Data Mining

V. Kumar, W. Reinartz, *Customer Relationship Management*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-662-55381-7_7

Overview

The way in which companies interact with their customers has changed dramatically over the past few years. Customers' expectations have risen, and it is becoming increasingly difficult to satisfy them. Customers have access to an array of alternative products to choose from and their loyalty is difficult to gain. At the same time, companies need to retain the profitable customers to succeed in a competitive and dynamic marketplace. As a result, companies have found they need to understand their customers better, and to respond to their wants and needs faster. The time frame in which these responses need to be made has been shrinking. More customers, more products, more competitors, and less time to react means understanding the customers is now much harder to do.

To succeed, companies must be proactive and anticipate customer desires. Many firms have realized this and are collecting information about their customers and their preferences. Firms collect, store, and process vast amounts of highly detailed information about customers, markets, products, and processes through different programs. Data mining this information gives businesses the ability to make knowledge-driven strategic business decisions to help predict future trends and behaviors and create new opportunities. Data mining can assist in selecting the right target customers or in identifying (previously unknown) customer segments with similar behaviors and needs.

This chapter describes the importance and benefits of data mining and gives a detailed overview of the underlying process. The data mining procedure breaks down into five subsections: defining the business objectives, getting the raw data, identifying relevant variables, gaining customer insight, and acting. The discussion of these steps will help the reader understand the overall process of data mining. The process steps are illustrated with the case study of Credite Est (name disguised), a French mid-tier bank. Finally, the case study, «Yapi Kredi—Predictive Model–Based cross-sell Campaign,» shows a comprehensive application of data mining.

7.1 The Need for Data Mining

Today, most companies do not suffer from lack of data about their customers, products, transactions, and markets. To the contrary, *data deluge* is a problem in many companies. This is especially challenging for information-intensive businesses, such as banking, telecommunication, and e-commerce, where large amounts of data can easily be recorded. The sheer amount of raw data is, for many, an obstacle to using it for extracting knowledge and for making critical business decisions. By default, educated guessing becomes the primary decision-making tool. It does not have to be that way.

Availability of computers and mass storage, statistical and data analysis methods, sophisticated reporting platforms, and online touch points with customers, now give companies access to a powerful asset: information. Data have become a company's most important—and in many cases, untapped—asset. To extract customer intelligence and value from that data, companies must implement a standardized data mining procedure. A successful data mining infrastructure consists of technology, human skills, and tight integration with enterprise operations to allow transforming new knowledge into business action and value. It is important to standardize the data mining process to assure the required quality of results, make it a repeatable process, better maintain and keep the knowledge inside the company, as well as training new employees more quickly.

7.2 The Business Value of Data Mining

In the context of customer management, data mining can help to gain a better understanding of customers and their needs. Marketing is still frequently associated only with creative and soft skills. But by scientifically enhancing targeting, we can obtain more impressive cost reductions and revenue growth than by working only with the creative aspects of marketing. Data mining can assist in selecting the right target custom-

ers or in identifying (previously unknown) customer segments with similar behaviors and needs. A good target list developed by using data mining techniques is likely to increase purchase rates and have a positive impact on revenue.

Applications of data mining include the following:

- Reducing churn with the help of predictive models (see Summary), which enable early identification of those customers likely to stop doing business with your company.
- Increasing customer profitability by identifying customers with a high growth potential.
- Reducing marketing costs by more selective targeting.

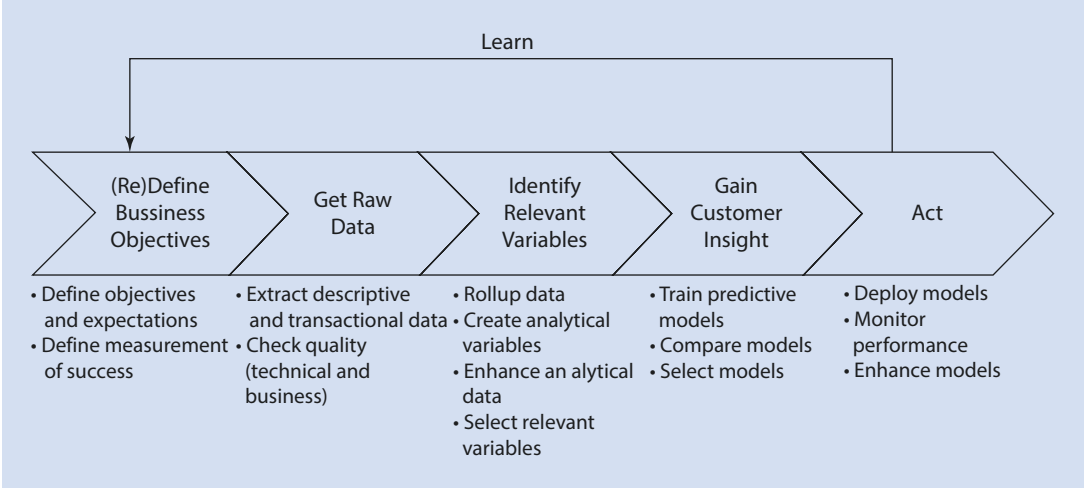
This chapter introduces a systematic approach to data mining projects.

7.3 The Data Mining Process

A complete data mining process does not *only* consist of building analytical models using techniques such as logistic regression (see ▶ Sect. 6.2.3). It includes assessing and specifying the business objectives, data sourcing, transformation, creation of analytical variables, selecting relevant variables, training predictive models, selecting the best suited model, and acting on the basis of the findings. These activities can be grouped into five process steps of defining the business objectives, getting raw data, indentifying relevant variables, gaining customer insight, and acting. ◘ Figure 7.1 presents an overview of the data mining process.

In many instances of current data mining projects we find data preparation steps easily take from 60% to 70% of the total project time. This is not due to the weaknesses of any specific data mining methodology. It is due to issues regarding unavailability of relevant variables describing customer behavior. An example of which is the difficult access to legacy data source systems managed by different departments which do not possess the customer centric views required for data mining projects. These departments are more likely geared towards transaction, product, contract, or other type of views more suited to fulfill the needs of their current operational systems. The graph shown in ◘ Fig. 7.2 helps to understand the timeframe of the individual steps of the data mining process.

It is important to automate the time-consuming data extraction and manipulation, as well as the data quality monitoring and enhancement steps. To achieve this goal, it is necessary to sequentially and systematically code data knowledge into programs that can be executed, for instance, in batch mode. This will free up time of highly qualified quantitative data analysts (data miners) to concentrate on the value-generating tasks such as the precise definition of business objectives, extraction of customer insight, and effective actions based on gained knowledge.



◘ **Fig. 7.1** Overview of the data mining process

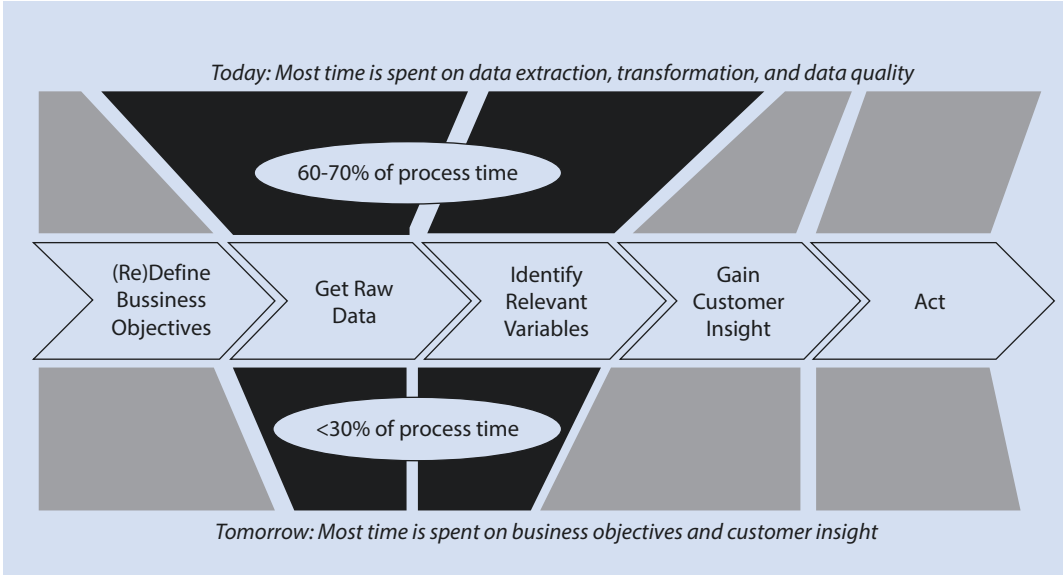


Fig. 7.2 Allocation of time for the steps in the data mining process

7.3.1 Involvement of Resources

Usually, different functional departments are involved in data mining projects. The main groups are the business group (e.g., marketing, product management), data mining, and IT. The business group is primarily involved in defining business objectives, and takes the lead when it comes to deploying the new insights into corporate action. The data mining group must understand the business objectives and support the business group in refining and sometimes correcting the scope of the project and aligning their expectations to fit the limitations posed by the available data. The data mining group is most active during the variable selection and modeling phase. It will share the obtained customer insights with the business group, who are strongly involved at this point to check the plausibility and soundness of the solution in business terms. IT resources are required for the sourcing and extraction of the required data used for modeling. Figure 7.3 shows the extent of involvement of the three main groups participating in a data mining project—business, data mining, and IT—during the different process steps.

7.3.2 Data Manipulation

As we move through the data mining process, the dimensionality of the data used may change dramatically. In a simple, two-dimensional data table we think of the columns as being the descriptive variables and the rows as being single observations, each pertaining to a collection of variables about the same primary object (e.g., customer identification number, transaction identification number).

Manipulations on columns can take several forms:

- **Transformation:** Transform birth date to age.
- **Derivation:** Create new variables based on existing ones (e.g., compute monthly profits from sales and cost information).
- **Elimination:** A whole variable may be excluded from further processing due to a variety of possible reasons (e.g., a variable that does not help in predicting or a variable that is correlated to one or more variables already in the model could be eliminated).

The number of variables used changes drastically during the data mining process. Figure 7.4 illustrates a typical example of the changes in the number of variables used at each step.

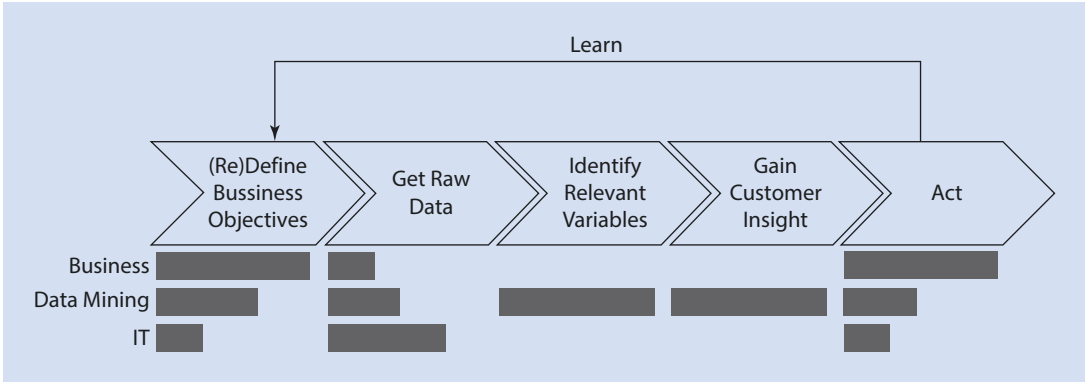


Fig. 7.3 Level of involvement of business, data mining, and IT resources in a typical data mining project

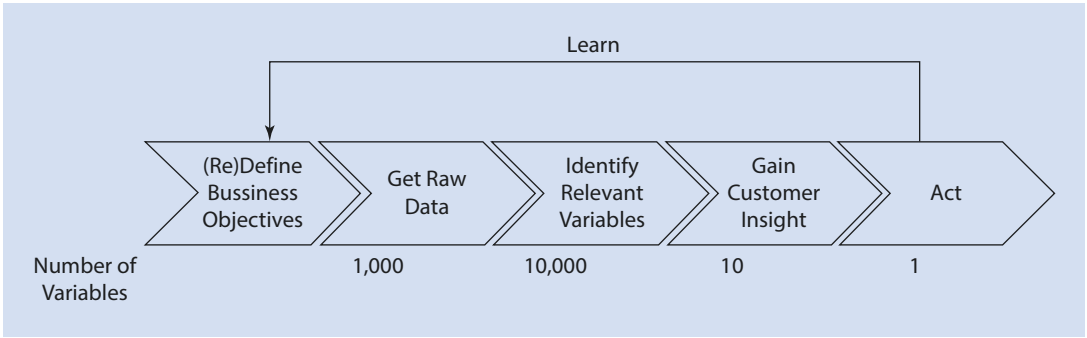


Fig. 7.4 Number of variables at different process steps

If we also take into consideration that we usually work with up to several millions of rows it becomes obvious that scalability and good sampling methods are essential for any data mining environment.

There are also several types of row manipulations, the most common ones can be classified into:

- **Aggregation:** Examples include counts, mean, and standard deviation of the number of transactions of a specific type over a given time period, for a specific customer, product type and many more.
- **Change detection:** This is used to detect when and if certain variables change their value such as ZIP code of customers domicile, or her credit rating.
- **Missing value detection:** It is common that raw data come with many data fields either totally missing or with some missing values. The reasons for this may be nonmandatory input fields in front-office systems, incomplete data migration from one system to another, and so on. There are various ways of treating a variable with missing values, including eliminating the whole row from further processing when a missing value is detected, replacing the missing value with a constant value, or replacing it with a randomly generated value based on the distribution of this data field's nonmissing values or based on correlations with: other data fields.[1]
- **Outlier detection:** In some cases observations may contain variables with extreme values, meaning values far away from the bulk of other values for the distinct variable. Sometimes these outliers are real; sometimes

1 An example would be the expectation maximization algorithm (EM) that takes into account the correlation of the data field for which a nonmissing value is to be generated with other data fields.

they are the consequence of bad data quality. Outlier detection is in its simple form univariate: we just look at one variable and try to find values that stand out. In its more sophisticated form we look at many variables at the same time and watch out for multivariate outliers (a data point might look like an outlier in the univariate case but not in the multivariate case). Outliers can be mapped to other values or the corresponding rows and be excluded from further processing.

When preparing the data for modeling, it is common to sample and split the incoming data into various streams for different purposes:

- **Train set:** Used to build the models.
- **Test set:** Used for out-of-sample tests of the model quality and to select the final model candidate.
- **Scoring data:** Used for model-based prediction. Typically, this data set is large as compared to the previous ones.

The data sets must be carefully examined and designed to assure statistical significance of the results obtained.

7.3.3 Define Business Objectives

Data mining finds application in many situations. Profitable customer acquisition requires modeling of expected customer potential, in order to target the acquisition of those customers who will be profitable over the lifetime of the business relationship (they might be unprofitable in the beginning and turn into very profitable customers later—e.g., a medicine student). In a cross-selling or up-selling model, we model the customer's affinity with a set of products or services translated into her purchase likelihood. In churn management, it is crucial to correctly model a customer's likelihood to defect based on past behavior. Some applications require predicting not only who will purchase which product or service, but also the expected amount spent on the transactions (Fig. 7.5).

Once it has been identified which customer behavior has to be predicted, we need to mathematically define this target variable (dependent variable). For example, while up-selling platinum credit cards to customers already owning a standard credit card, we might find several types of standard and platinum cards exist. The business objective might, in fact, only be to up-sell platinum cards of types P2 and P3 to customers not owning a card yet, or those already owning standard cards of type S1, S3, and S4. The target variable has to reflect these conditions when calculated. This has a later direct impact on the modeling process.

To prepare the modeling data sets we will look for two types of customers in the customer

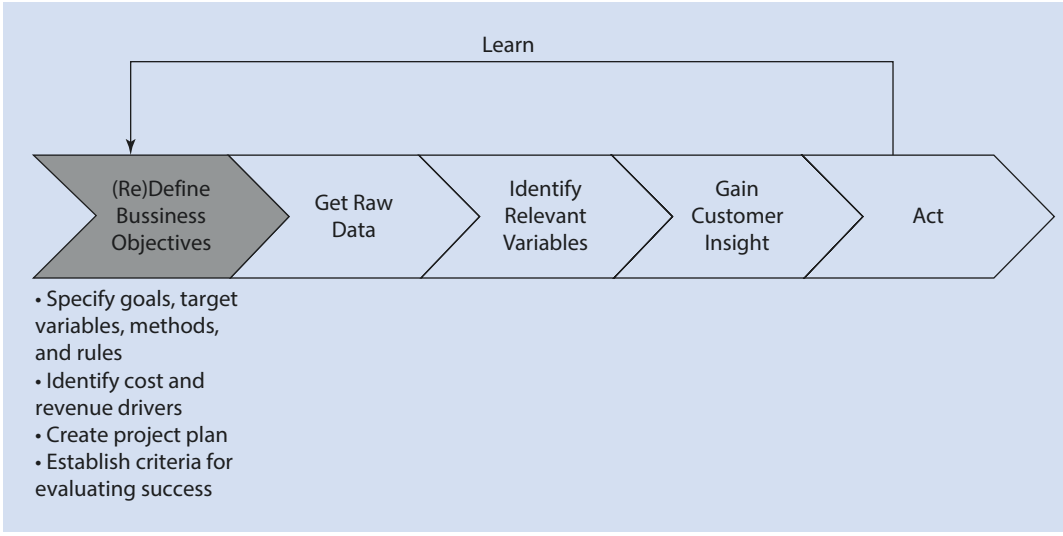


Fig. 7.5 Data mining process: define business objectives

database and set the value of the target variable accordingly:

- For all customers who (first purchased a standard card of type S1, S3 or S4 and then purchased a platinum card of type P2 or P3) or (purchased a platinum card of type P2 or P3 immediately) set target variable = 1.
- For all other customers set target variable = 0.
- Once these restrictions and considerations have been applied to the data, the model will be trained to distinguish between customers with a target variable equal to zero and customers with a target variable equal to one (for example using logistic regression, see ▶ Sect. 6.2.3) After training, the model will be applied to predict if a given customer is likely to buy the platinum cards. The business group must establish likelihood threshold levels above which they think a prospect should be included in the marketing campaign.

Another aspect of the campaign which should be defined during this project phase is the set of business or selection rules for a campaign. Rules define the customers that should be excluded from or included in the target groups: certain products or services might not be available for specific customer groups. Suppose that in certain countries only customers over 18 years are eligible to purchase a credit card. Credit products in general have restrictions with respect to the customer's credit rating. Companies have «black-lists» containing customers who should not receive any new offerings, either due to bad debt indicators (they do not have a good credit rating) or persons explicitly asking not to be contacted for marketing purposes. Some countries have centrally managed lists with all persons not wanting to receive unauthorized direct mails or calls. By contrast, there might also be customer groups that should be included at any price in the campaign—for instance, due to strategic issues such as need to gain market share in a specific region or otherwise defined group. In those cases, it is not relevant if members of that group get high model scores; they are included anyway.

To ensure a successful project, we need to define at this point the details of its execution. Therefore, we should create a project plan specifying, for instance, the start and delivery dates of the data mining process, as well as the responsible resources for each task. For the final model selection the business group must be available for reviewing the data mining results, perform consistency checks with the data mining group, and make the final decision about the model selected for deployment. Delivery dates for the final model or scores also need to be defined, along with dates for start and end of the supported campaign.

We need to carefully define the chosen experimental setup for the campaign; this is critical for correctly evaluating its success later. It is highly recommended to spend a significant amount of time getting this right. Usually, we split the target group into various cells. In the simplest case there will be only two cells:

1. **The control group contains only randomly selected customers:** This group will be needed to measure the baseline effect (i.e., what would have been the normal customer behavior without the influence of the campaign).
2. **The other cell will contain only the best customers according to the model used:** This simple setup allows measuring how the model-based selection is doing with respect to the average customer behavior.

More refined setups may generate more than two cells. As an example, take the control group and two target groups to which we communicate different content about the same offering during the campaign. This will show the impact of the communication content on the purchase behavior.

To get the business context into the data mining project, describe the nature of the business involving the data mining project, and the cost and revenue drivers of the business. This knowledge will affect the final model and target group selections. It is helpful to define a cost/revenue matrix describing how the business mechanics will work in the supported campaign and how it will affect the data mining process. As an example, consider a call center campaign to sell a mobile phone contract. ◘ Table 7.1 shows an example of the associated cost/revenue matrix.

Here, we assume the average cost per call is $5. Each positive responder (purchaser) will generate additional cost including administration work required to register him as a new customer and the cost of the delivered phone handset of, say, $100. Customers who respond positively will generate average revenue of $1000 a year. Putting all

Table 7.1 Cost/revenue matrix

Cost/revenue matrix	Prospect did not purchase		Prospect did purchase	
Model predicts prospect will not purchase (not contacted)	Cost:	$0	Lost business opportunity of + $895	
	First year revenue:	$0		
	Total:	$0		
Model predicts prospect will purchase (contacted)	Cost:	−$5	Cost:	−$105
	First year revenue:	$0	First year revenue:	+$1000
	Total:	−$5	Total:	+$895

these factors together defines the cost/revenue decision matrix, which will subsequently have an impact, on the choice of model parameters such as the cut-off point for the selected model scores. It will also give business users an immediately interpretable table.

Finally, we need to establish the criteria for evaluating the success of the campaign. This is a key aspect for the success or failure of the whole project. Often there is a misunderstanding on both sides—business and modelers—with respect to what is feasible from a business and statistical perspective. Clearly defining the expectations helps. If, for example, gaining market share is more important than obtaining high purchase rates, the measure of success changes from «percentage of sold cards per contacted customer» to «absolute number of cards sold during the campaign.» In this context, it is crucial to specify how the campaign results will be tracked and analyzed. Depending on the type of business, how it is structured into customer segments, regions, products, and others, we could be interested in measuring purchase rates per region, per sales channels, per product type, and so on as a function of time. When we consider a situation where we have defined various target groups for different communication contents or product offers in one campaign, it is worth measuring the purchase rates for each group.

Sometimes it is useful to look for a benchmark to compare results obtained in the past for the same or similar campaign setups using traditional targeting methods and not predictive models. We have to be careful when comparing the result of the old and new methods because there could be hidden differences due to different business (market) conditions, changes in the products or services, etc. In this chapter, we examine the French company Credite Est, a regional bank that implemented a data mining process. The following example looks at how Credite Est defined its business objectives.

CRM at Work 7.1

Defining Business Objectives at Credite Est

Credite Est (name disguised) is a regional mid-tier bank in France, serving roughly 600,000 customers. The company, which has been growing organically since its inception in 1965, has a quantitative approach to operations. Therefore, the use of quantitative methods in marketing via data mining is second nature to the company. The following example highlights a specific data mining project of the bank.

As is typical for financial services operations, the bank has a very diverse set of customers in terms of customer profitability. Besides using a segmentation scheme based on behavioral characteristics (e.g., product ownership), the company has an activity-based-costing system in place that allows individual customer-level contribution margins to be identified.

The project in question had the business goal to acquire new prospects by using the technique of profiling (see ▶ Sect. 6.2.1). Specifically, the objective was to identify the characteristics of profitable customers in Credite Est's mass-market segment. Once these characteristics are more closely identified, it could then efficiently target similar profiles in the prospect pool. The nature of this project required the bank to go beyond using firm-level data because behavioral (transaction) data are not available for prospects by definition. Since the company does all data mining projects in house, it has considerable experience in the process management of such a project.

7.3.4 Get Raw Data

Now that we have gained a clear understanding of the business objectives, we need to translate them into data requirements (i.e., which data are available that appropriately and accurately describe the problem thereby allowing us to model the targeted behavior?) Once the required data has been identified, it has to be extracted and consolidated in a database (often called analytical data mart) so it is readily available for subsequent data manipulation and data mining steps. Another important step is to check the quality of the analytical raw data. This includes technical checks as well as ensuring the data make sense in the given business context and that correct deductions can be obtained.

During this phase of the project (see Fig. 7.6), database administrators and IT professionals with knowledge of the data source systems will be asked to extract and provide all the data fields required for the data mining project. This is done in close cooperation with the data miners to ensure the extracted data corresponds to the initial requirements. Then, we also need to involve business resources to ensure and cross-check data quality.

Step 1: Looking for Data Sources

To start the acquisition of raw data, we look into data sourcing, a mixed top-down and bottom-up process driven by business requirements (top) and technical restrictions (bottom). Its main objective consists of searching for available data sources in your company (or externally) which describe the problem at hand. The availability of a data warehouse can sometimes speed up this process. Conflicting and bad quality of addresses and other demographic information is quite common. For example, you might find the same or similar information field resides in various source systems, but with contradictory content (e.g., in one database the gender code for a given customer is «male» and in another database it is «female» for the same customer). Data warehouse infrastructures with advanced data cleansing processes can help ensure you are working with high-quality data. It is also a good idea to ask for small sample data extractions from the sources to examine if the information represents what you thought it would. Make sure that you talk to many people from business and data management to understand which data sources are commonly used in certain contexts, but also to detect possible new sources that may contain valuable information. Collect all metadata available to fully understand data types, value ranges, and the primary/foreign key structures.

Once there is a better understanding of the data sources that need to be loaded, build a (simple) relational data model onto which the source data will be mapped. This model should be kept as simple and as close to a business data model as possible. Even though this data model might not be perfectly suited for data mining and analysis, it is important that all involved groups have a clear understanding of the data. Later in the process we will *denormalize* (flatten) the model to enable easier data analysis and predictive modeling.

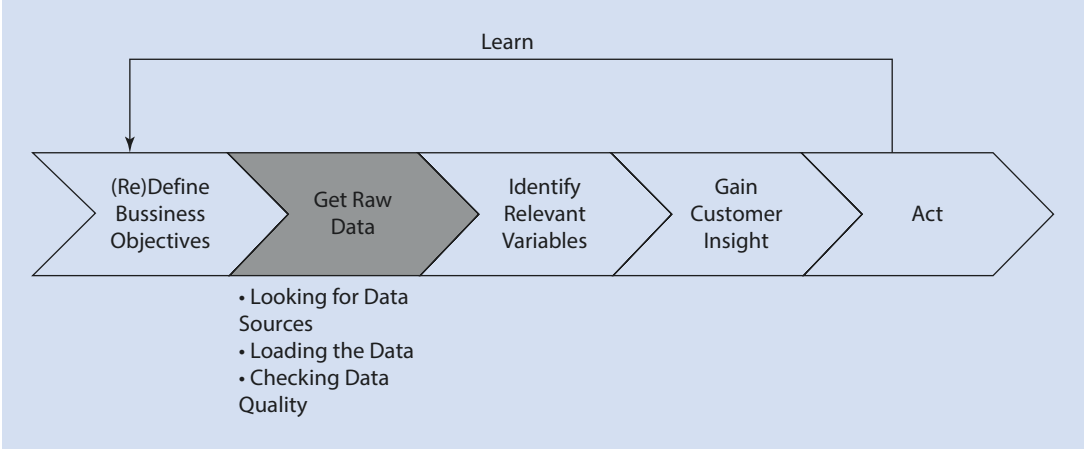


Fig. 7.6 Data mining process: get raw data

Step 2: Loading the Data

After specifying where and how the required data will be extracted, we still need to define further query restrictions because we might want to model only subsets of the full data (e.g., specific customer segments, geographical regions, time periods, etc.). Then it is time to request data management (IT) to deliver the specified data needs.[2] IT teams will prepare the necessary data queries, which will be executed during predefined time windows in batch mode (such as each night at midnight or each Sunday after completion of the accounting batch process).

Depending on the data miner's needs they might also get direct *asynchronous* access to the data so they can run extractions when necessary. The extracted data are then delivered to the data mining environment in a predefined format such as database tables in native format, or simply flat files in ASCII or XML (text) format with fixed or variable record lengths. In fact, flat files are still the most commonly used format for data mining due to their simplicity, enhanced definition of system boundaries, and interfaces. Data miners define how the data will be imported into the data mining environment. Delivery using an ftp protocol is common, or data may also be put onto a common file server to be accessed directly through the network. If a DB-link is preferred, a direct database connection from the data mining system to the source systems (or vice versa) will be established. After the data have been delivered to a defined landing area, they will be further processed and used to fill the previously defined data model in the data mining environment. The involved steps are part of the *ETL process* (Extract-Transform-Load) supported by dedicated software packages. Some data mining tools also offer quite advanced and comprehensive utilities for ETL.

Step 3: Checking Data Quality

It is often underestimated how seriously bad data quality may affect business decisions. According to Olson (2003) the costs of poor data quality are estimated at 15–25% of operating profit, for example through wrong inferences on customer attitudes, lost customers through poor services, or delays in delivering data to decision makers. We need to ensure that once the data for the data mining project have been loaded, we assess and understand their limitations resulting from their inherent quality (good or bad) aspects. We have to create an analytical database that all involved parties (business, data mining, IT) feel comfortable with, as it is the basis for subsequent analyses. Only then can the generated customer insights be trusted and applied in practice with maximum confidence regarding their effect on the organization.

Data quality crucially depends on the intended use and the data itself. Relevant aspects of data quality are:

- Accuracy (consistency and validity)
- Relevance
- Completeness
- Reliability

I.e., when checking data quality, we focus not only on technical aspects of the data (primary keys, duplicate records, missing values, etc.) but also on quality issues related to the business context (a customer should not be 200 years old or have a future birth date, customers should not be purchasing nonexisting or expired products, etc.).

A preliminary data quality assessment is carried out to ensure an acceptable level of quality of the delivered data and to ensure the data mining team has a clear understanding of how to interpret the data in business terms. All parties—business, data mining, and IT—are involved in this important task. Thus, the data available for the mining project must be analyzed to answer to the following questions: (1) Does the data correspond to the original sourcing requirements? (2) Is the quality sufficient? and (3) Do we understand the data?

Several iterations of data extractions may be necessary to satisfy the data requirements. The data miner represents the link between business and IT demands. Miscommunication between business and IT can lead to incorrect data extractions.

As already mentioned, data should have sufficient quality for achieving the project's objectives. A data field does not always have a clearly defined meaning (although available metadata might initially give you that impression). Sometimes the information it carries is different from its official description. This is a consequence of the accumu-

2 Sometimes, obstacles such as lacking authorization of the data mining team for accessing the required data might emerge. Data miners frequently work with data which other business departments do not have access to. There is a high level of secrecy and trust involved.

CRM at Work 7.2
Gathering Raw Data at Credite Est
The response variable for current customers is customer contribution margin. The company sorted customers by operating contribution and chose to profile the top 20% of them. Transaction information is not available for prospects. This is why the bank has to rely on information available for both existing customers and prospects. One type of information is geodemographic data, such as socioeconomic status of a region, average age, type of housing, and so on. They can be purchased from direct marketing agencies and then appended to individual records of existing customers. That is, depending on ZIP code, geodemographic information is added to existing customer records. The model attempts to predict customer operating margin as the dependent variable with geodemographic information as the independent variables. The rationale behind this process is to find the profile that best characterizes high-value clients, which is subsequently applied to prospects' information. Credite Est appended a total of 65 variables to existing customer records. They were procured from the French list manager CIFEA, as well as from Claritas.

lation of undocumented system changes over many years. Another common issue is missing data, which means that in some cases (i.e., for some records) a data field is not filled. We might also find wrong or contradictory information in a data field.

Finally, data miners must demonstrate they understand the data. To this end, it is useful to have them carry out some basic data interpretation and aggregation exercises where two things can be shown: (1) the data quality and (2) the ability to correctly interpret the data. As simple data interpretation examples, consider correctly counting the number of customers per region, customer segment, product ownership, total transaction volume per time periods, etc. Choose to include aggregations familiar to the business group and that can be easily cross-checked. We see an example of getting and combining data in ▶ CRM at Work 7.2 a further study of Credite Est.

Since the objective is to acquire prospects likely to be high-value customers, Credite Est must rely on customer characteristics common to both customers (the basis for establishing the critical profile) and the prospects (scored on the basis of their profile). For a more detailed description of the profiling process see ▶ Sect. 6.2.1.

7.3.5 Identify Relevant Predictive Variables

The raw data, now available for analysis, is not yet in a format suited to powerful predictive modeling. This is due to data formatting aspects, since the sourced data are still in a relational format, and do not yet represent a customer-centric view. During this step (see ◘ Fig. 7.7), we will (1) create a flattened view of the extracted raw data aggregating all facts about the customer behavior over time in a single observation (also called record or row). Also, it is a good practice to include a priori business knowledge by (2) creating new analytical variables which might have predictive power. This part will require imagination and participation from the business group. As a result, we might end up with thousands of variables describing each customer. Further analysis is likely to reveal that most variables do not possess predictive power at all. Therefore, we will (3) identify and select only those few variables with sufficient explanatory power for the modeled target behavior.

Step 1: Create Analytical Customer View: Flattening the Data

In the context of CRM, very often the individual customer is the central object analyzed by means of data mining. All data available for an individual customer must be gathered and consolidated because the individual customer constitutes an observational unit for data analysis and predictive modeling. The historical behavior of customers is obtained from the corresponding data queries in a time series–oriented relational transaction database.

Usually, we choose a simple, flat data model as the basis for predictive modeling. In this representation all data pertaining to an individual customer are contained in one observation (row, record). Individual columns (variables, fields) represent the conditions at specific points in time or a summary over a whole period. Creating such a customer view requires denormalizing the original relational data structures *(flattening)*. This task will involve data miners to define the details of the flattening process and use IT resources to obtain the targeted form of data.

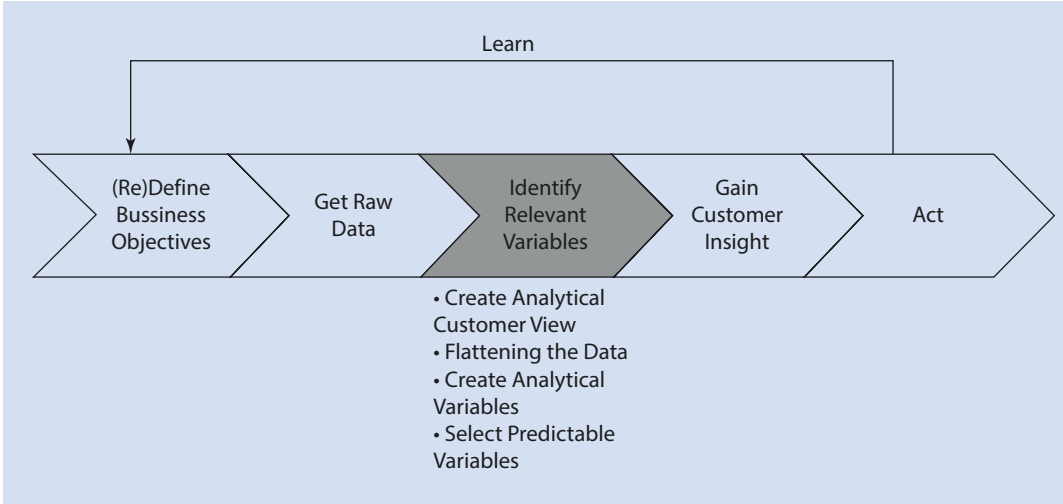


Fig. 7.7 Identify relevant variables

The business objectives for the data mining project determine which features of the customer's record need to be aggregated from the analytical raw data and how. The detail levels for calculating grouped sums (e.g., sum of monthly cash withdrawals from a bank account) and counts (e.g., number of address changes within a certain year) need to be defined. This includes specification of the temporal granularity of the time series in the flattened data table. Descriptive statistics such as sums, mean, median, and standard deviation will be employed to capture features of the related time series. As an example, consider raw data describing 1 year of customer transactions and create four new variables containing the average transaction volume per quarter. Different kinds of global transformations, combinations, or arithmetical operations can also be applied to selected variables such as currency exchange calculations, scaling factors, logarithmic transforms, and so on. Many new variables will be created through these types of operations, leading to very wide data tables. Later, we will use the newly created variables in addition to the raw data variables as predictors during the predictive modeling step.

Another key variable to be created during this step is the target or dependent variable. Its correct definition is extremely important for predictive modeling. In the example of modeling customer defection, a target value of zero is assigned if the customer was still maintaining a business relationship and a target value of one if the customer already terminated the business relationship. The definition of the target variable is not always as straightforward as it might seem. In the previous examples, we could also think about a customer who is inactive since a defined time period as a *lost* customer. There might be a multitude of business rules specifying the conditions under which the target variable is either one or zero. Once we've found a satisfactory definition of the target variable, its values should be generated for all customers and added to the existing data tables.

Step 2: Create Analytical Variables

The basic set of variables resulting from the previous flattening might not be enough to fully explore the data potential for predictive modeling. We might want to introduce additional variables derived from the original ones. For example, consider a variable resulting from the product of customer age and salary. This is often referred to as an *interaction term*. Transforming variables is another operation that might lead to new and more predictive variables. We could transform customer birth date into age, or use the number of days between two customer transactions instead of the absolute dates of each transaction. Variable *binning* (or categorization) is also often encountered. Here we take highly skewed variables (such as salary) and map the distribution to a few discrete classes such as low, medium, and high salary, each defined by its boundary values. More refined methods help to increase normality of variable distributions, which in turn help the predictive model training process. Many data mining tools provide support for increasing normality of the analytical variables. Finally, missing value management is key for enhancing the quality of the data set.

Numerous methods are available, including deleting each row with at least one missing value (the least preferred strategy), replacing a missing field with a constant value, randomly generating a value based on the variable's distribution, and randomly generating a value based on the variable's correlation with the other variables (such as the expectation maximization algorithm).

Step 3: Select Predictive Variables

At this point, we have a wealth of variables describing customer behavior; probably too many to enter the subsequent modeling phase of the data mining project. We now need to reduce the dimensionality (i.e., exclude variables) to get a more parsimonious model. Presenting all predictor variables to a neural network, for instance, might make the modeling phase extremely time consuming and sometimes results in *overfitting*, i.e., the model gives good results on the training data (in sample) but fails to be generally applicable to previously unseen data (out of sample). Exclusion of variables is usually possible without deteriorating the predictive power of the obtained models since many variables have no predictive power at all. To this end we inspect the descriptive statistics of all univariate distributions associated to all available variables. We can immediately exclude those variables, which take on only one value (i.e., the variable is a constant), since they will certainly not have any predictive power. We might also exclude variables with mostly missing values. A threshold missing value count level should be defined above which the field would be excluded from further analysis.

Variables directly or indirectly identifying an individual customer represent another type of candidate for exclusion. Examples are primary keys such as the customer ID number or name and address fields. Later, when deploying predictive models (i.e., when scoring customers), identifiers will usually be required. Otherwise you would not know whom to address with an offering. In some cases, collinear predictive variables can have a negative impact on the convergence and performance of the estimation process of certain types of models such as logistic regression. These collinearities must be identified and the respective variables excluded before proceeding. Finally, we also exclude variables showing little correlation with the target variable. To identify them we may carry out pair-wise chi-square tests, linear correlation analyses, or pair-wise simple linear regressions. Other frequently used techniques to support the variable selection process are histograms, scatter plots, box plots, and frequency tables.

Also notice that excluding variables from further processing does not automatically imply that the respective columns are deleted from the data sets. It could only mean flagging the respective columns to be temporarily ignored for further analysis. The exclusion should be easily reversible to readily test other variable scenario selections.

Before concluding the variable selection step, we should carefully check if all variables have been mapped to the appropriate data types. Some data fields might represent the data in an inappropriate format (e.g., ZIP codes stored as numerical integer variables should rather be categorical (or nominal) for the purpose of data mining, unless you have a ZIP code-based distance measure associated for your analysis). The following example examines this issue.

CRM at Work 7.3

Identifying Relevant Variables at Credite Est

Upon creating a single data file including all appended information, the next step is to start with exploratory analyses. A key concern with appended data is the amount of potentially missing information. All appended variables had almost 50% missing data. The next step was to assess whether the missing data could be meaningfully replaced. These operations improved the overall rate of missing values from 42% to 21%. The next step was to investigate univariate statistics (means, standard deviations, frequencies, outliers) for all variables to ensure the included variables have sufficient integrity. This step brought a reduction in variables from 65 to 54. The next step was to calculate all bivariate correlations (or mean analyses in case of categorical variables) of the existing independent variables with the dependent variable—customer value. This was an iterative process where independent variables were subjected to transformations and where new variables were created. For example, there were three variables which indicated whether a household has children in age brackets 0–4, 5–11, and 12–18. From that, a new variable was created that was a simple dummy indicator: children versus no children. In the end, this data evaluation process resulted in a total of 17 variables that had a reasonable correlation with the dependent variable. These were retained for the next step, the response model.

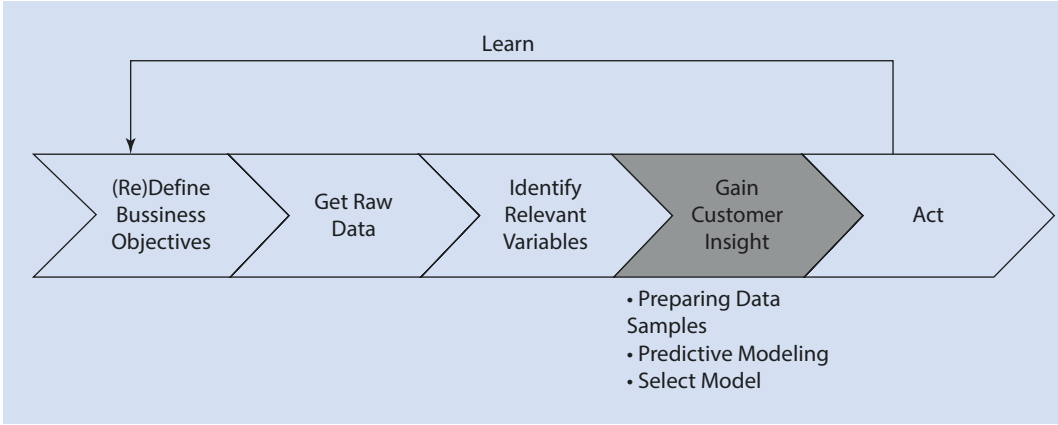


Fig. 7.8 Data mining process: gain customer insight

7.3.6 Gain Customer Insight

Once we have obtained a credible, good-quality set of descriptive data (i.e., we have prepared the data samples), the next step is to extract the knowledge about customer behavior and/or other properties needed for carrying out the planned campaign through predictive modeling (see Fig. 7.8).

Frequently, we distinguish between different types of predictive models obtained through different modeling paradigms: supervised and unsupervised modeling. In the case where we want to predict the likelihood of a customer purchasing a certain product, we would build a predictive model on a predefined test set containing customers who already purchased the product and customers who did not. In this case, we are applying the supervised learning paradigm, because for each customer in the modeling data set we know the correct answer to the question, i.e., did the customer purchase or not?

Building a model means finding the right relationships between the variables describing the customers to predict their respective group membership likelihood: purchaser or non-purchaser. This is usually also referred to as scoring (e.g., between 0 and 1). Since we know the purchase behavior for each customer in the train set, we can also measure the model's prediction quality, i.e., its misclassification rate (see ▶ Sect. 6.3.1). A different situation arises in the context, for instance, of a customer segmentation problem. Suppose you want to identify groups of customers having a similar general behavior, not only with respect to purchase behavior. In the beginning you don't know which groups will be identified. It is a process purely driven by the data and relationships between variables. Here, we would apply unsupervised modeling where group membership is not known beforehand. We're looking for new and unexpected patterns. Typical examples of statistical models in this context are self-organizing neural networks (Kohonen networks) and clustering algorithms.[3]

The output of this project phase can either consist in the predictive model itself, which is later applied in an online production environment (i.e., to predict next product recommendations for customers calling a call center), or directly in the customer score values (e.g., to select all customers with a score value above 90% purchase likelihood and send them a direct mail).

Step 1: Preparing Data Samples

Before we start building (or training) the models, it is necessary to analyze if sufficient data are available to obtain statistically significant results.

3 Kohonen networks belong to the family of neural network techniques. These are powerful data modeling tools able to capture and represent complex input/output relationships for example in target marketing, financial forecasting, or process control. In particular, the objective of a Kohonen network is to generate, out of complex input patterns of arbitrary dimension, a simplified (discrete) map with very few dimensions, say 1 or 2. Thus, the Kohonen network is an approach to quickly understand complex data as a result of a simplification of the structure. For a good overview of neural networks and Kohonen networks please refer to: Principe, Euliano, and Lefebvre (2000).

There are cases where there is only very little data available such as when modeling purchase behavior for a recently introduced product, with only very few customers having bought the product until now. If we have enough data available we split the data into two samples: the train set to fit the models and the test set to check the model's performance on observations that have not been used to build it. This will give an objective assessment of the model's generalization capability—a critical requirement before launching a product or campaign.

Step 2: Predictive Modeling

There are two steps of predictive modeling:

- The rules (or linear/nonlinear analytical models) are built based on a training set.
- These rules are then applied to a new dataset for generating the answers needed for the campaign.

Based on the training set, we develop predictive models that should minimize the prediction error. In the course of this process a set of optimal model parameters are obtained. Usually, several alternative models are trained together, applying different statistical methodologies such as neural networks, linear or logistic regression, survival analysis, principal component analysis, factor analysis, decision trees, or clustering.

Step 3: Select Model

When all alternative models have been trained, we start comparing their relative quality of prediction by comparing their respective misclassification rates (see ▶ Sect. 6.3.1) obtained on the test set or by performing a lift analysis (see ▶ Sect. 6.3.2). Some models will have more predictive power than others, and we will select the model we think generalizes best from the train to the test data.

We will also include the economic implications of a model by applying the previously defined cost/revenue matrix. Predictive models, for instance, deliver a score value, or likelihood, for each customer to show the modeled target behavior (e.g., purchase of a credit card). Nevertheless, determining the threshold level score to use for a given campaign is a business decision. It could be that you want to set it to the break-even point (see ▶ Sect. 6.1.1), or you may have a fixed budget for the campaign you want to fully use. This might lead to lowering the threshold until the point where your costs equal the budget. We continue to look at this issue in the following example.

7.3.7 Act

The final objective of a data mining project should be to act on its results (see ◘ Fig. 7.9). Sometimes we also refer to this as deployment of the results. This is crucial to the success of the whole project. The planning phase of the project must have addressed the issue of implementing the project's results into the respective business processes. The project plan must foresee involvement and availability of IT resources required to feed data mining results back into the process supporting IT systems (databases, Web sites, call centers, etc.). In practice, deployment can have numerous applications: score-based selection of customers to be addressed through a direct mailing cam-

CRM at Work 7.4
Gaining Customer Insight at Credite Est

The methodology chosen by the modelers was logistic regression. Since the goal was to either target or not target a certain individual in the prospect pool, classifying the dependent variable as 0/1 was appropriate. In the previous step, only those variables were retained with a minimum level of bivariate correlation. However, now the issue of multicollinearity came into play. Multicollinearity occurs when two variables convey essentially the same information, making one of them redundant. Thus, an important step was to make a theory-based elimination of those highly collinear variables. The final model was chosen on grounds of predictive ability while containing a low number of missing values. It contained five predictors of customer value: bourgeois cluster, technology cluster, children index, house value index, and managerial job position. The ability of the model to correctly classify was 75.5% in the estimation sample and 69.8% in the holdout sample, i.e., roughly 20% points higher than based on chance alone. This result was deemed successful, and thus it was decided to utilize this model for a prospecting campaign.

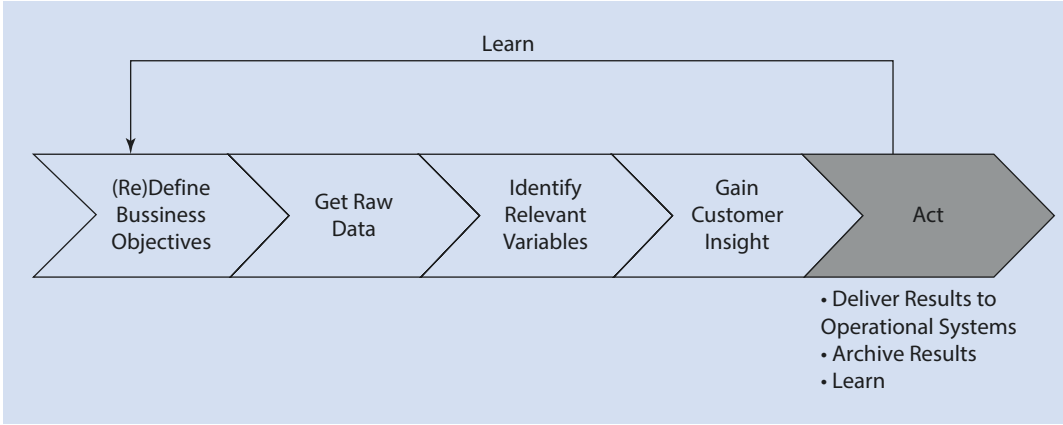


Fig. 7.9 Data mining process: act

paign, score-based next-product recommendation on an e-commerce Web site, optimization of marketing spending according to the model-based customer lifetime value prediction, choice of the appropriate communication channel for each customer, and so on.

In particular, acting can be subdivided into delivering results to operational systems, archiving the results, and learning.

Step 1: Deliver Results to Operational Systems

The final goal of prognostic modeling within the context of CRM is to select a subset of customers for a campaign to determine which customers are more likely to be responsive than others. To identify this subset, we apply the selected model to the entire customer base (unless restrictions have been previously defined limiting the total universe of modeled and targeted customers, such as geographical regions, a subset of customer segments, etc.). The obtained score value for each customer and the defined threshold value will determine whether the corresponding customer qualifies to participate in the campaign. We can either deploy the customer scores or, alternately, the scoring model itself, which implies that it is applied on demand—for example, when a customer calls the call center or visits the company's Web site.

Before scoring customers we need to prepare the score data set containing the most recent information available for each customer with the variables required by the model. This implies that the score set variables go through exactly the same variable transformation, derivation, and selection process as did the train and test data sets used for building the model. Data recency is an important requirement because otherwise we are scoring customers based on old information, which may, in turn, lead to wrong conclusions.

Imagine we are scoring customers for a direct mail campaign to sell a credit card to all those customers not yet owning one. If the scoring data are not reasonably current, we might be scoring customers although they have recently purchased a credit card and potentially (if the model works well) include them in your target group. As a result, we end up offering these customers a product they have just purchased, giving a rather poor image of how much our company knows about its customers.

Finally, when delivering the results to the operational systems, make sure to also provide the necessary customer identifiers required by those systems to unambiguously link the models score information to the correct customer.

Step 2: Archive Results

The data mining group is responsible for archiving all information related to each data mining project it executes. This is an important and often neglected or poorly followed piece of advice. Companies that do not archive their models cannot expect to learn from past experience as fast as those who do.

Each data mining project will produce a huge amount of information:

- Raw data used
- Transformations for each variable
- Formulas for creating derived variables
- Train, test, and score data sets
- Target variable calculation
- Models and their parameterizations
- Score threshold levels
- Final customer target selections

Knowing this information and having it readily available helps in understanding anomalies in model performance, and in learning what worked well and what did not (and why). It is also useful to preserve the details of the model when scoring has been done. The same model might be used to score different data sets obtained at different times.

Step 3: Learn

Learning from a data mining project is an integral part of the process. This is also sometimes referred to as *closing the loop*. It means learning from the actions you have executed to improve performance the next time. To learn from the data mining project, we must first obtain the facts describing its performance and business impact. In the ideal case, we would provide return on investment figures for the data mining project at hand.

Usually these facts are obtained by monitoring the campaign performance while it is running and from the final campaign performance analysis after the campaign has ended. Campaign monitoring is an important capability the data mining group must provide, since it avoids blind piloting of the campaign until its end, without any intermediate performance feedback. In rapidly changing environments, it is also required for detecting when a model should be retrained. Usually, monitoring provides some key performance indicators, such as the response and/or purchase rates by region, customer segment, product, and so on. These parameters will give early indication of undesired irregularities in model performance and enable early intervention. The final campaign performance analysis will produce similar performance indicators as the monitoring function. The main difference is that it is more complete and has a determined final time horizon of influence. This is required to ensure a correct measurement of the cause and effect of a campaign. It would, for instance, be unrealistic to positively attribute advertising to a customer's behavior when the customer purchases a product 1 year after seeing the advertisement of a direct mail campaign.

Sample learnings from campaign evaluation could be:

- Revealing that purchase rates depend on the choice of the communication channel.
- Discovering that a direct mail with a colorful and detailed product brochure sells less than one with a black and white one-page flyer.

Thus, the learning step requires data miners to generate the facts about campaign performance, and business resources to put them in context for correct interpretation. Our study of Credite Est concludes with the following example, showing how the company acted on the information it gleaned in the data mining process.

CRM at Work 7.5
Acting on the Information at Credite Est

The final model was rolled out in a sequential fashion to the target prospect audience. The goal was to iteratively refine the model in future rounds. As a first step, Credite Est purchased addresses from list brokers that had nonmissing values for at least three out of five variables in the final model. The prospects were scored with the model and then ranked by likelihood of being a high-value customer. From the resulting pool of 10,000 prospects, half were targeted with a money-market product, and halfwith a lending product. The objective was to assess the receptivity of the two samples for the respective products. In addition, a baseline scenario was conducted whereby the same prospecting campaigns were conducted for a random sample of households. Although both target mailings were significantly more successful than the baseline scenario, this was only the first step in a further refinement of the model and the offer. In particular, besides assessing response rate, it was now important to track and document the value of the acquired customers—the original goal of the project.

A comprehensive example of an application of data mining is given in the case study, which is taking a look at Yapi Kredi, a company that put the data mining tools to use to create a cross-selling campaign.

CRM at Work 7.6

Yapi Kredi—Predictive Model-Based Cross-Selling Campain

Established in 1944 as the first private bank in Turkey, Yapi Kredi has always been a pioneer in the Turkish financial sector. The bank has more than 860 domestic branches and various other subsidiaries, as well as affiliated companies active in leasing, factoring, investment banking insurance, brokerage and new economy companies. Yapi Kredi is positioned as the fourth largest privately owned commercial bank by asset size in Turkey, with leading positions in credit cards, assets under management, factoring, private pension funds, and life- and non-life insurances. As of 2010 it serves approximately six million customers.

The Challenge

To continue Yapi Kredi's development as the fastest-growing retail bank in Turkey, in terms of assets under management and retail profitability, it targets to maintain anintimate banking relationship with the top customer segment to fully explore the potential of its 5+ million customer base, and to increase the contract per customer ratio to five.

To this end, Yapi Kredi introduced a modern retail banking approach to enable serving all customers according to their specific needs through individual product packages. The capabilities required to achieve these goals were as follows:

- Advanced analytical customer segmentation.
- Segment-specific offering of product bundles.
- Conversion of customers to more profitable segments via targeted campaigns using advanced CRM tools such as predictive modeling.

Solution

To increase the product per customer ratio, to attract new money from customers, and to demonstrate the efficacy of the new analytical CRM methods, Yapi Kredi decided to carry out a set of pilot projects for cross-selling of consumer banking products. A reduced selection of target customers with a high propensity to positively respond would be included in a multichannel, two-step campaign. To illustrate the methodology we briefly describe the outcomes of the various project phases.

Define Business Objectives

Various cross-departmental workshops were held to define the business objectives, operational aspects of campaign execution, the basics of relevant data availability, and to measure the success of the campaign.

The first step was to find which product would be best suited for cross selling from a customer and bank perspective. After a deep analysis of potential products to be offered during this first predictive model based cross-selling campaign, it was decided to choose Yapi Kredi's B-type mutual funds, characterized by being low risk investment instruments based on fixed income securities. These funds can be easily purchased via the ATM, Web, and telephone channels.

Cross selling these mutual funds was considered to have a twofold positive business impact. It served the purpose of acquiring new money from customers, and even those customers transferring their existing investments from other Yapi Kredi products into mutual funds (*cannibalization effect)* was still considered beneficial to the bank. It was decided to offer this product to both customer groups:

- Customers already having invested into B-type mutual funds to stimulate an increase of the assets.
- Customers not yet owning any B-type fund to help increase product ratio and attract new money.

After fixing the product details, it had to be defined how the campaign would be carried out. The workshops helped define the start and end date of the campaign: a total duration of 5 weeks was considered appropriate. Communication channels for offering the product were agreed upon. A two-channel approach was deemed feasible since Yapi Kredi had just finished the implementation of a project integrating the call center and the bank's branch network. These were considered the right channels for the campaign.

Given the pilot project character of the campaign and the available resources, it had been

decided to contact 3000 customers based on outbound calls and active marketing during customer branch visits. A total of 16 branches in the Istanbul area were selected for participation in the campaign. Additionally, 1200 target customers were to be contacted by the call center.

It was decided to run a two-step campaign, where customers were first contacted with the B-type mutual fund offer. Then, positive responders received a follow-up call if they had not purchased 1 week after their initial positive response.

Response and purchase rates by contact channel (branch or call centre) were chosen as measures of the campaign's success.

Get Raw Data

A data mart was developed for supporting the activities of the CRM department. To this end, data were extracted from more than 50 source system tables. About 20 database tables were produced with 30 gigabytes of disk space for the initial project phase. The data mart included data most urgently needed for high-priority business activities (such as the pilot campaign) and assured that the data are readily available in a short time frame for subsequent data manipulation and data mining steps.

Identify Relevant Variables

Various aggregations and transformations were required to obtain the right customer-centric data format as needed for analysis and predictive modeling. Basic data-quality crosschecks were performed to assure the validity of the data and its suitability for further data mining activities.

Different types of attributes were found to be relevant and used to obtain a complete picture of customer behavior and preferences. These included the following customer attributes:

- **Demographics:** Age, gender, marital status, group memberships, address, profession, and other identifying characteristics belong in this category.
- **Product ownership:** This relates to product portfolio held by each customer, opening/closing dates, derived variables related to customer's tenure such as maximum tenure of owned products, and so on.
- **Product usage:** Variables are related to a customer's frequency of usage such as the average number of banking transactions.
- **Channel usage:** Variables are related to customer's automatic payment behavior,average amount of automatic payments, ratios of different channels' usage, and so on.
- **Assets:** Variables are related to savings and investment products such as the average balance invested in securities, time deposits, demand deposits, and so on.
- **Liabilities:** Variables are related to loan usage such as average balance on loans, average balance on credit cards, and so on.
- **Profitability:** For the pilot project, a profitability index was created, since profitability was not available for all customers at that time. The index was used for ranking customers according to their profitability without giving its absolute value.

Gain Customer Insight

Based on 6 months of historical customer data, five different predictive models were developed to estimate a customer's propensity to invest in a B-type mutual fund during the following 3-month period. The best model was found to be a logistic regression yielding a lift value of 2.9 for the top customer decile. The lift value measures the effect of the predictive model (see ▶ Sect. 6.3.2), and expresses the fact that in this case the logistic regression reaches 2.9 times more responders for the top customer decile than a random selection of the same size.

All customers were then scored using this model, and a set of 4200 customers with the highest propensity to purchase was selected as the target group for the pilot campaign.

Act

To roll out the campaign through the call center and the branches, each channel had to know exactly which customer to contact. Each channel needed a clear assignment of their respective target customers. A subset of 3000 customers was assigned to the 16 branches holding the responsibility for the respective relationships. The remaining 1200 customers were assigned to the call center. The target list with the corresponding channel assignment was then made available to the campaign management system. After preparing call scripts and training the staff involved in its execution, the campaign could start.

The following table summarizes the results (Table 7.2). Impressive response rates of 6.5% and 12.2%, respectively, were obtained with the branch-based and call-center-based part of the campaign. The pilot campaign could acquire more than €1 million into B-type mutual funds.

Table 7.2 Response rate and amount of funds sold

	Response rate (%)	Amount of funds sold (€)
Branches	6.5	582,000
Call center	12.2	452,000
Total	8.2	1,034,000

It is interesting to observe that although the branches obtained a lower response rate than the call center, they still acquired significantly more investment into the funds. This reflects the advantages of the more personal branch-based customer relationship. As a consequence of this successful pilot campaign, a large-scale rollout to a larger part of Yapi Kredi's customers looked very promising.

Summary

Data mining can assist in selecting the right target customers or in identifying previously unknown customers with similar behavior and needs. A good target list is likely to increase purchase rates and have a positive impact on revenue. A complete data mining process comprises assessing and specifying the business objectives, data sourcing, transformation and creation of analytical variables, as well as building analytical models using techniques such as logistic regression or neural networks. The number of variables used changes drastically during the data mining process. Types of row manipulation include aggregation, change, missing value, and outlier detection. Profitable customer acquisition requires modeling of expected customer potential over the lifetime of the business relationship. In a cross-selling or up-selling model, we try to predict the customer's affinity with a set of products or services translated into the customer's purchase likelihood.

Another aspect of the campaign that should be defined is the set of business or selection rules for a campaign, which specifies the customers who should be excluded from or included in the target groups. To measure how the model based selection is performing with respect to the average customer behavior, the target group can be split into various cells like the control group— containing only randomly selected customers and another cell containing only the best customers according to the implemented model. It is helpful to define a cost/revenue matrix describing how the business mechanics will work in the supported campaign and how it will impact the data mining process.

Once, the required data have been identified, extracted and consolidated, so that the data in a database (often also called *analytical data mart)* are readily available for subsequent data manipulation and data mining steps. Another important step is to check the quality of the analytical raw data. Data warehouse infrastructures with advanced data cleansing processes can help ensure you are working with high-quality data. Missing value management is a key element for enhancing the data quality. A preliminary data quality assessment is carried out to assure a good level of quality of the delivered data, and that the data mining team has a clear understanding of how to interpret the data in business terms.

It is important to identify and select only those variables with good explanatory power (relevant predictive power) for the modeled target behavior. Different methods are employed for selecting the predictor variables. These methods help us drop collinear variables and those with very low correlation with the target variable. In the context of Customer Management, very often the individual customer is the central object analyzed by means of data mining methods. Usually a very simple, flat data model is chosen as the basis for predictive modeling. In this representation, all data pertaining to an individual customer is contained in one observation (row). Individual columns (variables, fields) represent the conditions at specific points in time or a summary over a whole period. Descriptive statistics such as sums, mean, median, and standard deviation will be employed to capture features of the related time series. Another key variable to be created during this step is the target or dependent variable, needed for predictive modeling. Once a satisfactory definition of the target variable is achieved, its values will be generated for all customers and added to

the existing data tables. Many data mining tools provide support for increasing normality of the analytical variables.

The next step is to select the best model to predict the dependent variable. The performance of different competing models is compared using classification tables and lift analysis (see ► Sect. 6.3). The final step in the data mining project is acting based on final results. In this step, customers and prospects are scored and ranked to identify the right customers and prospects to target. Archiving and comparing the business results with the objectives initially set for the project are important activities of the data mining process in order to derive learnings for future data mining projects.

International Perspectives: Did You Know?

1. The Singapore based Oversea-Chinese Banking Corporation (OCBC) applies Data Mining to improve their marketing effectiveness. The publicly listed banking and financial services provider analyzed historic data of their clients to determine customer preferences and design marketing activities accordingly. More precisely, OCBC has designed an event-based marketing strategy that focuses on various coordinated, personalized marketing communications across multiple channels and touch points including email, call centers, branches, ATMs, direct mail, text messages and 3G mobile banking. The company launched its enterprise marketing management system in 2005 and has achieved a positive return on investment (ROI) within 18 months. With the use of data mining, the banking corporation has increased its conversion rates by 45 percent and cross-sales by 60 percent. Furthermore, OCBC was able to improve its overall marketing productivity and is running 12 times more campaigns compared to before (Turner, Schroeck, & Shockley, 2013).
2. The Swedish music and video streaming service Spotify is another great example of how to use the potential of data mining to improve its product offering. Other than most online companies that use the affinity model, which assumes that users who like some of the same music share the same tastes, Spotify analyzes audio «fingerprints» and crawls the web for information about music, including reviews, blogs and social media posting. The music service is able to identify «clusters» where a user is listening to the same kind of songs at different times every week and identify a commonality there. In the next step, Spotify aims to provide «situational playlisting», meaning that the company will use the understanding of music and users to figure out what listener want to hear at a given moment. This will enable Spotify «to connect the right playlist to the right listener at the right moment» (Levine, 2015).

Acknowledgments We thank Frank Block, Ph.D., of FinScore Corporation (Switzerland) for his collaboration on this chapter.

References

Levine, R. (2015, April 3). Data mining the digital gold rush: 4 companies that get it. *Billboard*. http://www.billboard.com/articles/business/6524078/big-data-mining-digital-gold-rush-companies-that-get-it. Accessed May 2, 2017.

Olson, J. (2003). *Data quality – The accuracy dimension*. Amsterdam, The Netherlands: Kaufmann.

Principe, J. C., Euliano, N. R., & Lefebvre, W. C. (2000). *Neural and adaptive systems: Fundamentals through simulations*. New York: Wiley.

Turner, D., Schroeck, M., & Shockley, R. (2013). Analytics: The real-world use of big data in financial services. *IBM Global Business Services*. https://www-935.ibm.com/services/multimedia/Analytics_The_real_world_use_of_big_data_in_Financial_services_Mai_2013.pdf. Accessed May 2, 2017.