Overview

The data mining process starts with problem understanding with a (business) objective. It is possibly the most crucial step in a data mining process. Without a well-defined and precise problem statement, it is impossible to develop the correct data set and determine suitable data mining methods(s). Furthermore, even though the data mining process is sequential, it is common to revisit previous steps and revise the assumptions, techniques, and tactics as we can observe in Figure 1 which diagrammatically presents the general process flow of Data Mining. Therefore, it is imperative to get the objective of the whole process right, even if it is exploratory data mining.

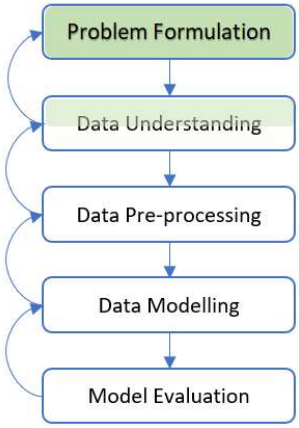

Figure 1 A General Data Mining Process and Phases

The week's topic covers mainly the first phase (i.e., Problem Formulation), and partially the second phase (i.e., Data Understanding) of the data mining process.
To formulate a problem(s) suitably using the Data-Mining approach, clearly understanding the problem's characteristics and the data for the solution is essential. Therefore, we can use several questions as a guiding checklist to gain the understanding for formulating the problem.

This week's topic focuses on establishing these questions as guidelines for understanding the problem(s) and data available for the solution. Then, we will use a mini case study to apply the questions as a checklist for problem formulation.

2.1 Problem Formulation

Asking the right questions for any data-mining project, gaining in-depth domain understanding, sourcing, preparing the data, and assessing implementation feasibilities are crucial to the success of the data mining process.

It is rare to see data mining problems defined in formal performance requirements, though it might be sensible for long-term system development efforts. Instead, the requirements are usually expressed in terms of performance metrics for the processes they facilitate. For example, it is common to see project goals: "Use customer transaction data to help operations managers to decrease churn by 30%." However, the means to realize the goals are usually left unstated.

The objective of data mining does not appear in isolation; it always develops from an existing subject matter and contextual information that is already known. Prior knowledge refers to information that is known about a subject. The prior knowledge helps in the data mining process define what problem we are solving, how it fits in the project context, and what data we need to use for the solution.

A data mining project is likely in trouble if it does not begin with focused discussions with domain experts who have prior knowledge of the subject, customers, and system users. Questions during the discussions should be addressed to the right person. For instance, a data-mining practitioner does not ask a business person detailed technical questions. Although not all questions apply to every data mining project, a project undertaking might begin with a short checklist.

A checklist can assist in problem formulation through problem understanding and assessing the feasibility of using the Data Mining approach.

2.1.2 Checklist for Problem Understanding

As discussed previously, a project undertaking might begin with a short checklist, although not all questions apply to every data mining project. A checklist is necessarily generic. As a data-mining practitioner develops experience, he/she will know which questions to ask and will add more specifics to the problem domain. This checklist serves as a reminder about the essential questions to ask.
To understand problems to solve and the feasibility of using a data mining approach, we can use the following checklist in the form of questions as guiding measures.

Checklist 1: What Problem are we addressing?

Success is unlikely without a clear understanding of the project's objective. Also, customers or users might not clearly understand what data mining is and possibly have unreasonable expectations.
The answer should be a working definition of success for the project. In addition, it should list project deliverables and be able to quantify measurements of the achievement of the objective(s).
If different users give divergent answers, the user organisation probably does not know what it wants.

Ask for requirements, specifically, what are the measures of performance or how to measure if the objective(s) is met? Get answers to these questions from the individual(s) evaluating the project. Their expectations are definitive.
For example, we might get to this question: "We want something that predicts whether a customer is about to churn or close their account."
The first step of a data mining process is to ask appropriate questions and make actionable suggestions for meeting the objective(s).

Checklist 2: Describe the ideal solution for the problem.

Usually, this question is not asked directly. However, customer expectations must be thoroughly comprehended. Presumably, the ideal solution will satisfy these.

The first response to this question is often vague. Press for details, specifically about the measures of performance.

Follow-up questions ask about the business case for the project. Someone in the management chain expects a return on this effort; what are the specifics of those expectations in business terms?

Special considerations Often, there is some unsolved problem or must-have feature that drives the decision to undertake a data mining project.

Try to determine what that is. In our customer churn example, the expectation is to build a model that uses the customer churn data to help operations managers to decrease churn by 30% as a measure of performance.

Checklist 3: What is the characteristic of the desired solution (associator, descriptor, classifier, or estimator)?

The user likely does not understand the difference between the functions of these different solutions. Asking this question enables the data-mining practitioner to work through the decision with them.

This question is seeking the choice of data mining solution and topology since classifiers are designed and built differently from estimators, for example.
In the customer churn example, the characteristic of the desired solution is a classifier, a data model that can classify a customer's class, either likely to churn or not.

Checklist 4: What Type of Data Mining Modeling Must Perform?

This question complements Checklist 3. Data mining processes work on data to provide two basic types of conclusive support: Firstly, they allow the user to obtain domain understanding, and secondly, they enable appropriate user action. Underlying these processes are various types of machine reasoning, each executed by a particular type of modeling. Furthermore, each model has a topology driven by

the type of modeling it must perform. Therefore, it is essential to decide what type of reasoning it will have to do before deciding what type of model to build.
This question needs to be answered precisely to ensure the data mining effort is to succeed. The followings describe the common types of models for particular problems.

Classifiers: Classifiers take a list of attributes and decide into which of many categories the entity (i.e., target class) is exhibiting these attributes falls. Automatic target class recognition and future event prediction are examples of a classifier.
Estimators: Estimators take a list of attributes and assign some numeric value to the target class demonstrating these attributes. Estimating a probability is an example of an estimator.

Associators or Descriptors: This type of model samples the entire corpus of domain data and identifies relationships or associations among entities. The outcome of the modeling is descriptive, not predictive. Data clustering to identify coherent and apparent subpopulations is a simple example. Another example is the market-basket analysis of items purchased and transaction records to infer the association of customers' preferences according to their buying patterns.

Checklist 5. What makes this problem challenging?

Data mining is not the first thing most organizations attempt when they have a complex problem. They have likely attempted other approaches; to answer this question to avoid repeating a past failure.
This question seeks to understand what has and has not worked. Consequently, this finding can provide plenty of information about the nature of a problem.
Usually, answers from users are too general and vague since previous efforts using other approaches might not be recent or fully comprehended. Follow-up questions ask about talking to those with domain knowledge and working on a problem or the data before.

Checklist 6: What is the data domain (time, frequency, both, image, text, web, other)?

This information is necessary to identify and extract the data as features or variables for the data mining process. Metadata is usually available, though it is usually not entirely updated.

Follow-up questions ask the user to extract the data and put it into an understandable and workable format for further investigation. A data-mining practitioner should ask and understand how the data extraction was done and know precisely the sources of the data received. It is critical to know the data types and relationships to extract and encode variables for analysis. If the data are in a database, a database schema and metadata show how the data are stored and linked should be available for gaining knowledge about the data.

Users may always think that accessing their data is easier than it is. Nevertheless, unfortunately, they also often overestimate the data quality, integrity, completeness, and usefulness for data mining.

More questions related to data understanding are discussed in Section 2.1.

Checklist 7: Describe the solution domain's base classes or range of estimates.

This information is the expected principal data output of the model selected in a data mining process. For example, questions like "How many classes is the expected classifier modeling? What are they? What are the units, precision, and range of estimator outputs?

Users with domain knowledge usually understand these subjects since they have probably had to manually produce and evaluate such outputs.

2.2 Data Understanding

The data understanding phase involves looking at the data available for mining. This phase is critical in avoiding unexpected problems during the next phase, i.e., the data preparation, typically the most prolonged phase of a data mining project.

There are two possible methods we can use to understand data:

1. Use guiding questions to understand the data with the input of domain experts or persons with background knowledge of the data.

   This method provides vital knowledge of the data sources, the meaning of data fields (i.e., attributes), the expected data type and value ranges, how the data is collected and stored, the appropriate data size, and the interrelationship between the data.

   We will learn several guiding questions as a checklist to understand data in the next section.
2. Use software tool(s) to explore data characteristics.

   This method is usually called data exploration. Previously we obtained prior knowledge about the datasets by asking guiding questions stated in the method above. Then, software tool(s)can further analyze the characteristics of the datasets used to solve the problem(s), especially data quality and distribution.

   We will study the data exploration method in Week 3.

2.3 A Mini Case study

In the following sections, we use a mini case study, a customer churn, as an example applying the guiding questions in practice to understand the problem(s) and the example data.

2.3.1 Checklist Example for Problem Understanding

This section uses the customer churn case study, as an example for applying the guiding questions in practice to understand the problem.

Checklist 1. What problem are we addressing?

For this case study, we want to use a data-mining approach to meet the following objective:
"We want an automated model that will categorize our clients as falling into one of two classes so that appropriate customer retention measures can be taken timely."

The two classes, for example, can be:

- Class one - the client is likely to churn (as the target positive class because the objective of this problem solving is to identify those customers likely to churn).
- Class two - the client is unlikely to churn (as the target negative class).

Based on the definition derived from domain knowledge (for example), a customer is labeled as 'yes' for likely to churn if he/she fulfilled either of the following conditions:

- The customer has terminated the service.
- The customer has been recorded as churned.

Checklist 2. Describe the ideal solution for the problem.

In our customer churn example, the expectation is to build a model that uses the customer demographics and transaction data to help decrease churn by 25% as a performance measure. Furthermore, the accuracy of the selected classifier must achieve at least 70% of accuracy.

Checklist 3. What is the characteristic of the desired solution (associator, descriptor, classifier, or estimator)?

In the customer churn example, the characteristic of the desired solution is a classifier, a data model that can classify a customer's class, either likely to churn or not.

Checklist 4. What Type of Data Mining Modeling Must Perform?

For this case study, the classifiers will take a list of attributes and decide on a predicted class (i.e., target class value, either likely (positive) or not (negative) to churn).

Checklist 5. What makes this problem challenging?

The challenges of this case study are that we cannot use simple data query methods such as SQL (Structured Query Language) or run statistics or machine learning as modeling techniques for the following reasons:

- Existing raw data sets do not have a complete target class attribute (i.e., churn) to recognize if a customer is likely to churn or not in a future event. Therefore, it is necessary to understand the problem objective and available data and preprocess them to derive this target class attribute from existing available data.
- Customer demographics and payment transaction data are stored in different files. Therefore, selecting relevant attributes to predict churn target classes (which requires domain knowledge) and combining them from different source files requires understanding their meaning and interrelationships.
- Selecting the best model as the classifier to produce the most accurate output requires an iterative modeling process using different techniques, evaluating and comparing their performance.
- Assuming the selected model as the best classifier requires a re-run analysis every two years, data understanding and preprocessing and modeling will need to be re-run to suit the most recent available data to reflect the current customer churn characteristics.

Therefore, Data mining can be an appropriate approach to solving the problem with the complexity involved.

Checklist 6. What is the data domain (video, image, text, web, other)?

The available datasets are all having a textual data domain.

Checklist 7. Describe the solution domain's base classes or range of estimates.

The data output of the models as classifiers are either "yes" (as the positive target class because the objective of this problem-solving is to identify those customers likely to churn) or "no" (as the negative target class).

The value of "yes" representing a customer is likely to churn, whereas "no" for unlikely to churn. Therefore, there are only two possible classes the expected classifier modeling should produce.

2.3.2 Checklist Example for Data Understanding

This section uses the customer churn case study as an example for applying the guiding questions in practice to understand the data.

Checklist 1. What constitutes a data collection, and how many collections are there?

There are three data sources available for mining.
The following basic metadata in Figure 2 presents the data characteristics of source name, attribute names, data type domain including data type and expected values, and a brief description of each attribute.

Source File: **CUSTOMER** (in .xlsx file format)
Source Description: Customer demographics information

Attribute Name	Data Type	Expected Values	Description
CUSTOMERID*	NUMBERS	Any number starts from 1	Customer ID
FIRSTNAME	CHARACTERS	Any characters	Customer's first name
POSTALCODE	CHARACTERS	Any combination of characters or numbers	Customer's postal code
HASHCODE	CHARACTERS	Any combination of characters or numbers	System generated hash-code
BIRTHDATE	DATE	Date	Customer's birthdate
GENDER	CHARACTERS	female, male	Indication of the gender type

Source File: **CHURN** (in .xlsx file format)
Source Description: A list of customer IDs who has terminated service

Attribute Name	Data Type	Expected Values	Description
CUSTOMERID*	NUMBERS	Any number starts from 1	Customer ID who has churned

Source File: **PTRANSACTION** (in .xlsx file format)
Source Description: Customer's payment transaction aggregated details

Attribute Name	Data Type	Expected Values	Description
CUSTOMERID*	NUMBERS	Any number starts from 1	Customer ID who has churned
LASTTRANSACTION	DATE	Date	The last payment transaction of a customer
MINTRXVALUE	DECIMAL	Any numbers from 0	The lowest payment amount paid by a customer
MAXTRXVALUE	DECIMAL	Any numbers from 0	The lowest payment amount paid by a customer
TOTALTRXVALUE	DECIMAL	Any numbers from 0	The total payment amount paid by a customer
CASH	NUMBERS	Binary number 0 or 1	If a customer paid by cash before, then 1 else 0
CHEQUE	NUMBERS	Binary number 0 or 1	The number of times a customer paid by cheque
CREDITCARD	NUMBERS	Binary number 0 or 1	The number of times a customer paid by credit
SINCELASTTRX	NUMBERS	Any numbers from 0	Number of days since the last payment date

* Unique key of a dataset

Figure 2 Metadata of The Customer Churn Datasets

Checklist 2. What constitutes a data sample?

The internal structure of the data samples that constitute the various attributes, a.k.a. data fields, and their data types (i.e., formats) can be derived from the metadata which is described in Figure 2.

Checklist 3. What is the data size?

For this case study example, to estimate the data size for mining, the largest dataset is the CUSTOMER source file containing 1000 records, collected from March 2008 till the end of February 2022.

Checklist 4. How is this data currently stored, and how to get it?

For this case study as an example to demonstrate the possibility of having different types of source file formats, the datasets are all available to be accessed via any platforms or software tools that can support XLSX and CVS files.

Checklist 5. What are the baseline sample rates and sizes (samples per second, bits per sample)?

To know the data formats and structure, we can refer to the metadata described in Figure 2.

The data collection rate from a periodicity perspective is roughly 15 years from March 2008 till the end of February 2022. For the case study example, we consider the classification only needs one-shot modeling for this period.

Checklist 6. What is the duration/rate of the sampling activity that produced the data?

The timeframe of data collected is from March 2008 till the end of February 2022.

Checklist 7. Were collections taken under varying settings or conditions?

The original datasets were collected in Germany. For the case study example, we assume the data samples represent all the possible scenarios, and data collection is done at once.

Checklist 8. What is the source of this data?

The original datasets were collected in Germany. The data source information can be obtained from the metadata description in Figure 2.

Checklist 9. What does this data field mean?

The data field (a.k.a. attribute) information can be obtained from the metadata description in Figure 2.

Checklist 10. What are the possible values for this data field (nominal, numerical, ordinal)?

The possible values for the data field (a.k.a. attribute) information can be obtained from the metadata description in Figure 2.

Checklist 11. Why is this data field collected?

The data fields collected are used as the possible input as predictors and the target for predicting a classification.

Checklist 12. Is this data field related to other data fields?

The data field of CUSTOMERID in each source file represents the same meaning. Therefore, it can be used as the common data field to interlink between datasets to derive a more complete view of customer demographics information, payment transaction details, and churn indication statuses.

Checklist 13. Does the data field always have a value? What does it mean when the data field is missing? What to do when this data field is empty?

The possible values for the data field (a.k.a. attribute) and context information can be obtained from the metadata description in Figure 2.