Commercial Data Mining

Commercial Data Mining

Processing, Analysis and Modeling for Predictive Analytics Projects

David Nettleton

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Morgan Kaufmann is an imprint of Elsevier

MK
MORGAN KAUFMANN

Acquiring Editor: Andrea Dierna
Editorial Project Manager: Kaitlin Herbert
Project Manager: Punithavathy Govindaradjane
Designer: Matthew Limbert

Morgan Kaufmann is an imprint of Elsevier
225 Wyman Street, Waltham, MA 02451, USA

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods or professional practices, may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information or methods described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data
Nettleton, David, 1963-
Commercial data mining : processing, analysis and modeling for predictive analytics projects / David Nettleton.
pages cm
Includes bibliographical references and index.
ISBN 978-0-12-416602-8 (paperback)
1. Data mining. 2. Management–Mathematical models. 3. Management–Data processing. I. Title.
HD30.25.N48 2014
658′.056312–dc23
2013045341

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-0-12-416602-8

Printed and bound in the United States of America
14 15 16 17 18 10 9 8 7 6 5 4 3 2 1

Acknowledgments

Dr. David Nettleton is a contract researcher at the Pompeu Fabra University, Barcelona, Catalonia, Spain, and at the IIIA-CSIC, Bellaterra, Catalonia, Spain.

I would first like to thank the reviewers of the book proposal: Xavier Navarro Arnal, Dr. Anton Dries (KU Leuven, Belgium), Tim Holden (Imagination Technologies, United Kingdom) and Colin Shearer (Advanced Analytic Solutions, IBM). Second, I would like to thank the reviewers of the book manuscript: Dr. Vicenç Torra (IIIA-CSIC, Catalonia, Spain), Tim Holden, Dr. Anton Dries, Joan Gómez Escofet (Autonomous University of Barcelona, Catalonia, Spain), and the anonymous reviewers for their suggestions, corrections, and constructive criticisms, which have helped enhance the content.

Third, I would like to thank the following people and companies for their permission to adapt and include material for the book: IBM, Newprosoft, Professor Ian Witten, Dr. Ricardo Baeza-Yates (Yahoo! Research and Pompeu Fabra University, Catalonia, Spain), Dr. Joan Codina-Filba (Pompeu Fabra University, Catalonia, Spain), Dra. Liliana Calderón Benavides (Autonomous University of Bucaramanga, Columbia), and Néstor Martinez.

I would also like to thank Kaitlin Herbert and Andrea Dierna of Morgan Kaufmann Publishers for their belief in this project and their support throughout the preparation of the book.

Finally, I would like to thank my supporting institutions: Pompeu Fabra University, Catalonia, Spain, and the IIIA-CSIC, Catalonia, Spain.

Chapter 1

Introduction

This book is intended to benefit a wide audience, from those who have limited experience in commercial data analysis to those who already analyze commercial data, offering a vision of the whole process and its related topics. The author includes material from over 20 years of professional business experience as well as a diversity of research projects he was involved in, in order to enrich the content and give an original approach to commercial data analysis. In the appendix, practical case studies derived from real-world projects are used to illustrate the concepts and techniques that are explained throughout the book. Numerous references are included for those readers who wish to go into greater depth about a given topic.

Many of the methods, techniques, and ideas presented, such as data quality, data mart, customer relationship management, data sources, and Internet searches, can be applied by small business owners, freelance professionals, or medium to large-sized companies. The reader will see that it is not a prerequisite to have large volumes of data, and many tools used for data analysis are available for a nominal cost.

Although the steps in Chapters 2 through 10 can be carried out sequentially, note that, in practice, aspects such as data sources, data representation, and data quality are often carried out in parallel and reiteratively. This also applies to the variable/factor selection, analysis, and modeling steps. However, note that the better each step is performed, the fewer iterations will be necessary.

In order to obtain meaningful results, data analysis requires an attention to detail, an adequate project definition, meticulous preparation of the data, investigative capacity, patience, rigor, and objectives that are well defined from the beginning. If these requirements are taken together as a starting point, then a basis can be built from which a data warehouse is converted into a high-value asset. One of the motivators for data analysis is to realize a return on investment for the database infrastructures that many businesses have installed. Another is to gain competitive leverage and insight for products and services by better understanding the marketplace, including customer and competitor behavior.

The analysis and comprehension of business data are fundamental parts of all organizations. Monitoring national economies and retail sales tendencies depend on data analysis, as does measuring the profitability, costs, and competitiveness of commercial organizations and businesses. Analyzing customer data has become easier due to data management infrastructures that separate the operational data from the analytical data, and from Internet applications and cloud computing, which facilitate the gathering of large-volume historical data logs.

On the other hand, computer systems have swamped us with large volumes of data and information, much of which is irrelevant for a specific analysis objective. Also, customer behavior has become more complex due to the diversity of applications that compete in the marketplace, especially for mobile devices. Thus, the objective of data analysis should be that of discovering useful and meaningful knowledge and separating the relevant from the irrelevant.

Chapters 2 through 10 follow the sequential steps for a typical data mining project. A scheme of the organization of these chapters can be seen in

Chapter 2, "Business Objectives," discusses the definition of a data mining project, including its initial concept, motivation, business objectives, viability, estimated costs, and expected benefit (returns). Key considerations are defined and a way of quantifying the cost and benefit is presented in terms of the factors that most influence the project. Finally, two case studies illustrate how the cost/benefit evaluation can be applied to real-world projects.

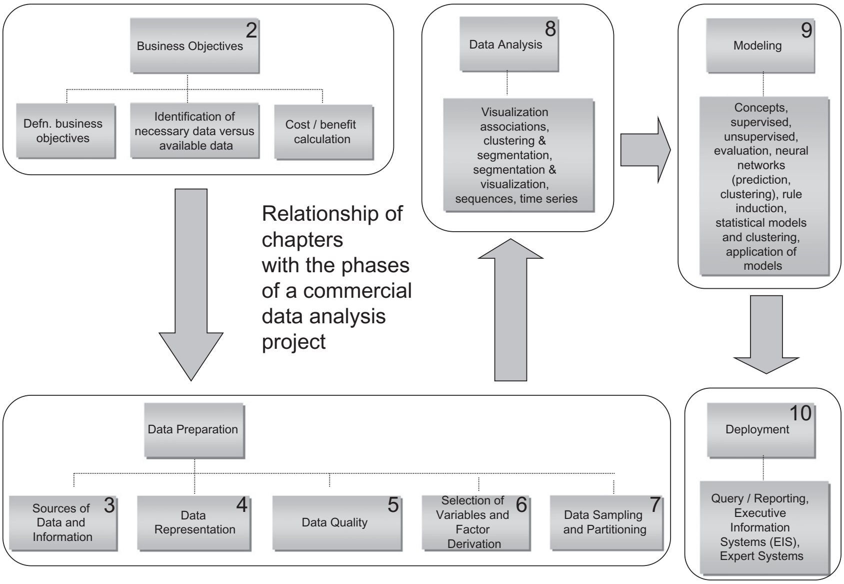
2
Business Objectives
Defn. business objectives
Identification of necessary data versus available data
Cost / benefit calculation
8
Data Analysis
Visualization associations, clustering & segmentation, segmentation & visualization, sequences, time series
9
Modeling
Concepts, supervised, unsupervised, evaluation, neural networks (prediction, clustering), rule induction, statistical models and clustering, application of models
Relationship of chapters with the phases of a commercial data analysis project
Data Preparation
3
Sources of Data and Information
4
Data Representation
5
Data Quality
6
Selection of Variables and Factor Derivation
7
Data Sampling and Partitioning
10
Deployment
Query / Reporting, Executive Information Systems (EIS), Expert Systems

FIGURE 1.1 Relationship between chapters and the phases of a commercial data analysis project

Chapter 3, “Incorporating Various Sources of Data and Information,” discusses possible sources of data and information that can be used for a commercial data mining project and how to establish which data sources are available and can be accessed for a commercial data analysis project. Data sources include a business’s own internal data about its customers and its business activities, as well as external data that affects a business and its customers in different domains and in given sectors: competitive, demographic, and macro-economic.

Chapter 4 “Data Representation,” looks at the different ways data can be conceptualized in order to facilitate its interpretation and visualization. Visualization methods include pie charts, histograms, graph plots, and radar diagrams. The topics covered in this chapter include representation, comparison, and processing of different types of variables; principal types of variables (numerical, categorical ordinal, categorical nominal, binary); normalization of the values of a variable; distribution of the values of a variable; and identification of atypical values or outliers. The chapter also discusses some of the more advanced types of data representation, such as semantic networks and graphs.

Chapter 5, “Data Quality,” discusses data quality, which is a primary consideration for any commercial data analysis project. In this book the definition of “quality” includes the availability or accessibility of data. The chapter discusses typical problems that can occur with data, errors in the content of the data (especially textual data), and relevance and reliability of the data and addresses how to quantitatively evaluate data quality.

Chapter 6, “Selection of Variables and Factor Derivation,” considers the topics of variable selection and factor derivation, which are used in a later chapter for analysis and modeling. Often, key factors must be selected from a large number of variables, and to do this two starting points are considered: (i) data mining projects that are defined by looking at the available data, and (ii) data mining projects that are driven by considering what the final desired result is. The chapter also discusses techniques such as correlation and factor analysis.

Chapter 7, “Data Sampling and Partitioning,” discusses sampling and partitioning methods, which is often done when the volume of data is too great to process as a whole or when the analyst is interested in selecting data by specific criteria. The chapter considers different types of sampling, such as random sampling and sampling based on business criteria (age of client, length of time as client, etc.).

With Chapters 2 through 7 having laid the foundation for obtaining and defining a dataset for analysis, Chapter 8, “Data Analysis,” describes a selection of the most common types of data analysis for data mining. Data visualization is discussed, followed by clustering and how it can be combined with visualization techniques. The reader is also introduced to transactional analysis and time series analysis. Finally, the chapter considers some common mistakes made when analyzing and interpreting data.

Chapter 9, “Modeling,” begins with the definition of a data model and what its inputs and outputs are, then goes on to discuss concepts such as supervised and unsupervised learning, cross-validation, and how to evaluate the precision of modeling results. The chapter then considers various techniques for modeling data, from AI (artificial intelligence) approaches, such as neural networks and rule induction, to statistical techniques, such as regression. The chapter explains which techniques should be used for various modeling scenarios. It goes on to discuss how to apply models to real-world production data and how to evaluate and use the results. Finally, guidelines are given for how to perform and reiterate the modeling phase, especially when the initial results are not the desired or optimal ones.

Chapter 10, “Deployment Systems: From Query Reporting to EIS and Expert Systems,” discusses ways that the results of data mining can be fed into the decision-making and operative processes of the business.

Chapters 11 through 19 address various background topics and specific data mining domains. A scheme of the organization of these chapters can be seen in

Chapter 11, “Text Analysis,” discusses both simple and more advanced text processing and text analysis: basic processing takes into account format checking based on pattern identification, and more advanced techniques consider named entity recognition, concept identification based on synonyms and hyponyms, and information retrieval concepts.

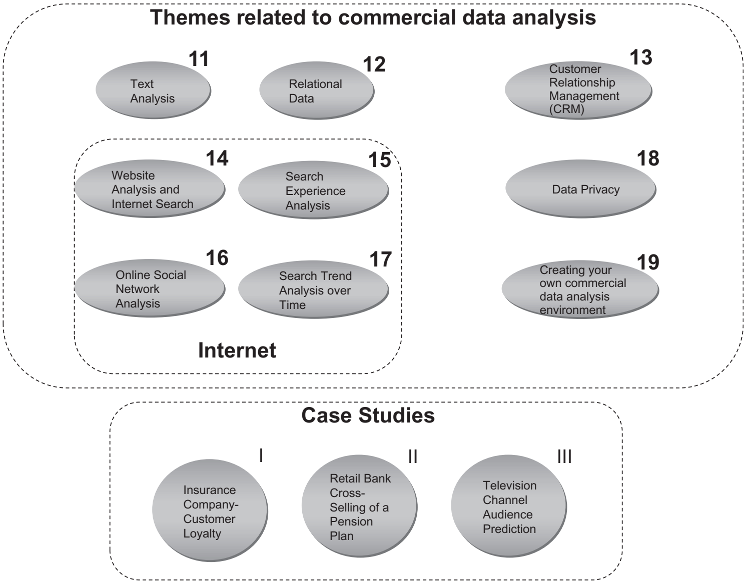
Themes related to commercial data analysis
11
Text Analysis
12
Relational Data
13
Customer Relationship Management (CRM)
14
Website Analysis and Internet Search
15
Search Experience Analysis
18
Data Privacy
16
Online Social Network Analysis
17
Search Trend Analysis over Time
19
Creating your own commercial data analysis environment
Internet
Case Studies
I
Insurance Company-Customer Loyalty
II
Retail Bank Cross-Selling of a Pension Plan
III
Television Channel Audience Prediction

FIGURE 1.2 Chapter topics related to commercial data analysis and projects based on real-world cases

Chapter 12, “Data Mining from Relationally Structured Data, Marts, and Warehouses,” deals with extracting a data mining file from relational data. The chapter reviews the concepts of “data mart” and “data warehouse” and discusses how the informational data is separated from the operational data, then describes the path of extracting data from an operational environment into a data mart and finally into a unique file that can then be used as the starting point for data mining.

Chapter 13, “CRM – Customer Relationship Management and Analysis,” introduces the reader to the CRM approach in terms of recency, frequency, and latency of customer activity, and in terms of the client life cycle: capturing new clients, potentiating and retaining existing clients, and winning back ex-clients. The chapter goes on to discuss the characteristics of commercial CRM software products and provides examples and functionality from a simple CRM application.

Chapter 14, “Analysis of Data on the Internet I – Website Analysis and Internet Search,” first discusses how to analyze transactional data from customer visits to a website and then discusses how Internet search can be used as a market research tool.

Chapter 15, “Analysis of Data on the Internet II – Search Experience Analysis,” Chapter 16, “Analysis of Data on the Internet III – Online Social Network Analysis,” and Chapter 17, “Analysis of Data on the Internet IV – Search Trend Analysis over Time,” continue the discussion of data analysis on the Internet, going more in-depth on topics such as search experience analysis, online social network analysis, and search trend analysis over time.

Chapter 18, “Data Privacy and Privacy-Preserving Data Publishing,” addresses data privacy issues, which are important when collecting and analyzing data about individuals and organizations. The chapter discusses how well-known Internet applications deal with data privacy, how they inform users about using customer data on websites, and how cookies are used. The chapter goes on to discuss techniques used for anonymizing data so the data can be used in the public domain.

Chapter 19, “Creating an Environment for Commercial Data Analysis,” discusses how to create an environment for commercial data analysis in a company. The chapter begins with a discussion of powerful tools with high price tags, such as the IBM Intelligent Miner, the SAS Enterprise Miner, and the IBM SPSS Modeler, which are used by multinational companies, banks, insurance companies, large chain stores, and so on. It then addresses a low-cost and more artisanal approach, which consists of using ad hoc, or open source, software tools such as Weka and Spreadsheets.

Chapter 20, “Summary,” provides a synopsis of the chapters.

The appendix details three case studies that illustrate how the techniques and methods discussed throughout the book are applied in real-world situations. The studies include: (i) a customer loyalty project in the insurance industry, (ii) cross-selling a pension plan in the retail banking sector, and (iii) an audience prediction for a television channel.

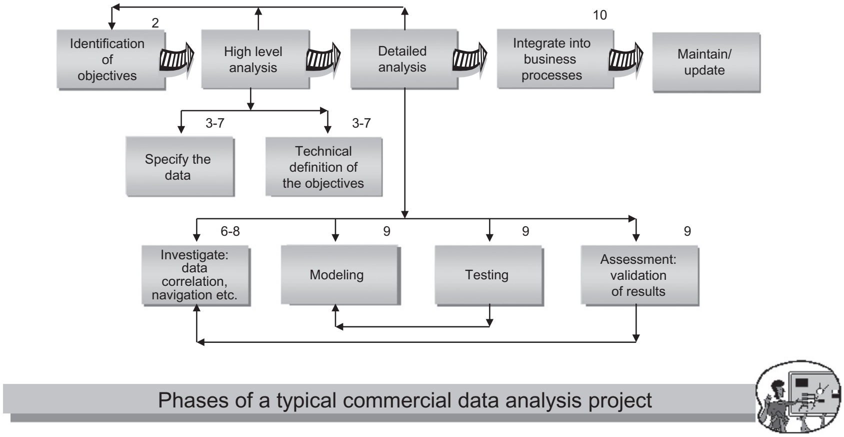
2
Identification of objectives
High level analysis
Detailed analysis
10
Integrate into business processes
Maintain/ update
3-7
3-7
Specify the data
Technical definition of the objectives
6-8
9
9
9
Investigate: data correlation, navigation etc.
Modeling
Testing
Assessment: validation of results
Phases of a typical commercial data analysis project

FIGURE 1.3 Life cycle of a typical commercial data analysis project (numbers correspond to chapters)

For readers who wish to focus on business aspects, the following reading plan is recommended: Chapters 2, 3, 6, 8 (specifically the sections titled "Visualization," "Associations," and "Segmentation and Visualization"), 9, 10, 11 (specifically the sections titled "Basic Analysis of Textual Information" and "Advanced Analysis of Textual Information"), 13, 14, 18, 19, and the case studies in the appendix. For readers who want more technical details, the following chapters are recommended: Chapters 4 through 9, 11, 12, 15 through 17, 19, and the case studies in the appendix.

The demographic data referred to in this book has been randomly anonymized and aggregated; that is, individual people cannot be identified by first and last names or by any unique or derivable identifier. Any resemblance to a real person or entity is purely coincidental.

Chapter 2

Business Objectives

INTRODUCTION

This chapter discusses the definition of a data mining project, including its initial concept, motivation, objective, viability, estimated costs, and expected benefit (returns). Key considerations are defined, and a way of quantifying the cost and benefit is presented in terms of the factors that most influence the project. Two case studies illustrate how the cost/benefit evaluation can be applied for real-world projects.

A commercial data analysis project that lives up to its expectations will probably do so because sufficient time was dedicated at the outset to defining the project's business objectives. What is meant by business objectives? The following are some examples:

- Reduce the loss of existing customers by 3 percent.
- Augment the contract signings of new customers by 2 percent.
- Augment the sales from cross-selling products to existing customers by 5 percent.
- Predict the television audience share with a probability of 70 percent.
- Predict, with a precision of 75 percent, which clients are most likely to contract a new product.
- Identify new categories of clients and products.
- Create a new customer segmentation model.

The first three examples define a specific percentage of precision and improvement as part of the objective.

Business Objective

Assigning a Value for Percent Improvement

The percentage improvement should always be considered with regard to the current precision of an existing index as a baseline. Also, the new precision objective should not get lost in the error bars of the current precision. That is, if the current precision has an error margin of ± 3% in its measurement or calculation, this should be taken into account.

In the fourth and fifth examples, an absolute value is specified for the desired precision for the data model. In the final two examples the desired improvement is not quantified; instead, the objective is expressed in qualitative terms.

CRITERIA FOR CHOOSING A VIABLE PROJECT

This section enumerates some main issues and poses some key questions relevant to evaluating the viability of a potential data mining project. The checklists of general and specific considerations provided here are the bases for the rest of the chapter, which enters into a more detailed specification of benefit and cost criteria and applies these definitions to two case studies.

Evaluation of Potential Commercial Data Analysis Projects – General Considerations

The following is a list of questions to ask when considering a data analysis project:

- Is data available that is consistent and correlated with the business objectives?
- What is the capacity for improvement with respect to the current methods? (The greater the capacity for improvement, the greater the economic benefit.)
- Is there an operational business need for the project results?
- Can the problem be solved by other techniques or methods? (If the answer is no, the profitability return on the project will be greater.)
- Does the project have a well-defined scope? (If this is the first instance of a project of this type, reducing the scale of the project is recommended.)

Evaluation of Viability in Terms of Available Data – Specific Considerations

The following list provides specific considerations for evaluating the viability of a data mining project in terms of the available data:

- Does the necessary data for the business objectives exist, and does the business have access to it?
- If part or all of the data does not exist, can processes be defined to capture or obtain it?
- What is the coverage of the data with respect to the business objectives?
- What is the availability of a sufficient volume of data over a required period of time, for all clients, product types, sales channels, and so on? (The data should cover all the business factors to be analyzed and modeled. The historical data should cover the current business cycle.)

- Is it necessary to evaluate the quality of the available data in terms of reliability? (The reliability depends on the percentage of erroneous data and incomplete or missing data. The ranges of values must be sufficiently wide to cover all cases of interest.)
- Are people available who are familiar with the relevant data and the operational processes that generate the data?

FACTORS THAT INFLUENCE PROJECT BENEFITS

There are several factors that influence the benefits of a project. A qualitative assessment of current functionality is first required: what is the current grade of satisfaction of how the task is being done? A value between 1 and 0 is assigned, where 1 is the highest grade of satisfaction and 0 is the lowest, where the lower the current grade of satisfaction, the greater the improvement and, consequently, the benefit, will be.

The potential quality of the result (the evaluation of future functionality) can be estimated by three aspects of the data: coverage, reliability, and correlation:

- The coverage or completeness of the data, assigned a value between 0 and 1, where 1 indicates total coverage.
- The quality or reliability of the data, assigned a value between 0 and 1, where 1 indicates the highest quality. (Both the coverage and the reliability are normally measured variable by variable, giving a total for the whole dataset. Good coverage and reliability for the data help to make the analysis a success, thus giving a greater benefit.)
- The correlation between the data and its grade of dependence with the business objective can be statistically measured. A correlation is typically measured as a value from –1 (total negative correlation) through 0 (no correlation) to 1 (total positive correlation). For example, if the business objective is that clients buy more products, the correlation would be calculated for each customer variable (age, time as a customer, zip code of postal address, etc.) with the customer's sales volume.

Once individual values for coverage, reliability, and correlation are acquired, an estimation of the future functionality can be obtained using the formula:

$$\text{Future functionality} = (\text{correlation} + \text{reliability} + \text{coverage})/3$$

An estimation of the possible improvement is then determined by calculating the difference between the current and the future functionality, thus:

$$\text{Estimated improvement} = \text{Future functionality} - \text{Current functionality}$$

A fourth aspect, volatility, concerns the amount of time the results of the analysis or data modeling will remain valid.

Volatility of the environment of the business objective can be defined as a value of between 0 and 1, where 0 = minimum volatility and 1 = maximum volatility. A high volatility can cause models and conclusions to become quickly out of date with respect to the data; even the business objective can lose relevance. Volatility depends on whether the results are applicable over the long, medium, or short terms with respect to the business cycle.

Note that this *a priori* evaluation gives an idea for the viability of a data mining project. However, it is clear that the quality and precision of the end result will also depend on how well the project is executed: analysis, modeling, implementation, deployment, and so on. The next section, which deals with the estimation of the cost of the project, includes a factor (expertise) that evaluates the availability of the people and skills necessary to guarantee the *a posteriori* success of the project.

FACTORS THAT INFLUENCE PROJECT COSTS

There are numerous factors that influence how much a project costs. These include:

- Accessibility: The more data sources, the higher the cost. Typically, there are at least two different data sources.
- Complexity: The greater the number of variables in the data, the greater the cost. Categorical-type variables (zones, product types, etc.) must especially be taken into account, given that each variable may have many possible values (for example, 50). On the other hand, there could be just 10 other variables, each of which has only two possible values.
- Data volumes: The more records there are in the data, the higher the cost. A data sample extracted from the complete dataset can have a volume of about 25,000 records, whereas the complete database could contain between 250,000 and 10 million records.
- Expertise: The more expertise available with respect to the data, the lower the cost. Expertise includes knowledge about the business environment, customers, and so on that facilitates the interpretation of the data. It also includes technical know-how about the data sources and the company databases from which the data is extracted.

EXAMPLE 1: CUSTOMER CALL CENTER – OBJECTIVE: IT SUPPORT FOR CUSTOMER RECLAMATIONS

Mr. Strong is the operations manager of a customer call center that provides outsourced customer support for a diverse group of client companies. In the last quarter, he has detected an increase of reclamations by customers for erroneous billing by a specific company. By revising the bills and speaking with the client company, the telephone operators identified a defective batch software program in the batch billing process and reported the incident to Mr. Strong, who, together with the IT manager of the client company, located the defective process. He determined the origin of the problem, and the IT manager gave instructions to the IT department to make the necessary corrections to the billing software. The complete process, from identifying the incident to the corrective actions, was documented in the call center's audit trail and the client company. Given the concern for the increase in incidents, Mr. Strong and the IT manager decided to initiate a data mining project to efficiently investigate reclamations due to IT processing errors and other causes.

Hypothetical values can be assigned to the factors that influence the benefit of this project, as follows: The available data has a high grade of correlation (0.9) with the business objective. Sixty-two percent of the incidents (which are subsequently identified as IT processing issues) are solved by the primary corrective actions; thus, the current grade of precision is 0.62. The data captured represents 85 percent of the modifications made to the IT processes, together with the relevant information at the time of the incident. The incidents, the corrections, and the effect of applying the corrections are entered into a spreadsheet, with a margin of error or omission of 8 percent. Therefore, the degree of coverage is 0.85 and the grade of reliability is (1 – 0.08) = 0.92.

The client company's products and services that the call center supports have to be continually updated due to changes in their characteristics. This means that 10 percent of the products and services change completely over a one year period. Thus a degree of volatility of 0.10 is assigned. The project quality model, in terms of the factors related to benefit, is summarized as follows:

- Qualitative measure of the current functionality: 0.62 (medium)
- Evaluation of future functionality:
  - Coverage: 0.85 (high)
  - Reliability: 0.92 (high)
  - Correlation of available data with business objective: 0.9 (high)
- Volatility of the environment of the business objective: 0.10 (low)

Values can now be assigned for factors that influence the cost of the project.

Mr. Strong's operations department has an Oracle database that stores the statistical summaries of customer calls. Other historical records are kept in an Excel spreadsheet for the daily operations, diagnostics arising from reclamations, and corrective actions. Some of the records are used for operations monitoring. The IT manager of the client company has a DB2 database of software maintenance that the IT department has performed. Thus there are three data sources: the Oracle database, the data in the call center's Excel spreadsheets, and the DB2 database from the client IT department.

There are about 100 variables represented in the three data sources, 25 of which the operations manager and the IT manager consider relevant for the data model. Twenty of the variables are numerical and five are categorical (service type, customer type, reclamation type, software correction type, and priority level). Note that the correlation value used to estimate the benefit and the future functionality is calculated as an average for the subset of the 25 variables evaluated as being the most relevant, and not the 100 original variables.

The operations manager and the IT manager agree that, with three years' worth of historical data, the call center reclamations and IT processes can be modeled. It is clear that the business cycle does not have seasonal cycles; however, there is a temporal aspect due to peaks and troughs in the volume of customer calls at certain times of the year. Three years' worth of data implies about 25,000 records from all three data sources. Thus the data volume is 25,000.

The operations manager and the IT manager can make time for specific questions related to the data, the operations, and the IT processes. The IT manager may also dedicate time to technical interpretation of the data in order to extract the required data from the data sources. Thus there is a high level of available expertise in relation to the data.

Factors that influence the project costs include:

- Accessibility: three data sources, with easy accessibility
- Complexity: 25 variables
- Data volume: 25,000 records
- Expertise: high

Overall Evaluation of the Cost and Benefit of Mr. Strong's Project

In terms of benefit, the evaluation gives a quite favorable result, given that the current functionality (0.62) is medium, thus giving a good margin for improvement on the current precision. The available data for the model has a high level of coverage of the environment (0.85) and is very reliable (0.92); these two factors are favorable for the success of the project. The correlation of the data with the business objective is high (0.9), again favorable, and a low volatility (0.10) will prolong the useful life of the data model. Using the formula defined earlier (factors that influence the benefit of a project), the future functionality is estimated by taking the average of the correlation, reliability, and coverage $(0.9+0.92+0.85)/3=0.89$, and subtracting the current precision (0.62), which gives an estimated improvement of 0.27, or 27 percent. Mr. Strong can interpret this percentage in terms of improvement of the operations process or he can convert it into a monetary value.

In terms of cost, there is reasonable accessibility to the data, since there are only three data sources. However, as the Oracle and DB2 databases are located in different companies (the former in the call center and the latter in the client company), the possible costs of unifying any necessary data will have to be evaluated in more detail. The complexity of having 25 descriptive variables is considered as medium; however, the variables will have to be studied individually to see if there are many different categories and whether new factors need to be derived from the original variables. The data volume (25,000 records) is medium-high for this type of problem. In terms of expertise, the participating managers have good initial availability, although they will need to commit their time given that, in practice, the day-to-day workings of the call center and IT department may reduce their dedication. The project would have a medium level of costs.

As part of the economic cost of the project, two factors must be taken into account: the services of an external consultant specializing in data analysis, and the time dedicated by the call center's own employees (Mr. Strong, the operations manager; the IT manager; a call center operator; and a technician from the IT department). Also, for a project with the given characteristics and medium complexity, renting or purchasing a data analysis software tool is recommended. With a benefit of 27 percent and a medium cost level, it is recommended that Mr. Strong go ahead with his operations model project.

EXAMPLE 2: ONLINE MUSIC APP – OBJECTIVE: DETERMINE EFFECTIVENESS OF ADVERTISING FOR MOBILE DEVICE APPS

Melody-online is a new music streaming application for mobile devices (iPhone, iPad, Android, etc.). The commercial basis of the application is to have users pay for a premium account with no publicity, or have users connect for free but with publicity inserted before the selected music is played. The company's application was previously available only on non-mobile computers (desktop, laptop, etc.), and the company now wishes to evaluate the effectiveness of advertising in this new environment. There is typically a minimum time when non-paying users cannot deactivate the publicity, after which they can switch it off and the song they selected starts to play. Hence, Melody-online wishes to evaluate whether the listening time for users of the mobile device app is comparable to the listening time for users of the fixed computer applications. The company also wishes to study the behavior of mobile device app users in general by incorporating new types of information, such as geolocation data.

Values are assigned to the factors that influence the benefit of this project. The available data has a high grade of correlation (0.9) with the business objectives. Fifty percent of users are currently categorized in terms of the available data, thus the current grade of precision is 0.50.

The data available represents 100 percent of users, but only six months of data is available for the mobile app, whereas five years' worth of data has been accumulated for the non-mobile app. A minimum of two years' worth of data is needed to cover all user types and behaviors, hence only a quarter of the required data is available. User data is automatically registered by cookies and then sent to the database, with a margin of error or omission of 5 percent. Therefore, the degree of coverage is 0.25 and the grade of reliability is $(1 - 0.05) = 0.95$.

The music genres and artists that Melody-online has available have to be continually updated for changing music tendencies and new artists. This means 25 percent of the total music offering changes completely over a one-year period. Thus a degree of volatility of 0.25 is assigned. The project quality model, in terms of the factors related to benefit, are summarized as follows:

- Qualitative measure of the current functionality: 0.50 (medium-low)
- Evaluation of future functionality:
  - Coverage: 0.25 (low)
  - Reliability: 0.95 (high)
  - Correlation of available data with business objective: 0.9 (high)
- Volatility of the environment of the business objective: 0.25 (medium)

Values can now be assigned to factors that influence the cost of the project.

Melody-online maintains an Access database containing statistical summaries of user sessions and activities. Some records are transferred to an Excel spreadsheet and are used for management monitoring. Thus, there are two data sources: the Access database and the Excel spreadsheets.

There are about 40 variables represented in the two data sources, 15 of which the marketing manager considers relevant for the data model. Ten of the variables are numerical (average ad listening time, etc.) and five are categorical (user type, music type, ad type, etc.). As with the previous case study, note that the correlation value used to estimate the benefit and the future functionality is calculated as an average for the subset of 15 variables, not for the 40 original variables.

The IT manager and the marketing manager agree that user behavior can be modeled with two years' worth of historical data, taking into account the characteristics of the business cycle. This much data implies about 500 thousand user sessions with an average of 20 records per session, totaled from both data sources. Thus the data volume is 10 million data records for the 2 year time period considered.

The IT manager and the marketing director can make some time for specific questions related to the data and the production process, but the IT manager has very limited time available. (The marketing manager is the main driver behind the project.) Thus there is a medium level of available expertise in relationship to the data.

Factors that influence the project costs include:

- Accessibility: two data sources, with easy accessibility
- Complexity: 15 variables
- Data volume: 10 million records
- Expertise: medium

Overall Evaluation of the Cost and Benefit of Melody-online's Project

In terms of benefit, the evaluation gives a quite favorable result, given that the current functionality (0.50) is at a medium-low level, thus giving a good margin for improvement on the current precision. The available data is very reliable

(0.95); however, the low coverage of the environment (0.25) is a major drawback. The values for these two factors are critical for the project's success. The correlation of the data with the business objective is high (0.9), which is favorable, but a medium volatility (0.25) will reduce the useful life of the analysis results. The future functionality is estimated by taking the average of the correlation, reliability, and coverage (0.9 + 0.95 + 0.25)/3 = 0.7), and subtracting the current precision (0.50) which gives an estimated improvement of 0.2, or 20 percent. Melody-online can interpret this percentage in terms of users having increased exposure times to advertising, or it can convert the percentage into a monetary value (e.g., advertising revenues).

In terms of cost, there is good accessibility to the data, given that there are only two data sources. However, there is a serious problem with low data coverage, with only 25 percent of the required business period covered. The complexity of having 15 descriptive variables is considered medium-low, but the variables will have to be studied individually to see if there are many different categories, and whether new factors must be derived from the original variables. The data volume, at 10 million records, is high for this type of problem. In terms of expertise, the IT manager stated up front that he will not have much time for the project, so there is a medium initial availability. The project would have a medium-high level of costs.

Renting or purchasing a data analysis software tool for a project with these characteristics is recommended. As part of the economic cost of the project, the services of an external consultant for the data analysis tool and the time dedicated by the company's employees (the IT manager and the marketing manager) must be taken into account.

With a benefit of 20 percent, a medium-high cost level, the lack of the IT manager's availability, and especially the lack of available data (only 25 percent), it is recommended that Melody-online postpone its data mining project until sufficient user behavior data for its mobile app has been captured. The IT manager should delegate his participation to another member of the IT department (for example, an analyst-programmer) who has more time available to dedicate to the project.

SUMMARY

In this chapter, some detailed guidelines and evaluation criteria have been discussed for choosing a commercial data mining business objective and evaluating its viability in terms of benefit and cost. Two examples were examined that applied the evaluation criteria in order to quantify expected benefit and cost, and then the resulting information was used to decide whether to go ahead with the project. This method has been used by the author to successfully evaluate real-world data mining projects and is specifically designed for an evaluation based on the characteristics of the data and business objectives.

FURTHER READING

Boardman, A.E., Greenberg, D.H., Vining, A.R., Weimer, D.L., 2008. *Cost–Benefit Analysis*, fourth ed. Pearson Education, New Jersey, ISBN: 0132311488.

Zerbe, R.O., Bellas, A.S., 2006. *A Primer for Benefit–Cost Analysis*. Edward Elgar Publishing, Northampton, MA, ISBN: 1843768976.