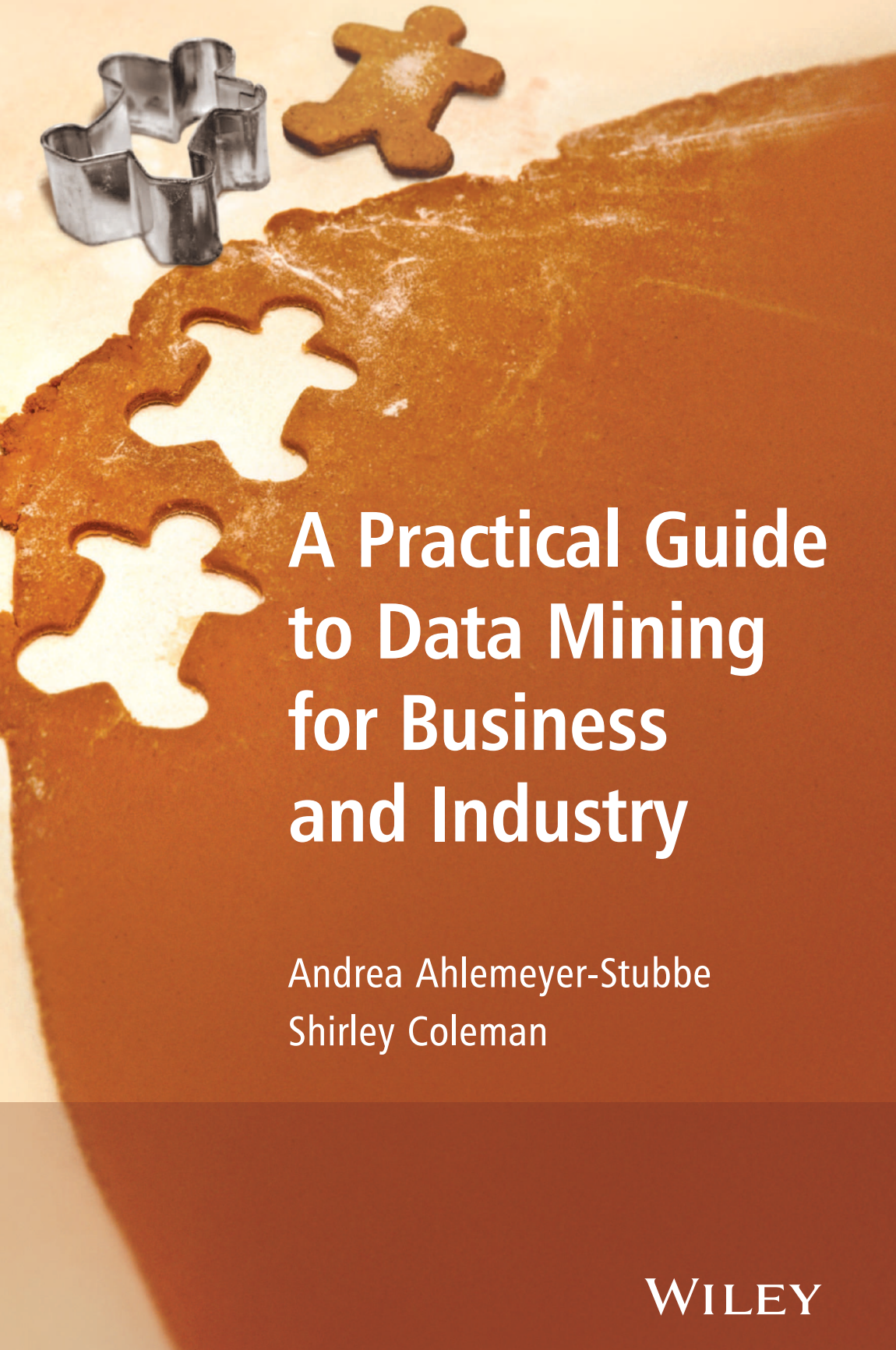

A Practical Guide to Data Mining for Business and Industry

Andrea Ahlemeyer-Stubbe
Shirley Coleman

WILEY

A Practical Guide to Data Mining for Business and Industry

A Practical Guide to Data Mining for Business and Industry

Andrea Ahlemeyer-Stubbe
Director Strategic Analytics, DRAFTFCB München GmbH, Germany

Shirley Coleman
Principal Statistician, Industrial Statistics Research Unit
School of Maths and Statistics, Newcastle University, UK

WILEY

This edition first published 2014

Registered Office
John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data

Ahlemeyer-Stubbe, Andrea.
A practical guide to data mining for business and industry / Andrea Ahlemeyer-Stubbe, Shirley Coleman.
pages cm
Includes bibliographical references and index.
ISBN 978-1-119-97713-1 (cloth)
1. Data mining. 2. Marketing–Data processing. 3. Management–Mathematical models. I. Title.
HF5415.125.A42 2014
006.3′12–dc23

2013047218

A catalogue record for this book is available from the British Library.

ISBN: 978-1-119-97713-1

Set in 10.5/13pt Minion by SPi Publisher Services, Pondicherry, India

1 2014

Contents

Glossary of terms

Accuracy | A measurement of the match (degree of closeness) between predictions and real values.

Address | A unique identifier for a computer or site online, usually a URL for a website or marked with an @ for an email address. Literally, it is how your computer finds a location on the information highway.

Advertising | Paid form of a non-personal communication by industry, business firms, non-profit organisations or individuals delivered through the various media. Advertising is persuasive and informational and is designed to influence the purchasing behaviour and thought patterns of the audience. Advertising may be used in combination with sales promotions, personal selling tactics or publicity. This also includes promotion of a product, service or message by an identified sponsor using paid-for media.

Aggregation | Form of segmentation that assumes most consumers are alike.

Algorithm | The process a search engine applies to web pages so it can accurately produce a list of results based on a search term. Search engines regularly change their algorithms to improve the quality of the search results. Hence, search engine optimisation tends to require constant research and monitoring.

Analytics | A feature that allows you to understand (learn more) a wide range of activity related to your website, your online marketing activities and direct marketing activities. Using analytics provides you with information to help optimise your campaigns, ad groups and keywords, as well as your other online marketing activities, to best meet your business goals.

API | Application Programming Interface, often used to exchange data, for example, with social networks.

Attention | A momentary attraction to a stimulus, something someone senses via sight, sound, touch, smell or taste. Attention is the starting point of the perceptual process in that attention of a stimulus will either cause someone to decide to make sense of it or reject it.

B2B | Business To Business – Business conducted between companies rather than between a company and individual consumers. For example, a firm that makes parts that are sold directly to an automobile manufacturer.

B2C | Business To Consumer – Business conducted between companies and individual consumers rather than between two companies. A retailer such as Tesco or the greengrocer next door is an example of a B2C company.

Banner | Banners are the 468-by-60 pixels ad space on commercial websites that are usually 'hotlinked' to the advertiser's site.

Banner ad | Form of Internet promotion featuring information or special offers for products and services. These small space 'banners' are interactive: when clicked, they open another website where a sale can be finalized. The hosting website of the banner ad often earns money each time someone clicks on the banner ad.

Base period | Period of time applicable to the learning data.

Behavioural targeting | Practice of targeting and ads to groups of people who exhibit similarities not only in their location, gender or age but also in how they act and react in their online environment: tracking areas they frequently visit or subscribe to or subjects or content or shopping categories for which they have registered. Google uses behavioural targeting to direct ads to people based on the sites they have visited.

Benefit | A desirable attribute of goods or services, which customers perceive that they will get from purchasing and consuming or using them. Whereas vendors sell features ('a high-speed 1cm drill bit with tungsten-carbide tip'), buyers seek the benefit (a 1cm hole).

Bias | The expected value differs from the true value. Bias can occur when measurements are not calibrated properly or when subjective opinions are accepted without checking them.

Big data | Is a relative term used to describe data that is so large in terms of volume, variety of structure and velocity of capture that it cannot be stored and analysed using standard equipment.

Blog | A blog is an online journal or 'log' of any given subject. Blogs are easy to update, manage and syndicate, powered by individuals and/or corporations and enable users to comment on postings.

BOGOF | Buy One, Get One Free. Promotional practice where on the purchase of one item, another one is given free.

Boston matrix | A product portfolio evaluation tool developed by the Boston Consulting Group. The matrix categorises products into one of four classifications based on market growth and market share.

The four classifications are as follows:

- Cash cow – low growth, high market share
- Star – high growth, high market share
- Problem child – high growth, low market share
- Dog – low growth, low market share

Brand | A unique design, sign, symbol, words or a combination of these, employed in creating an image that identifies a product and differentiates or positions it from competitors. Over time, this image becomes associated with a level of credibility, quality and satisfaction in the consumers' minds. Thus, brands stand for certain benefits and value. Legal name for a brand is trademark, and when it identifies or represents a firm, it is called a brand name. (Also see Differentiation and Positioning.)

Bundling | Combining products as a package, often to introduce other products or services to the customer. For example, AT&T offers discounts for customers by combining 2 or more of the following services: cable television, home phone service, wireless phone service and Internet service.

Buttons | Objects that, when clicked once, cause something to happen.

Buying behaviour | The process that buyers go through when deciding whether or not to purchase goods or services. Buying behaviour can be influenced by a variety of external factors and motivations, including marketing activities.

Campaign | Defines the daily budget, language, geographic targeting and location of where the ads are displayed.

Cash cow | See 'Boston matrix'.

Category management | Products are grouped and managed by strategic business unit categories. These are defined by how consumers view goods rather than by how they look to the seller, for example, confectionery could be part of either a 'food' or 'gifts' category and marketed depending on the category into which it is grouped.

Channels | The methods used by a company to communicate and interact with its customers, like direct mail, telephone and email.

Characteristic | Distinguishing feature or attribute of an item, person or phenomenon that usually falls into either a physical, functional or operational category.

Churn rate | Rate of customers lost (stopped using the service) over a specific period of time, often over the course of a year. Used to compare against new customers gained.

Click | The opportunity for a visitor to be transferred to a location by clicking on an ad, as recorded by the server.

Clusters | Customer profiles based on lifestyle, demographic, shopping behaviour or appetite for fashion. For example, ready-to-eat meals may be heavily influenced by the ethnic make-up of a store's shoppers, while beer, wine and spirits categories in the same store may be influenced predominantly by the shopper's income level and education.

Code | Anything written in a language intended for computers to interpret.

Competitions | Sales promotions that allow the consumer the possibility of winning a prize.

Competitors | Companies that sell products or services in the same marketplace as one another.

Consumer | A purchaser of goods or services at retail, or an end user not necessarily a purchaser, in the distribution chain of goods or services (gift recipient).

Contextual advertising | Advertising that is targeted to a web page based on the page's content, keywords or category. Ads in most content networks are targeted contextually.

Cookie | A file on your computer that records information such as where you have been on the World Wide Web. The browser stores this information which allows a site to remember the browser in future transactions or requests. Since the web's protocol has no way to remember requests, cookies read and record a user's browser type and IP address and store this information on the user's own computer. The cookie can be read only by a server in the domain that stored it. Visitors can accept or deny cookies by changing a setting in their browser preferences.

Coupon | A ticket that can be exchanged for a discount or rebate when procuring an item.

CRM | Customer Relationship Management – Broad term that covers concepts used by companies to manage their relationships with customers, including the capture, storage and analysis of customer, vendor, partner and internal process information. CRM is the coherent management of contacts and interactions with customers. This term is often used as if it related purely to the use of Information Technology (IT), but IT should in fact be regarded as a facilitator of CRM.

Cross-selling | A process to offer and sell additional products or services to an existing customer.

Customer |A person or company who purchases goods or services (not necessarily the end consumer).

Customer Lifetime Value (CLV) | The profitability of customers during the lifetime of the relationship, as opposed to profitability on one transaction.

Customer loyalty | Feelings or attitudes that incline a customer either to return to a company, shop or outlet to purchase there again or else to repurchase a particular product, service or brand.

Customer profile | Description of a customer group or type of customer based on various geographic, demographic, and psychographic characteristics; also called shopper profile (may include income, occupation, level of education, age, gender, hobbies or area of residence). Profiles provide knowledge needed to select the best prospect lists and to enable advertisers to select the best media

Data | Facts/figures pertinent to customer, consumer behaviour, marketing and sales activities.

Data processing | The obtaining, recording and holding of information which can then be retrieved, used, disseminated or erased. The term tends to be used in connection with computer systems and today is often used interchangeably with 'information technology'.

Database marketing | Whereby customer information, stored in an electronic database, is utilised for targeting marketing activities. Information can be a mixture of what is gleaned from previous interactions with the customer and

what is available from outside sources. (Also see 'Customer Relationship Management (CRM)'.)

Demographics | Consumer statistics regarding socio-economic factors, including gender, age, race, religion, nationality, education, income, occupation and family size. Each demographic category is broken down according to its characteristics by the various research companies.

Description | A short piece of descriptive text to describe a web page or website. With most search engines, they gain this information primarily from the metadata element of a web page. Directories approve or edit the description based on the submission that is made for a particular URL.

Differentiation | Ensuring that products and services have a unique element to allow them to stand out from the rest.

Digital marketing | Use of Internet-connected devices to engage customers with online products and service marketing/promotional programmes. It includes marketing mobile phones, iPads and other Wi-Fi devices.

Direct marketing | All activities which make it possible to offer goods or services or to transmit other messages to a segment of the population by post, telephone, email or other direct means.

Distribution | Movement of goods and services through the distribution channel to the final customer, consumer or end user, with the movement of payment (transactions) in the opposite direction back to the original producer or supplier.

Dog | See 'Boston matrix'.

Domain |A domain is the main subdivision of Internet addresses and the last three letters after the final dot, and it tells you what kind of organisation you are dealing with. There are six top-level domains widely used: .com (commercial), .edu (educational), .net (network operations), .gov (US government), .mil (US military) and .org (organisation). Other two-letter domains represent countries: .uk for the United Kingdom, .dk for Denmark, .fr for France, .de for Germany, .es for Spain, .it for Italy and so on.

Domain knowledge | General knowledge about in-depth business issues in specific industries that is necessary to understand idiosyncrasies in the data.

ENBIS | European Network of Business and Industrial Statistics.

ERP | | Enterprise Resource Planning includes all the processes around billing, logistics and real business processes.

ETL | Extraction, Transforming and Loading processes which cover all processes and algorithms that are necessary to take data from the original source to the data warehouse.

Forecast | The use of experience and/or existing data to learn/develop models that will be used to make judgments about future events and potential results. Often used interchangeably with prediction.

Forms | The pages in most browsers that accept information in text-entry fields. They can be customised to receive company sales data and orders, expense reports or other information. They can also be used to communicate.

Freeware | Shareware, or software, that can be downloaded off the Internet – for free.

Front-end applications | Interfaces and applications mainly used in customer service and help desks, especially for contacts with prospects and new customers.

ID | Unique identity code for cases or customers used internally in a database.

Index | The database of a search engine or directory.

Input or explanatory variable | Information used to carry out prediction and forecasting. In a regression, these are the X variables.

Inventory | The number of ads available for sale on a website. Ad inventory is determined by the number of ads on a page, the number of pages containing ad space and the number of page requests.

Key Success Factors (KSF) and Key Performance Indicators (KPIs) | Those factors that are a necessary condition for success in a given market. That is, a company that does poorly on one of the factors critical to success in its market is certain to fail.

Knowledge | A customer's understanding or relationship with a notion or idea. This applies to facts or ideas acquired by study, investigation, observation or experience, not assumptions or opinions.

Knowledge Management (KM) | The collection, organisation and distribution of information in a form that lends itself to practical application. Knowledge management often relies on IT to facilitate the storage and retrieval of information.

Log or log files | File that keeps track of network connections. These text files have the ability to record the amount of search engine referrals that is being delivered to your website.

Login | The identification or name used to access – log into – a computer, network or site.

Logistics | Process of planning, implementing and controlling the efficient and effective flow and storage of goods, services and related information from point of origin to point of consumption for the purpose of conforming to customer requirements, internal and external movements and return of materials for environmental purposes.

Mailing list | Online, a mailing list is an automatically distributed email message on a particular topic going to certain individuals. You can subscribe or unsubscribe to a mailing list by sending a message via email. There are many good professional mailing lists, and you should find the ones that concern your business.

Market research | Process of making investigations into the characteristics of given markets, for example, location, size, growth potential and observed attitudes.

Marketing | Marketing is the management process responsible for identifying, anticipating and satisfying customer requirements profitably.

Marketing dashboard | Any information used or required to support marketing decisions – often drawn from a computerised 'marketing information system'.

Needs | Basic forces that motivate a person to think about and do something/take action. In marketing, they help explain the benefit or satisfaction derived from a product or service, generally falling into the physical (air > water > food > sleep > sex > safety/security) or psychological (belonging > esteem > self-actualisation > synergy) subsets of Maslow's hierarchy of needs.

Null hypothesis | A proposal that is to be tested and that represents the baseline state, for example, that gender does not affect affinity to buy.

OLAP | Online Analytical Processing which is a convenient and fast way to look at business-related results or to monitor KPIs. Similar words are Management Information Systems (MIS) and Decision Support Systems (DSS).

Outlier | Outliers are unusual values that show up as very different to other values in the dataset.

Personal data | Data related to a living individual who can be identified from the information; includes any expression of opinion about the individual.

Population | All the customers or cases for which the analysis is relevant. In some situations, the population from which the learning sample is taken may necessarily differ from the population that the analysis is intended for because of changes in environment, circumstances, etc.

Precision | A measurement of the match (degree of uncertainty) between predictions and real values.

Prediction | Uses statistical models (learnt on existing data) to make assumptions about future behaviour, preferences and affinity. Prediction modelling is a main part of data mining. Often used interchangeably with forecast.

Primary key | A primary key is a field in a table in a database. Primary keys must contain unique, non-null values. If a table has a primary key defined on any field(s), then you cannot have two records having the same value of that field(s).

Probability | The chance of something happening.

Problem child | See 'Boston matrix'.

Product | Whatever the customer thinks, feels or expects from an item or idea. From a 'marketing-oriented' perspective, products should be defined by what they satisfy, contribute or deliver versus what they do or the form utility involved in their development. For example, a dishwasher cleans dishes but it's what the consumer does with the time savings that matters most. And ultimately, a dishwasher is about 'clean dishes', not the act of cleaning them.

Prospects | People who are likely to become users or customers.

Real Time | Events that happen in real time are happening virtually at that particular moment. When you chat in a chat room or send an instant message, you are interacting in real time since it is immediate.

Recession | A period of negative economic growth. Common criteria used to define when a country is in a recession are two successive quarters of falling GDP or a year-on-year fall in GDP.

Reliability | Research study can be replicated and get some basic results (free of errors).