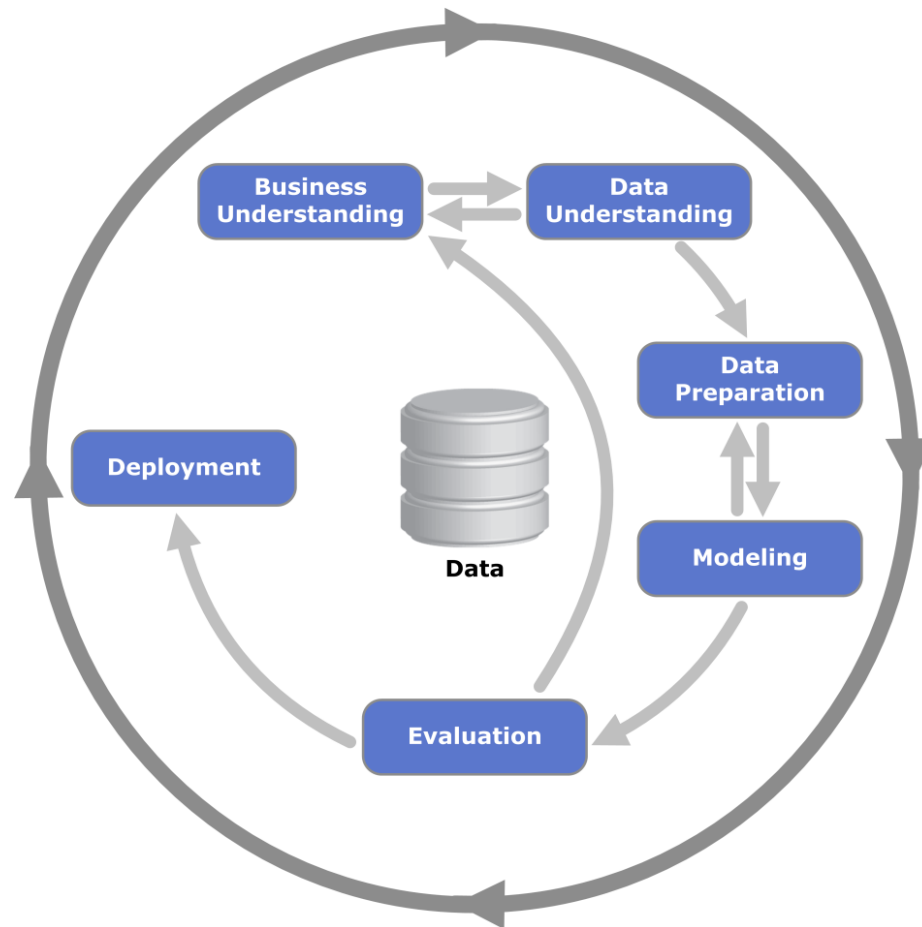
(Data Mining)

Project Specifications

Ph.D. Ettore Ritacco



CRISP-DM Methodology





Instructions

- Form a group of 3 students (exceptions are allowed, but must be reported)
- Pick up a project theme from the following list:



Project list 1 / 3

- UK Road Safety: Traffic Accidents and Vehicles
 - <http://www.kaggle.com/tsiaras/uk-road-safety-accidents-and-vehicles>
- Stack Overflow 2018 Developer Survey
 - <http://www.kaggle.com/stackoverflow/stack-overflow-2018-developer-survey>
- Credit Card Fraud Detection
 - <http://www.kaggle.com/mlg-ulb/creditcardfraud>



Project list 2/3

- Smart meters in London
 - <http://www.kaggle.com/jeanmidev/smart-meters-in-london>
- Los Angeles Parking Citations
 - <http://www.kaggle.com/cityofLA/los-angeles-parking-citations>
- Austin Animal Center Shelter Intakes and Outcomes
 - <http://www.kaggle.com/aaronschlegel/austin-animal-center-shelter-intakes-and-outcomes>



Project list 3/3

- MyAnimeList Dataset
 - <http://www.kaggle.com/azathoth42/myanimelist>
- Heart Disease and Stroke Prevention
 - <http://www.kaggle.com/mazharkarimi/heart-disease-and-stroke-prevention>
- Avocado Prices
 - <http://www.kaggle.com/neuromusic/avocado-prices>
- Star Wars Movie Scripts
 - <http://www.kaggle.com/xvivancos/star-wars-movie-scripts>



Instructions

- Choose a business and mining goal about the theme you have chosen
- Analyze the related dataset according to what you learnt during the course, trying to achieve your goals
 - You can use any kind of analysis tool the best fits your needs
- Produce 2 documents:
 - A CRISP-DM documentation (.doc, .docx or .pdf), **max 15 pages**
 - A Presentation (.ppt, .pptx or .pdf)
- The maximum score you can achieve in this phase is 25/30
- If the project will be approved, you will have the oral proof



Documentation - Example

Business Understanding

Determine Business Objectives

Background
Business Objectives
Business Success
Criteria

Situation Assessment

Inventory of Resources
Requirements,
Assumptions, and
Constraints
Risks and
Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goal

Data Mining Goals
Data Mining Success
Criteria

Produce Project Plan

Project Plan
Initial Assessment of
Tools and Techniques

Data Understanding

Collect Initial Data

Initial Data Collection
Report

Describe Data

Data Description Report

Explore Data

Data Exploration Report

Verify Data Quality

Data Quality Report

Data Preparation

Data Set

Data Set Description

Select Data

Rationale for Inclusion /
Exclusion

Clean Data

Data Cleaning Report

Construct Data

Derived Attributes
Generated Records

Integrate Data

Merged Data

Format Data

Reformatted Data

Modeling

Select Modeling Technique

Modeling Technique
Modeling Assumptions

Generate Test Design

Test Design

Build Model

Parameter Settings
Models
Model Description

Assess Model

Model Assessment
Revised Parameter
Settings

Evaluation

Evaluate Results

Assessment of Data
Mining Results w.r.t.
Business Success
Criteria
Approved Models

Review Process

Review of Process

Determine Next Steps

List of Possible Actions
Decision



Presentation

- The presentation is a summary of what you did during the dataset analysis (**once again you can take inspiration from the CRISP-DM methodology**)
 - Don't forget any step
- It must last 15 minutes at most
- Focus on your analysis