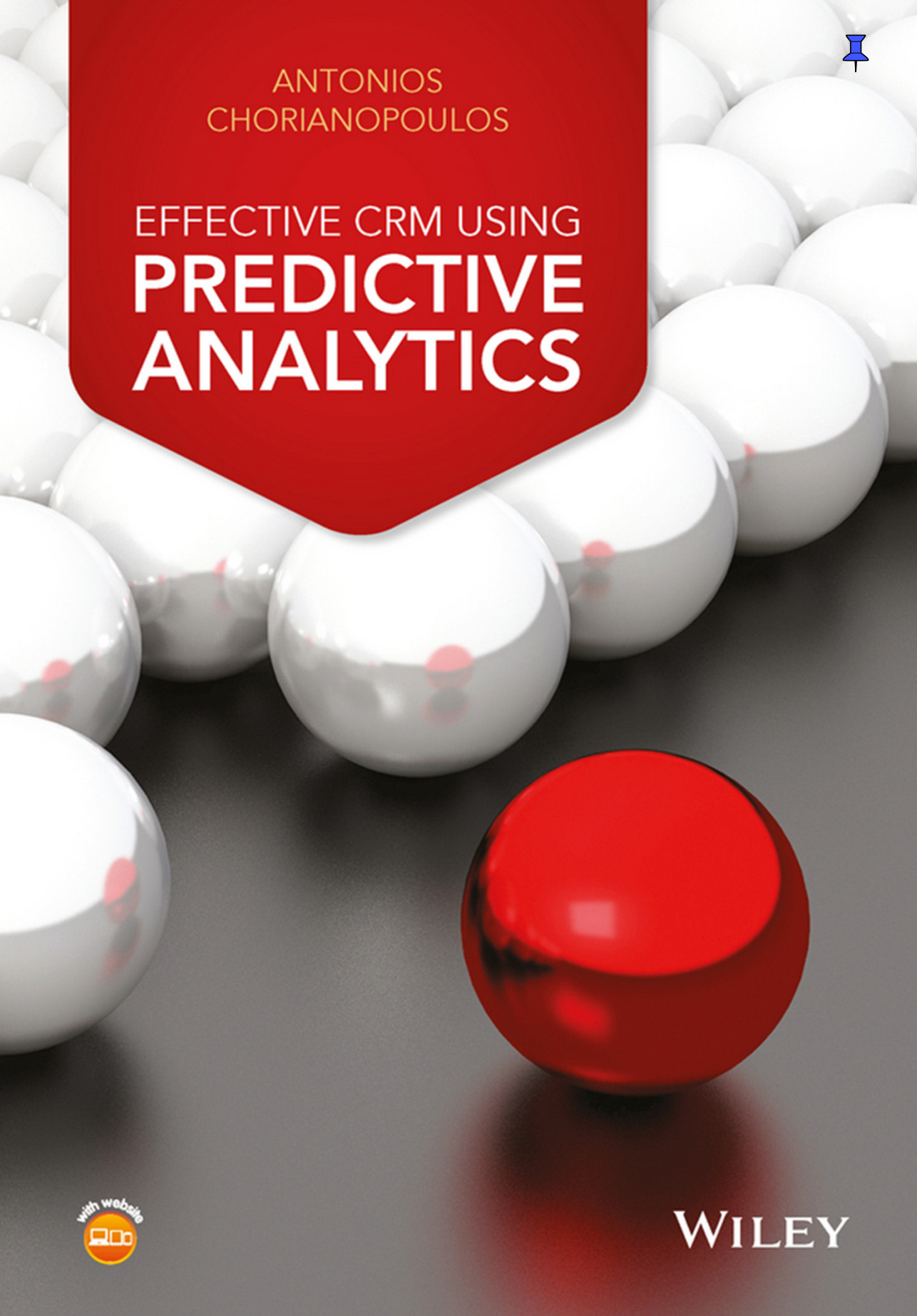
ANTONIOS
CHORIANOPOULOS
EFFECTIVE CRM USING
PREDICTIVE
ANALYTICS
with website
WILEY

Effective CRM Using Predictive Analytics

Effective CRM Using Predictive Analytics

Antonios Chorianopoulos

WILEY

This edition first published 2016

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

Library of Congress Cataloging-in-Publication Data

Chorianopoulos, Antonios.
Effective CRM using predictive analytics / Antonios Chorianopoulos.
pages cm
Includes bibliographical references and index.
ISBN 978-1-119-01155-2 (cloth)
1. Customer relations–Management–Data processing. 2. Data mining. I. Title.
HF5415.5.C4786 2015
658.8′12–dc23

2015020804

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

Cover image: Baitong333/iStockphoto

Set in 10/12pt Times by SPi Global, Pondicherry, India

1 2016

To my daughter Eugenia and my wife Virginia

Contents

Preface

This book is in a way the "sequel" of the first book that I wrote together with Konstantinos Tsiptsis. It follows the same principles, aiming to be an applied guide rather than a generic reference book on predictive analytics and data mining. There are many excellent, well-written books that succeed in presenting the theoretical background of the data mining algorithms. But the scope of this book is to enlighten the usage of these algorithms in marketing applications and to transfer domain expertise and knowledge. That's why it is packed with real-world case studies which are presented with the use of three powerful and popular software tools: IBM SPSS Modeler, RapidMiner, and Data Mining for Excel.

Here are a few words on the book's structure and some tips on "how to read the book." The book is organized in three main parts:

Part I, the Methodology. Chapters 2 and 3: I strongly believe that these sections are among the strong points of the book. Part I provides a methodological roadmap, covering both the technical and the business aspects for designing and carrying out optimized marketing actions using predictive analytics. The data mining process is presented in detail along with specific guidelines for the development of targeted acquisition, cross-/deep-/up-selling and retention campaigns, as well as effective customer segmentation schemes.

Part II, the Algorithms. Chapters 4 and 5: This part is dedicated in introducing the main concepts of some of the most popular and powerful data mining algorithms for classification and clustering. The data mining algorithms are explained in a simple and comprehensive language for business users with no technical expertise. The intention is to demystify the main concepts of the algorithms rather than "diving" deep in mathematical explanations and formulas so that data mining and marketing practitioners can confidently deploy them in their everyday business problems.

Part III, the Case Studies. Chapters 6, 7, and 8: And then it's "action time"! The third part of the book is the "hands-on" part. Three case studies from banking, retail, and telephony are presented in detail following the specific methodological steps explained in the previous chapters. The concept is to apply the methodological "blueprints" of Chapters 2 and 3 in real-world applications and to bridge the gap between analytics and their use in CRM. Given the level of detail and the accompanying material, the case studies can be considered as "application templates" for developing similar applications. The software tools are presented in that context.

In the book's companion website, you can access the material from each case study, including the datasets and the relevant code. This material is an inseparable part of the book, and I'd strongly suggest exploring and experimenting with it to gain full advantage of the book.

Those interested in segmentation and its marketing usage are strongly encouraged to look for the previous title: Konstantinos Tsiptsis and Antonios Chorianopoulos. *Data Mining Techniques in CRM: Inside Customer Segmentation*. Wiley, New York, 2009.

Finally, I would really like to thank all the readers of the first book for their warm acceptance, all those who read or reviewed the book, and all those who contacted us to share kind and encouraging words about how much they liked it. They truly inspired the creation of this new book. I really hope that this title meets their expectations.

Acknowledgments

Special thanks to Ioanna Koutrouvis and Vassilis Panagos at PREDICTA (http://www.predicta.gr) for their support.

1

An overview of data mining: The applications, the methodology, the algorithms, and the data

1.1 The applications

Customers are the most important asset of an organization. That's why an organization should plan and employ a clear strategy for customer handling. Customer relationship management (CRM) is the strategy for building, managing, and strengthening loyal and long-lasting customer relationships. CRM should be a customer-centric approach based on customer insight. Its scope should be the "personalized" handling of the customers as distinct entities through the identification and understanding of their differentiated needs, preferences, and behaviors.

CRM aims at two main objectives:

1. Customer retention through customer satisfaction
2. Customer development

Data mining can provide customer insight which is vital for these objectives and for establishing an effective CRM strategy. It can lead to personalized interactions with customers and hence increased satisfaction and profitable customer relationships through data analysis. It can offer individualized and optimized customer management throughout all the phases of the customer life cycle, from acquisition and establishment of a strong relationship to attrition prevention and win-back of lost customers. Marketers strive to get a greater market share and a greater share of their customers. In plain words, they are responsible for getting, developing, and keeping the customers. Data mining can help them in all these tasks, as shown in Figure 1.1.

Effective CRM using Predictive Analytics, First Edition. Antonios Chorianopoulos.

Companion website: www.wiley.com/go/chorianopoulos/effective_crm

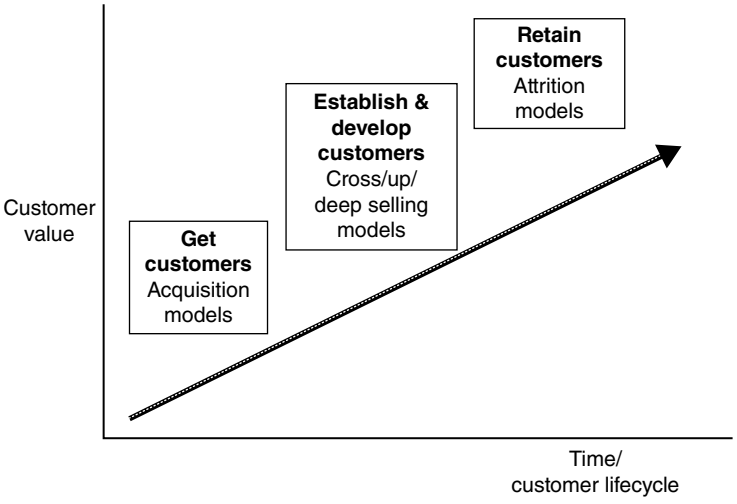

Figure 1.1 Data mining and customer life cycle management. Source: Tsiptsis and Chorianopoulos (2009). Reproduced with permission from Wiley

More specifically, the marketing activities that can be supported with the use of data mining include:

Customer segmentation

Segmentation is the process of dividing the customer base in distinct and homogeneous groups in order to develop differentiated marketing strategies according to their characteristics. There are many different segmentation types according to the specific criteria/attributes used for segmentation. In behavioral segmentation, customers are grouped based on behavioral and usage characteristics. Although behavioral segments can be created using business rules, this approach has inherent disadvantages. It can handle only a few segmentation fields, and its objectivity is questionable as it is based on the personal perceptions of a business expert. Data mining on the other hand can create data-driven behavioral segments. Clustering algorithms can analyze behavioral data, identify the natural groupings of customers, and suggest a grouping founded on observed data patterns. Provided it is properly built, it can uncover groups with distinct profiles and characteristics and lead to rich, actionable segmentation schemes with business meaning and value.

Data mining can also be used for the development of segmentation schemes based on the current or expected/estimated value of the customers. These segments are necessary in order to prioritize the customer handling and the marketing interventions according to the importance of each customer.

Direct marketing campaigns

Marketers carry out direct marketing campaigns to communicate a message to their customers through mail, Internet, e-mail, telemarketing (phone), and other direct channels in order to prevent churn (attrition) and drive customer acquisition and purchase of add-on products. More specifically, acquisition campaigns aim at drawing new and potentially valuable customers from the competition. Cross/deep/up-selling campaigns are rolled out

to sell additional products, more of the same product, or alternative but more profitable products to the existing customers. Finally, retention campaigns aim at preventing valuable customers from terminating their relationship with the organization.

These campaigns, although potentially effective, when not refined can also lead to a huge waste of resources and to the annoyance of customers with unsolicited communication. Data mining and classification (propensity) models in particular can support the development of targeted marketing campaigns. They analyze the customer characteristics and recognize the profile of the target customers. New cases with similar profiles are then identified, assigned a high propensity score, and included in the target lists. Table 1.1 summarizes the use of data mining models in direct marketing campaigns.

When properly built, propensity models can identify the right customers to contact and lead to campaign lists with increased concentrations of target customers. They outperform random selections as well as predictions based on business rules and personal intuitions.

Table 1.1 Data mining models and direct marketing campaigns

Business objective	**Marketing campaign**	**Data mining models**
Getting customers	• Acquisition: finding new customers and expanding the customer base with new and potentially profitable customers	• Acquisition classification models can be used to recognize potentially profitable prospect customers by finding "clones" of existing valuable customers in lists of contacts
Developing customers	• Cross selling: promoting and selling additional products/services to existing customers • Up selling: offering and switching customers to premium products, other products more profitable than the ones that already have • Deep selling: increasing usage of the products/ services that customers already have	• Cross/up/deep-selling classification models can reveal the existing customers with purchase potentials
Retaining customers	• Retention: prevention of voluntary churn, with priority given to presently or potentially valuable customers at risk	• Voluntary attrition (churn) models can identify early churn signals and discern the customers with increased likelihood of voluntary churn

Source: Tsiptsis and Chorianopoulos (2009).

Market basket and sequence analysis Data mining and association models in particular can be used to identify related products, typically purchased together. These models can be used for market basket analysis and for the revealing of bundles of products/services that can be sold together. Sequence models take into account the order of actions/purchases and can identify sequences of events.

1.2 The methodology

The modeling phase is just one phase in the implementation process of a data mining project. Steps of critical importance precede and follow the model building and have a significant effect in the success of the project. An outline of the basic phases in the development of a data mining project, according to the Cross Industry Standard Process for Data Mining (CRISP-DM) process model, is presented in Table 1.2.

Data mining projects are not simple. They may end in business failure if the engaged team is not guided by a clear methodological framework. The CRISP-DM process model charts the steps that should be followed for successful data mining implementations. These steps are:

Business understanding. The data mining project should start with the understanding of the business objective and the assessment of the current situation. The project's parameters should be considered, including resources and limitations. The business objective should be translated to a data mining goal. Success criteria should be defined and a project plan should be developed.

Data understanding. This phase involves considering the data requirements for properly addressing the defined goal and an investigation on the availability of the required data. This phase also includes an initial data collection and exploration with summary statistics and visualization tools to understand the data and identify potential problems of availability and quality.

Data preparation. The data to be used should be identified, selected, and prepared for inclusion in the data mining model. This phase involves the data acquisition, integration, and formatting according to the needs of the project. The consolidated data should then be cleaned and properly transformed according to the requirements of the algorithm to be applied. New fields such as sums, averages, ratios, flags, etc. should be derived from the original fields to enrich the customer information, better summarize the customer characteristics, and therefore enhance the performance of the models.

Modeling. The processed data are then used for model training. Analysts should select the appropriate modeling technique for the particular business objective. Before the training of the models and especially in the case of predictive modeling, the modeling dataset should be partitioned so that the model's performance is evaluated on a separate validation dataset. This phase involves the examination of alternative modeling algorithms and parameter settings and a comparison of their performance in order to find the one that yields the best results. Based on an initial evaluation of the model results, the model settings can be revised and fine-tuned.

Evaluation. The generated models are then formally evaluated not only in terms of technical measures but, more importantly, in the context of the business success criteria set in the business understanding phase. The project team should decide whether the

Table 1.2 The CRISP-DM phases

1. Business understanding • Understanding of the business goal • Situation assessment • Translating the business goal to a data mining objective • Development of a project plan	**2. Data understanding** • Considering data requirements • Initial data collection/ exploration and quality assessment	**3. Data preparation** • Selection of required data • Data acquisition • Data integration and formatting (merge/joins, aggregations) • Data cleaning • Data transformations and enrichment (regrouping/ binning of existing fields, creation of derived attributes, and KPIs: ratios, flag fields, averages, sums, etc.)
4. Modeling • Selection of the appropriate modeling technique • Especially in the case of predictive models, splitting of the dataset into training and testing subsets for evaluation purposes • Development and examination of alternative modeling algorithms and parameter settings • Fine-tuning of the model settings according to an initial assessment of the model's performance	**5. Model evaluation** • Evaluation of the models in the context of the business success criteria • Model approval	**6. Deployment** • Create a report of findings • Planning and development of the deployment procedure • Deployment of the data mining model • Distribution of the model results and integration in the organization's operational CRM system • Development of a maintenance–update plan • Review of the project • Planning of next steps

Source: Tsiptsis and Chorianopoulos (2009). Reproduced with permission from Wiley.

results of a given model properly address the initial business objectives. If so, this model is approved and prepared for deployment.

Deployment. The project's findings and conclusions are summarized in a report, but this is hardly the end of the project. Even the best model will turn out to be a business failure if its results are not deployed and integrated in the organization's everyday marketing operations. A procedure should be designed and developed that will enable the scoring of customers and the update of the results. The deployment procedure should also enable the distribution of the model results throughout the enterprise and their incorporation in the organization's data warehouse and operational CRM system. Finally, a maintenance plan should be designed and the whole process should be reviewed. Lessons learned should be taken into account and next steps should be planned.

The aforementioned phases present strong dependencies, and the outcomes of a phase may lead to revisiting and reviewing the results of preceding phases. The nature of the process is cyclical since the data mining itself is a never-ending journey and quest, demanding continuous reassessment and update of completed tasks in the context of a rapidly changing business environment.

This book contains two chapters dedicated in the methodological framework of classification and behavioral segmentation modeling. In these chapters, the recommended approach for these applications is elaborated and presented as a step-by-step guide.

1.3 The algorithms

Data mining models employ statistical or machine-learning algorithms to identify useful data patterns and understand and predict behaviors. They can be grouped in two main classes according to their goal:

1. **Supervised/predictive models**
   In supervised, also referred to as predictive, directed, or targeted, modeling, the goal is to predict an event or estimate the values of a continuous numeric attribute. In these models, there are input fields and an output or target field. Inputs are also called predictors because they are used by the algorithm for the identification of a prediction function for the output. We can think of predictors as the "X" part of the function and the target field as the "Y" part, the outcome.

   The algorithm associates the outcome with input data patterns. Pattern recognition is "supervised" by the target field. Relationships are established between the inputs and the output. An input–output "mapping function" is generated by the algorithm that associates predictors with the output and permits the prediction of the output values, given the values of the inputs.

2. **Unsupervised models**
   In unsupervised or undirected models, there is no output, just inputs. The pattern recognition is undirected; it is not guided by a specific target field. The goal of the algorithm is to uncover data patterns in the set of inputs and identify groups of similar cases, groups of correlated fields, frequent itemsets, or anomalous records.

1.3.1 Supervised models

Models learn from past cases. In order for predictive algorithms to associate input data patterns with specific outcomes, it is necessary to present them cases with known outcomes. This phase is called the training phase. During that phase, the predictive algorithm builds the function that connects the inputs with the target. Once the relationships are identified and the model is evaluated and proved of satisfactory predictive power, the scoring phase follows. New records, for which the outcome values are unknown, are presented to the model and scored accordingly.

Some predictive algorithms such as regression and Decision Trees are transparent, providing an explanation of their results. Besides prediction, these algorithms can also be used for insight and profiling. They can identify inputs with a significant effect on the target attribute, for example, drivers of customer satisfaction or attrition, and they can reveal the type and the magnitude of their effect.

According to their scope and the measurement level of the field to be predicted, supervised models are further categorized into:

1. **Classification or propensity models**
   Classification or propensity models predict categorical outcomes. Their goal is to classify new cases to predefined classes, in other words to predict an event. The classification algorithm estimates a propensity score for each new case. The propensity score denotes the likelihood of occurrence of the target event.

2. **Estimation (regression) models**
   Estimation models are similar to classification models with one big difference. They are used for predicting the value of a continuous output based on the observed values of the inputs.

3. **Feature selection**
   These models are used as a preparation step preceding the development of a predictive model. Feature selection algorithms assess the predictive importance of the inputs and identify the significant ones. Predictors with trivial predictive power are discarded from the subsequent modeling steps.

1.3.1.1 Classification models

Classification models predict categorical outcomes by using a set of inputs and a historical dataset with preclassified data. Generated models are then used to predict event occurrence and classify unseen records. Typical examples of target categorical fields include:

- Accepted a marketing offer: yes/no
- Defaulted: yes/no
- Churned: yes/no

In the heart of all classification models is the estimation of confidence scores. These scores denote the likelihood of the predicted outcome. They are estimates of the probability of occurrence of the respective event, typically ranging from 0 to 1. Confidence scores can be translated to propensity scores which signify the likelihood of a particular target class: the propensity of a customer to churn, to buy a specific add-on product, or to default on his loan. Propensity scores allow for the rank ordering of customers according to their likelihood. This feature enables marketers to target their lists and optimally tailor their campaign sizes according to their resources and marketing objectives. They can expand or narrow their target lists on the base of their particular objectives, always targeting the customers with the relatively higher probabilities.

Popular classification algorithms include:

- **Decision Trees**. Decision Trees apply recursive partitions to the initial population. For each split (partition), they automatically select the most significant predictor, the predictor that yields the best separation in respect to the target filed. Through successive partitions, their goal is to produce "pure" subsegments, with homogeneous behavior in terms of the output. They are perhaps the most popular classification technique. Part of their popularity is because they produce transparent results that are easily interpretable, offering insight in the event under study. The produced results can have

two equivalent formats. In a rule format, results are represented in plain English, as ordinary rules:

IF (PREDICTOR VALUES) ***THEN*** (TARGET OUTCOME & CONFIDENCE SCORE)

For example:

IF (Gender = Male and Profession = White Colar and SMS_Usage > 60 messages per month) ***THEN*** Prediction = Buyer and Confidence = 0.95

In a tree format, rules are graphically represented as a tree in which the initial population (root node) is successively partitioned into terminal (leaf) nodes with similar behavior in respect to the target field.

Decision Tree algorithms are fast and scalable. Available algorithms include:

- C4.5/C5.0
- CHAID
- Classification and regression trees (CART)

- **Decision rules**. They are quite similar to Decision Trees and produce a list of rules which have the format of human understandable statements: IF (PREDICTOR VALUES) THEN (TARGET OUTCOME & CONFIDENCE SCORES). Their main difference from Decision Trees is that they may produce multiple rules for each record. Decision Trees generate exhaustive and mutually exclusive rules which cover all records. For each record, only one rule applies. On the contrary, decision rules may generate an overlapping set of rules. More than one rule, with different predictions, may hold true for each record. In that case, through an integrated voting procedure, rules are evaluated and compared or combined to determine the final prediction and confidence.

- **Logistic regression**. This is a powerful and well-established statistical algorithm that estimates the probabilities of the target classes. It is analogous to simple linear regression but for categorical outcomes. Logistic regression results have the form of continuous functions that estimate membership probabilities of the target classes:

$$\ln\left(\frac{p_j}{p_k}\right) = b_0 + \sum_i b_i X_i$$

where p_j = probability of the target class $\boldsymbol{j}$, p_k probability of the reference target class k, X_i the predictors, b_i the regression coefficients, and b_0 the intercept of the model. The regression coefficients represent the effect of predictors.

For example, in the case of a binary target denoting churn,

$$\ln\left(\frac{\text{churn probability}}{\text{no churn probability}}\right) = b_0 + b_1 \cdot \text{tenure} + b_2 \cdot \text{num of products} + \cdots$$

In order to yield optimal results, it may require special data preparation, including potential screening and transformation (optimal binning) of the predictors. It demands some statistical experience yet, provided it is built properly, it can produce stable and understandable results.

- **Neural networks**. Neural networks are powerful machine-learning algorithms that use complex, nonlinear mapping functions for estimation and classification. They

consist of neurons organized in layers. The input layer contains the predictors or input neurons. The output layer includes the target field. These models estimate weights that connect predictors (input layer) to the output. Models with more complex topologies may also include intermediate, hidden layers, and neurons. The training procedure is an iterative process. Input records, with known outcome, are presented to the network, and model prediction is evaluated in respect to the observed results. Observed errors are used to adjust and optimize the initial weight estimates. They are considered as opaque or "black box" solutions since they do not provide an explanation of their predictions. They only provide a sensitivity analysis, which summarizes the predictive importance of the input fields. They require minimum statistical knowledge but, depending on the problem, may require long processing times for training.

- **Support Vector Machine (SVM)**. SVM is a classification algorithm that can model highly nonlinear complex data patterns and avoid overfitting, that is, the situation in which a model memorizes patterns only relevant to the specific cases analyzed. SVM works by mapping data to a high-dimensional feature space in which records become more easily separable (i.e., separated by linear functions) in respect to the target categories. Input training data are appropriately transformed through nonlinear kernel functions, and this transformation is followed by a search for simpler functions, that is, linear functions, which optimally separate cases. Analysts typically experiment with different kernel functions and compare the results. Overall, SVM is an effective yet demanding algorithm, in terms of processing time and resources. Additionally, it lacks transparency since the predictions are not explained, and only the importance of predictors is summarized.

- **Bayesian networks**. Bayesian networks are statistical models based on the Bayes theorem. They are probabilistic models as they estimate the probabilities of belonging to each target class. Bayesian belief networks, in particular, are graphical models which provide a visual representation of the attribute relationships, ensuring transparency and explanation of the model rationale.

1.3.1.2 Estimation (regression) models

Estimation models, also referred to as regression models, deal with continuous numeric outcomes. By using linear or nonlinear functions, they use the input fields to estimate the unknown values of a continuous target field.

Estimation algorithms can be used to predict attributes like the following:

- The expected balance of the savings accounts of the customers of a bank in the near future
- The estimated loss given default (LGD) incurred after a customer has defaulted
- The expected revenue from a customer within a specified time period

A dataset with historical data and known values of the continuous output is required for the model training. A mapping function is then identified that associates the available inputs to the output values. These models are also referred to as regression models, after the well-known

and established statistical algorithm of *ordinary least squares regression (OLSR)*. The OLSR estimates the line that best fits the data and minimizes the observed errors, the so-called least squares line. It requires some statistical experience, and since it is sensitive to possible violations of its assumptions, it may require specific data examination and processing before building. The final model has the intuitive form of a linear function with coefficients denoting the effect of predictors to the outcome. Although transparent, it has inherent limitations that may affect its performance in complex situations of nonlinear relationships and interactions between predictors.

Nowadays, traditional regression is not the only available estimation algorithm. New techniques, with less stringent assumptions, which also capture nonlinear relationships, can also be employed to handle continuous outcomes. More specifically, *polynomial regression, neural networks, SVM, and regression trees such as CART* can also be employed for the prediction of continuous attributes.

1.3.1.3 Feature selection (field screening)

The feature selection (field screening) process is a preparation step for the development of classification and estimation (regression) models. The situation of having hundreds of candidate predictors is not an unusual case in complicated data mining tasks. Some of these fields though may not have an influence to the output that we want to predict.

The basic idea of feature selection is to use basic statistical measures to assess and quantify the relationship of the inputs to the output. More specifically, feature selection is used to:

- Assess all the available inputs and rank them according to their association with the outcome.
- Identify the key predictors, the most relevant features for classification or regression.
- Screen the predictors with marginal importance, reducing the set of inputs to those related to the target field.

Some predictive algorithms, including Decision Trees, integrate screening mechanisms that internally filter out the unrelated predictors. A preprocessing feature selection step is also available in Data Mining for Excel, and it can be invoked when building a predictive model. Feature selection can efficiently reduce data dimensionality, retaining only a subset of significant inputs so that the training time is reduced with no or insignificant loss of accuracy.

1.3.2 Unsupervised models

In unsupervised modeling, only input fields are involved. The scope is the identification of groupings and associations. Unsupervised models include:

1. **Cluster models**
   In cluster models, the groups are not known in advance. Instead, the algorithms analyze the input data patterns and identify the natural groupings of instances/cases. When new cases are scored by the generated cluster model, they are assigned into one of the revealed clusters.

2. **Association (affinity) and sequence models**
   Association and sequence models also belong to the class of unsupervised algorithms. Association models do not involve direct prediction of a single field. In fact, all fields have a double role, since they act as inputs and outputs at the same time. Association algorithms detect associations between discrete events, products, and attributes. Sequence algorithms detect associations over time.

3. **Dimensionality reduction models**
   Dimensionality reduction algorithms "group" fields into new compound measures and reduce the dimensions of data without sacrificing much of the information of the original fields.

1.3.2.1 Cluster models

Cluster models automatically detect the underlying groups of cases, the clusters. The clusters are not known in advance. They are revealed by analyzing the observed input data patterns. Clustering algorithms assess the similarity of the records/customers in respect to the clustering fields, and they assign them to the revealed clusters accordingly. Their goal is to detect groups with internal homogeneity and interclass heterogeneity.

Clustering algorithms are quite popular, and their use is widespread from data mining to market research. They can support the development of different segmentation schemes according to the clustering attributes used: behavioral, attitudinal, or demographical segmentation.

The major advantage of the clustering algorithms is that they can efficiently manage a large number of attributes and create data-driven segments. The revealed segments are not based on personal concepts, intuitions, and perceptions of the business people. They are induced by the observed data patterns, and provided they are properly built, they can lead to results with real business meaning and value. Clustering models can analyze complex input data patterns and suggest solutions that would not otherwise be apparent. They reveal customer typologies, enabling tailored marketing strategies.

Nowadays, various clustering algorithms are available which differ in their approach for assessing the similarity of the cases. According to the way they work and their outputs, the clustering algorithms can be categorized in two classes, the hard and the soft clustering algorithms. The hard clustering algorithms assess the distances (dissimilarities) of the instances. The revealed clusters do not overlap and each case is assigned to a single cluster.

Hard clustering algorithms include:

- **Agglomerative or hierarchical**. In a way, it is the "mother" of all clustering algorithms. It is called hierarchical or agglomerative since it starts by a solution where each record comprises a cluster and gradually groups records up to the point where all records fall into one supercluster. In each step, it calculates the distances between all pairs of records and groups the ones most similar. A table (agglomeration schedule) or a graph (dendrogram) summarizes the grouping steps and the respective distances. The analyst should then consult this information, identify the point where the algorithm starts to group disjoint cases, and then decide on the number of clusters to retain. This algorithm cannot effectively handle more than a few thousand cases. Thus, it cannot be directly applied in most business clustering tasks. A usual workaround is to a use it on a sample of the clustering population. However, with numerous

other efficient algorithms that can easily handle even millions of records, clustering through sampling is not considered an ideal approach.

- **K-means**. K-means is an efficient and perhaps the fastest clustering algorithm that can handle both long (many records) and wide datasets (many data dimensions and input fields). In K-means, each cluster is represented by its centroid, the central point defined by the averages of the inputs. K-means is an iterative, distance-based clustering algorithm in which cases are assigned to the "nearest" cluster. Unlike hierarchical, it does not need to calculate distances between all pairs of records. The number of clusters to be formed is predetermined and specified by the user in advance. Thus, usually a number of different solutions should be tried and evaluated before approving the most appropriate. It best handles continuous clustering fields.

- **K-medoids**. K-medoids is a K-means variant which differs from K-means in the way clusters are represented during the model training phase. In K-means, each cluster is represented by the averages of inputs. In K-medoids, each cluster is represented by an actual, representative data point instead of using the hypothetical point defined by the cluster means. This makes this algorithm less sensitive to outliers.

- **TwoStep cluster**. A scalable and efficient clustering model, based on the BIRCH algorithm, included in IBM SPSS Modeler. As the name implies, it processes records in two steps. The first step of preclustering makes a single pass of the data, and records are assigned to a limited set of initial subclusters. In the second step, initial subclusters are further grouped, into the final segments.

- **Kohonen Network/Self-Organizing Map (SOM)**. Kohonen Networks are based on neural networks, and they typically produce a two-dimensional grid or map of the clusters, hence the name SOM. Kohonen Networks usually take longer time to train than K-means and TwoStep, but they provide a different and worth trying view on clustering.

The soft clustering techniques on the other end use probabilistic measures to assign the cases to clusters with a certain probabilities. The clusters can overlap and the instances can belong to more than one cluster with certain, estimated probabilities. The most popular probabilistic clustering algorithm is *Expectation Maximization (EM) clustering*.

1.3.2.2 Association (affinity) and sequence models

Association models analyze past co-occurrences of events and detect associations and frequent itemsets. They associate a particular outcome category with a set of conditions. They are typically used to identify purchase patterns and groups of products often purchased together. Association algorithms generate rules of the following general format:

IF (ANTECEDENTS) ***THEN*** CONSEQUENT

For example:

IF (product A and product C and product E and…) ***THEN*** product B

More specifically, a rule referring to supermarket purchases might be:

IF EGGS & MILK ***&*** FRESH FRUIT ***THEN*** VEGETABLES

This simple rule, derived by analyzing past shopping carts, identifies associated products that tend to be purchased together: when eggs, milk, and fresh fruit are bought, then there is an