



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
© 2014 John Wiley & Sons, Ltd

Registered Office

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

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. It is sold on the understanding that the publisher is not engaged in rendering professional services and neither the publisher nor the author shall be liable for damages arising herefrom. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

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

Contents

Glossary of terms	xii
Part I Data Mining Concept	1
1 Introduction	3
1.1 Aims of the Book	3
1.2 Data Mining Context	5
1.2.1 Domain Knowledge	6
1.2.2 Words to Remember	7
1.2.3 Associated Concepts	7
1.3 Global Appeal	8
1.4 Example Datasets Used in This Book	8
1.5 Recipe Structure	11
1.6 Further Reading and Resources	13
2 Data Mining Definition	14
2.1 Types of Data Mining Questions	15
2.1.1 Population and Sample	15
2.1.2 Data Preparation	16
2.1.3 Supervised and Unsupervised Methods	16
2.1.4 Knowledge-Discovery Techniques	18
2.2 Data Mining Process	19
2.3 Business Task: Clarification of the Business	
Question behind the Problem	20
2.4 Data: Provision and Processing of the Required Data	21
2.4.1 Fixing the Analysis Period	22
2.4.2 Basic Unit of Interest	23

2.4.3	Target Variables	24
2.4.4	Input Variables/Explanatory Variables	24
2.5	Modelling: Analysis of the Data	25
2.6	Evaluation and Validation during the Analysis Stage	25
2.7	Application of Data Mining Results and Learning from the Experience	28

Part II Data Mining Practicalities 31

3	All about Data	33
3.1	Some Basics	34
3.1.1	Data, Information, Knowledge and Wisdom	35
3.1.2	Sources and Quality of Data	36
3.1.3	Measurement Level and Types of Data	37
3.1.4	Measures of Magnitude and Dispersion	39
3.1.5	Data Distributions	41
3.2	Data Partition: Random Samples for Training, Testing and Validation	41
3.3	Types of Business Information Systems	44
3.3.1	Operational Systems Supporting Business Processes	44
3.3.2	Analysis-Based Information Systems	45
3.3.3	Importance of Information	45
3.4	Data Warehouses	47
3.4.1	Topic Orientation	47
3.4.2	Logical Integration and Homogenisation	48
3.4.3	Reference Period	48
3.4.4	Low Volatility	48
3.4.5	Using the Data Warehouse	49
3.5	Three Components of a Data Warehouse: DBMS, DB and DBCS	50
3.5.1	Database Management System (DBMS)	51
3.5.2	Database (DB)	51
3.5.3	Database Communication Systems (DBCS)	51
3.6	Data Marts	52
3.6.1	Regularly Filled Data Marts	53
3.6.2	Comparison between Data Marts and Data Warehouses	53
3.7	A Typical Example from the Online Marketing Area	54
3.8	Unique Data Marts	54
3.8.1	Permanent Data Marts	54
3.8.2	Data Marts Resulting from Complex Analysis	56

3.9	Data Mart: Do's and Don'ts	58
3.9.1	Do's and Don'ts for Processes	58
3.9.2	Do's and Don'ts for Handling	58
3.9.3	Do's and Don'ts for Coding/Programming	59
4	Data Preparation	60
4.1	Necessity of Data Preparation	61
4.2	From Small and Long to Short and Wide	61
4.3	Transformation of Variables	65
4.4	Missing Data and Imputation Strategies	66
4.5	Outliers	69
4.6	Dealing with the Vagaries of Data	70
4.6.1	Distributions	70
4.6.2	Tests for Normality	70
4.6.3	Data with Totally Different Scales	70
4.7	Adjusting the Data Distributions	71
4.7.1	Standardisation and Normalisation	71
4.7.2	Ranking	71
4.7.3	Box-Cox Transformation	71
4.8	Binning	72
4.8.1	Bucket Method	73
4.8.2	Analytical Binning for Nominal Variables	73
4.8.3	Quantiles	73
4.8.4	Binning in Practice	74
4.9	Timing Considerations	77
4.10	Operational Issues	77
5	Analytics	78
5.1	Introduction	79
5.2	Basis of Statistical Tests	80
5.2.1	Hypothesis Tests and <i>P</i> Values	80
5.2.2	Tolerance Intervals	82
5.2.3	Standard Errors and Confidence Intervals	83
5.3	Sampling	83
5.3.1	Methods	83
5.3.2	Sample Sizes	84
5.3.3	Sample Quality and Stability	84
5.4	Basic Statistics for Pre-analytics	85
5.4.1	Frequencies	85
5.4.2	Comparative Tests	88
5.4.3	Cross Tabulation and Contingency Tables	89
5.4.4	Correlations	90

5.4.5	Association Measures for Nominal Variables	91
5.4.6	Examples of Output from Comparative and Cross Tabulation Tests	92
5.5	Feature Selection/Reduction of Variables	96
5.5.1	Feature Reduction Using Domain Knowledge	96
5.5.2	Feature Selection Using Chi-Square	97
5.5.3	Principal Components Analysis and Factor Analysis	97
5.5.4	Canonical Correlation, PLS and SEM	98
5.5.5	Decision Trees	98
5.5.6	Random Forests	98
5.6	Time Series Analysis	99
6	Methods	102
6.1	Methods Overview	104
6.2	Supervised Learning	105
6.2.1	Introduction and Process Steps	105
6.2.2	Business Task	105
6.2.3	Provision and Processing of the Required Data	106
6.2.4	Analysis of the Data	107
6.2.5	Evaluation and Validation of the Results (during the Analysis)	108
6.2.6	Application of the Results	108
6.3	Multiple Linear Regression for Use When Target is Continuous	109
6.3.1	Rationale of Multiple Linear Regression Modelling	109
6.3.2	Regression Coefficients	110
6.3.3	Assessment of the Quality of the Model	111
6.3.4	Example of Linear Regression in Practice	113
6.4	Regression When the Target is Not Continuous	119
6.4.1	Logistic Regression	119
6.4.2	Example of Logistic Regression in Practice	121
6.4.3	Discriminant Analysis	126
6.4.4	Log-Linear Models and Poisson Regression	128
6.5	Decision Trees	129
6.5.1	Overview	129
6.5.2	Selection Procedures of the Relevant Input Variables	134
6.5.3	Splitting Criteria	134
6.5.4	Number of Splits (Branches of the Tree)	135
6.5.5	Symmetry/Asymmetry	135
6.5.6	Pruning	135
6.6	Neural Networks	137
6.7	Which Method Produces the Best Model? A Comparison of Regression, Decision Trees and Neural Networks	141

6.8	Unsupervised Learning	142
6.8.1	Introduction and Process Steps	142
6.8.2	Business Task	143
6.8.3	Provision and Processing of the Required Data	143
6.8.4	Analysis of the Data	145
6.8.5	Evaluation and Validation of the Results (during the Analysis)	147
6.8.6	Application of the Results	148
6.9	Cluster Analysis	148
6.9.1	Introduction	148
6.9.2	Hierarchical Cluster Analysis	149
6.9.3	K-Means Method of Cluster Analysis	150
6.9.4	Example of Cluster Analysis in Practice	151
6.10	Kohonen Networks and Self-Organising Maps	151
6.10.1	Description	151
6.10.2	Example of SOMs in Practice	152
6.11	Group Purchase Methods: Association and Sequence Analysis	155
6.11.1	Introduction	155
6.11.2	Analysis of the Data	157
6.11.3	Group Purchase Methods	158
6.11.4	Examples of Group Purchase Methods in Practice	158
7	Validation and Application	161
7.1	Introduction to Methods for Validation	161
7.2	Lift and Gain Charts	162
7.3	Model Stability	164
7.4	Sensitivity Analysis	167
7.5	Threshold Analytics and Confusion Matrix	169
7.6	ROC Curves	170
7.7	Cross-Validation and Robustness	171
7.8	Model Complexity	172

Part III Data Mining in Action 173

8	Marketing: Prediction	175
8.1	Recipe 1: Response Optimisation: To Find and Address the Right Number of Customers	176
8.2	Recipe 2: To Find the $x\%$ of Customers with the Highest Affinity to an Offer	186
8.3	Recipe 3: To Find the Right Number of Customers to Ignore	187

8.4	Recipe 4: To Find the $x\%$ of Customers with the Lowest Affinity to an Offer	190
8.5	Recipe 5: To Find the $x\%$ of Customers with the Highest Affinity to Buy	191
8.6	Recipe 6: To Find the $x\%$ of Customers with the Lowest Affinity to Buy	192
8.7	Recipe 7: To Find the $x\%$ of Customers with the Highest Affinity to a Single Purchase	193
8.8	Recipe 8: To Find the $x\%$ of Customers with the Highest Affinity to Sign a Long-Term Contract in Communication Areas	194
8.9	Recipe 9: To Find the $x\%$ of Customers with the Highest Affinity to Sign a Long-Term Contract in Insurance Areas	196
9	Intra-Customer Analysis	198
9.1	Recipe 10: To Find the Optimal Amount of Single Communication to Activate One Customer	199
9.2	Recipe 11: To Find the Optimal Communication Mix to Activate One Customer	200
9.3	Recipe 12: To Find and Describe Homogeneous Groups of Products	206
9.4	Recipe 13: To Find and Describe Groups of Customers with Homogeneous Usage	210
9.5	Recipe 14: To Predict the Order Size of Single Products or Product Groups	216
9.6	Recipe 15: Product Set Combination	217
9.7	Recipe 16: To Predict the Future Customer Lifetime Value of a Customer	219
10	Learning from a Small Testing Sample and Prediction	225
10.1	Recipe 17: To Predict Demographic Signs (Like Sex, Age, Education and Income)	225
10.2	Recipe 18: To Predict the Potential Customers of a Brand New Product or Service in Your Databases	236
10.3	Recipe 19: To Understand Operational Features and General Business Forecasting	241
11	Miscellaneous	244
11.1	Recipe 20: To Find Customers Who Will Potentially Churn	244
11.2	Recipe 21: Indirect Churn Based on a Discontinued Contract	249
11.3	Recipe 22: Social Media Target Group Descriptions	250

<i>Contents</i>	<i>xi</i>
11.4 Recipe 23: Web Monitoring	254
11.5 Recipe 24: To Predict Who is Likely to Click on a Special Banner	258
12 Software and Tools: A Quick Guide	261
12.1 List of Requirements When Choosing a Data Mining Tool	261
12.2 Introduction to the Idea of Fully Automated Modelling (FAM)	265
12.2.1 Predictive Behavioural Targeting	265
12.2.2 Fully Automatic Predictive Targeting and Modelling Real-Time Online Behaviour	266
12.3 FAM Function	266
12.4 FAM Architecture	267
12.5 FAM Data Flows and Databases	268
12.6 FAM Modelling Aspects	269
12.7 FAM Challenges and Critical Success Factors	270
12.8 FAM Summary	270
13 Overviews	271
13.1 To Make Use of Official Statistics	272
13.2 How to Use Simple Maths to Make an Impression	272
13.2.1 Approximations	272
13.2.2 Absolute and Relative Values	273
13.2.3 % Change	273
13.2.4 Values in Context	273
13.2.5 Confidence Intervals	274
13.2.6 Rounding	274
13.2.7 Tables	274
13.2.8 Figures	274
13.3 Differences between Statistical Analysis and Data Mining	275
13.3.1 Assumptions	275
13.3.2 Values Missing Because 'Nothing Happened'	275
13.3.3 Sample Sizes	276
13.3.4 Goodness-of-Fit Tests	276
13.3.5 Model Complexity	277
13.4 How to Use Data Mining in Different Industries	277
13.5 Future Views	283
Bibliography	285
Index	296

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 meta-data 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).