**Tutorial: Data Modeling**

1. There are several proposals about what the intelligent data analysis process should look like. CRoss Industry Standard Process for Data mining (CRISPDM) is one of them. Describe the components of the process and the relationships among them. Also discuss questions associated with each of the components.

CRIPS (Cross-Industry standard process for data mining, known as Crisp-DM), is an opens standard process model that describes common approaches used by data mining experts. It is the most widely-used analytics mode.

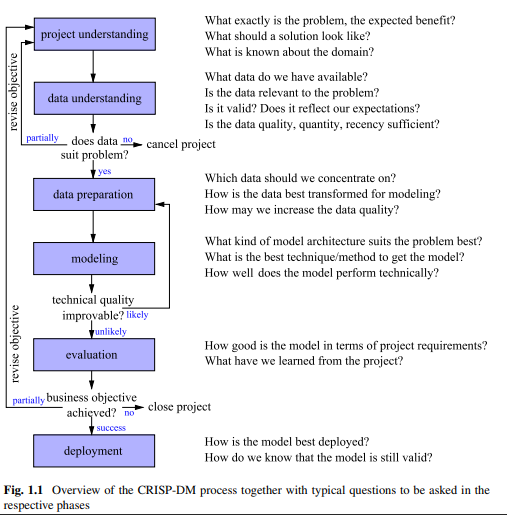
CRISP-DM consists of 6 phases (Fig.1), that are usually repeated more than once; The main objective of the first project understanding step is to identify the potential benefit as well as the risk and efforts of a successful project, such that a deliberate decision on conducting the full project can be made. In this step, we basically convert the project in a more technical notion and this step is usually called business understanding; Next we need to make sure that we have ll data to solve the problem and here, we are in the data understanding phase; In this step, we analyze data, see if they are clear and understand the relationship between the attributes; If data are not clear, go back to the check the project bjectives;

Then we have data preparation, where data is selected, corrected and even new attributes are generated; here, in this step, all the deficits seen in the understanding data step are fixed;

Depends on what we want to do with the model, we may choose techniques that are easily interpretable or less demonstrative black-box models, which may perform better. Background knowledge may provide hints on useful transformation that simplify the representation of the solution.

When technical benchmarks cannot be improved anymore, the obtained results are analyzed in the **evaluation phase** from the perspective of the problem owner. At this point, the project may stop de to unsatisfactory results, the objectives may be revised to succeed under a slightly different setting, or the found and optimized model may be deployed. After deployment, wich ranges from wriing a report to the creation of a software system that applied th model automatically to aid or make decision, the project is not necessary finished.

Once data is ready, apply modeling tools (this in another step), to extract knowledge out of data In form of a model. Depending on what we want to do with the model, we may vchoose techniques that are



2. There are mainly four method categories for data analysis problems: classification, regression, clustering and association. Explain and discuss each of them. What are differences among them? Examples are expected to support your discussion.

Each category heko ti categorize the large number of different tools and algorithms that oslve a specific task;

* **Classification:** Predict the outcome of an experiment with a finite number of possible results (like yes/no or unacceptable/acceptable/good/very good)
* **Regression**: Regression is, just like classification, also a prediction task, but this time the value of interest is numerical in nature. Typical questions: How will the EUR/USD exchange rate develop?
* **Clustering, segmentation:** Summarize the data to get a better overview by forming groups of similar cases (called clusters or segments). Instead of examining a large number of similar records, we need to inspect the group summary only. We may also obtain some insight into the structure of the whole data set. Cases that do not belong to any group may be considered as abnormal or outliers. Typical questions: Do my customers divide into different groups?
* **Association analysis”** Find any correlations or associations to better understand or describe the interdependencies of all the attributes. The focus is on relationships between all attributes rather than focusing on a single target variable or the cases (full record). Typical questions: Which optional equipment of a car often goes together? How do the various qualities influence each other?

3. Most of data analysis methods share the same procedure. a. Select the Model Class b. Select the Score Function c. Apply the algorithm d. Validate the Results Explain and discuss the purpose and main issues of the four steps (what are they? What are the purpose and main issues).

After we have gone thought the phases of project and data understanding, we have to option: have a working model or go back to the project understanding;

Most of the data analysis methods can be viewed withing 4 stress:

* **Select Model Class:** we must specify the general structure of the analysis result; We call this the architecture or “model class”; In a regression problem it could be considered the linear function or a quadratic function for example
* **Select the score function:** used to find the best mode; . In the case of the simple linear regression function, our score function will tell us which specific choice of the coefficients is better when we compare different linear functions
* **Apply the Algorithm:** The score function enables us to compare models, but it does not tell us how to find the model that obtains the best evaluation from the score function. Therefore, an algorithm is required that solves this optimization problem
* **Validate the Results:** Even if our optimization algorithm is so cleverly designed that it will find the best model in the model class we consider, we still do not know, whether this is the best model among very good models for our data or just the best one among even worse choices. Therefore, we still need other means than the score function to validate our model

4. Simplicity of the model is one of the important aspects. What are the reasons?

. Simpler models will be preferred for various reasons:

• They are usually easier to understand and to be interpreted.

• Their computational complexity is normally lower.

• A model can also be a summary of the data. A complex model might not be a summary but, in the worst case, just more or less a one-to-one representation of the data. General structures in the data can only be discovered by a model when it summarizes the data to a certain extent. The problem that too complex models often fail to reveal the general relations and structures in the data is called overfitting

5. A perfect fit with zero error is suspicious in most of cases. Once we have fitted a model to given data, the fitting error can be composed into 4 components: a. the experimental error; b. the sample error; c. the model error; d. the algorithm error. Explain and discuss each of them.

a model with a smaller error is not necessarily a better fit for the data. In general, more complex models can fit the data better but have a higher tendency to show this bad effect called overfitting. Once we have fitted a model to given data, the fitting error can be composed into four components. • The pure or experimental error, • the sample error, • the lack of fit or model error, and • the algorithmic error

* Experimental Error: The pure error or experimental error is inherent in the data and is due to noise, random variations, imprecise measurements, or the influence of hidden variables that cannot be observed. It is also called intrinsic error.
* Bayes Error
* Sample error: The sample error is caused by the fact that the data is only an imperfect representation of the underlying distribution of the data. A finite sample, especially when its size is quite
* Model Error: A large error may be caused by a high pure error, but it may also be due to a lack of fit. When the set of considered models is too simple for the structure inherent in the data, no model will yield a small error. Such an error is also called model error.

6. Describe the cross-validation method. Discuss in which situation it should be employed.

* Cross validation
  + Standard tool in analytics and an important feature for helping me develop a fine-rune data mining models.
  + Cross validation has the following applications
    - Validating he robustness of a particular mining model
    - Evaluating multiple models from a single statement.

Cross-validation consists of two phases, training and result generation. These phases include the following steps:

* You select a target mining structure.
* You specify the models you want to test. This step is optional; you can test just the mining structure as well.
* You specify the parameters for testing the trained models.
  + The predictable attribute, predicted value, and accuracy threshold.
  + The number of folds into which to partition the structure or model data.
* Analysis Services creates and trains as many models as there are folds.
* Analysis Services returns a set of accuracy metrics for each fold in each model, or for the data set as a whole.