On the Way from a Knowledge Discovery in Databases to a Predictive Analytics

Claudia Koschtial and Carsten Felden

Abstract Business Intelligence has "decision support" as a characterizing element. Decisions are done as a selection process based on alternatives. The choice depends on prospective developments whereby those developments are predicted with uncertainty. Due to this reason, forecasts are getting more into focus of the strategic and tactical level. But forecasts, usually based on Knowledge Discovery in Databases (KDD), are limited, yet. They often produce non-adequate results, which can lead to wrong decisions. Such a forecast quality demands further research in identifying improvements to increase reliability of forecast results and its usage in practice. This chapter modifies the Knowledge Discovery in Databases to improve the forecast quality. The associated process is supplemented by further steps to enhance the analyzed data set with additional future oriented data by using the KDD markup language. First results of an evaluation implementation at a German saving and loans bank shows motivating results.

Keywords Business Intelligence • Knowledge Discovery in Databases • Forecast • Data Mining • Predictive Analytics

1 Introduction

The daily work of managers is specified by the task of decision making. Decision makers act thereby in an area of conflict between organizational structures, budget restrictions, production changes, product innovations, changing customer needs, future oriented investment decisions, efficiency of production, selection of future oriented investments, etc. Management literature suggests that such a decision making can be supported by forecast techniques. But the financial crisis has shown that the complexity and dynamics of a crisis obstruct a reliable prognosis. Models are often multilayered and a little comprehensible, so that an appropriate usage is not possible. Also practical experiences show that the computed results do

C. Koschtial (✉) • C. Felden
Technische Universität Bergakademie Freiberg, Freiberg, Germany
e-mail: claudia.koschtial@bwl.tu-freiberg.de; carsten.felden@bwl.tu-freiberg.de

J.H. Mayer, R. Quick (eds.), *Business Intelligence for New-Generation Managers*,
DOI 10.1007/978-3-319-15696-5_2

not satisfy the decision makers due to the unsatisfactory prognosis quality. This leads to the situation that small and medium-size enterprises, but also large consolidated enterprises, avoid using prognosis models. Therefore, it is the goal of this chapter to improve the quality and thus the use of forecast models by an extension of Businesses Intelligence in sense of a Predictive Analytics process within enterprises.

Abilities and talents of decision makers, who determine the quality of decisions and problem solutions, can be embedded into a socio technical system. This frames the usage of information system by individuals or functions. Within this socio technical system it is the task of decision makers to specify goals, to evaluate alternatives, and to choose between possible alternatives. This is supported by Business Intelligence (BI) systems, which are described as abilities, technologies, applications, and procedures for supporting enterprise activities to assist the understanding of the business environment [5]. The more future oriented an analysis is, the more the concept of Knowledge Discovery in Databases (KDD) and in particular the Predictive Analytics, as a component of the Businesses Intelligence, are of importance and used in favor for decision support. Predictive Analytics is a kind of data analysis for strategic management. It is a process, which defines not only the collection of mass data, but also their processing by suitable statistical/analytical methods [6]. Its success depends on the available volume and quality of relevant historical and future oriented data and their modeling and concomitantly the reliability of the model in dynamic environments [20]. This research contributes to the discussion of forecast model quality to add value to the practice of information systems and to the scientific discussion about the quality improvement of forecast models and overall Businesses Intelligence.

The chapter is organized as follows: after a presentation of Business Intelligence based forecast process, application related obstacles are outlined (Sect. 2). Section 3 presents a modification of the KDD process in order to solve the addressed challenges. Section 4 uses the proposed concept in a case study. This case study describes Customer Relationship Management of a German saving and loans bank. The chapter ends with conclusions in Sect. 5.

2 Practical Implications of a Business Intelligence Based Forecast Process

This section discusses the forecast process support by Business Intelligence concepts and deals thereby with their obstacles. Based on these findings, the research objective is specified.

2.1 Forecasts within the Decision Process

A forecast task is to meet statements about the future. The necessity for a forecast is based on the uncertainty about future developments [2]. For this purpose, individual process steps are processed in order to gain a basis for decision making. The forecast process is just one element of an entire decision making process [18]. Simon identified the steps of decision making processes framed by rational behavior. A decision maker has to collect data about future conditions or consequences of defined events. Additionally, alternatives are to be determined, which consider appropriate probabilities. Distinctions to other existing models exist in the following aspects:

1. Decision makers look for available alternatives, since search costs have to be kept small. This leads to a prioritization and finally to non-optimal results.
2. It is not possible to occupy all alternatives with the probabilities of incidence and disbursements.
3. The value of a disbursement does not have to be a scalar parameter. It can be also a combination of individual values, which are defined as a vector.

The aspects mentioned by Simon led to the finding that decision making is restricted by limitations and not rational. He derived the following process steps, which are also seen as fundamental for the Business Intelligence concept [18]:

1. Provision of information (information acquisition and problem classification).
2. Design (determination of relevant criteria of the alternative selection).
3. Selection (determination of the most effective alternative).

These three phases were supplemented by the implementation phase of Simon himself and the monitoring phase by Book [10]. In addition, during the implementation of such a decision making process in Business Intelligence systems, the entwinement between users and information system has to be considered, too [1]. This includes for example, which information are offered to the user, for which purpose, or how to extent the model of the user.

2.2 Business Intelligence in the Forecast Process

Hans Peter Luhn defined the term Businesses Intelligence already in 1958 [14]. Intensive dissemination in theory and practice found this term however by a Gartner Group study in 1996 [1]. “Business Intelligence is providing decision makers with valuable information and knowledge by leveraging a variety of sources of data as well as structure and unstructured information [...] the term Business Intelligence has been in two different ways. It is sometimes used to refer to the product of the process or the information and knowledge that are useful to organizations for their business activities and decision making. On other occasions, BI is

Table 1 Information/methods of a business intelligence within the decision process

Decision phase	Information	Business intelligence
Information acquisition	Internal and external sources for planning and consolidation measures	Extraction/transformation/loading (ETL), Data Warehouse, KDD
Design	Knowledge and Experience for problem solving	Information analysis and modeling, intelligente agents
Selection	Business and implementation know how	
Implementation	Integration of potential users	Portals, analytical CRM
Monitoring	Understanding about socio technical systems	Reporting, online analytical processing (OLAP), visualization

used to refer to the process through which organizations obtains, analyzes, and distributes such information and knowledge" [17]. The definition shows the area of conflict between tools and individuals [13]. Following Table 1 clarifies the task supporting character of a Business Intelligence solution.

The determination, selection, and usage of relevant information and presented results depend on the respective user. In particular, during the forecast process, KDD has a high impact. KDD, as "nontrivial process of identifying valid, novel, potential useful, and ultimately understandable patterns in data" [9] is used for pattern recognition in order to be able to meet statements about the future.

2.3 Problem Statement

In particular practice oriented journals discuss regularly that the quality of the used forecast models is not sufficient. Prognoses are classified marginally better than dicing cubes [16]. It refers to the fact that forecasts are only suitable for a situation determination to date [15]. The DIW marked critically in 2009 that prognosis models provide only a small information value. Such discussions lead to the trend not to use these tools in decision making any longer [7]. In sense of designs science [11], we are proposing a modified KDD process in order to improve the forecast quality.

3 Conceptual Modification of the KDD Process to Improve the Forecast Quality

The usually used analytical model for supporting the forecast process is KDD, whose processing concept contains the processing and analysis of historical data. Goal of such a processing is finding samples and connections, which assist the forecast of future developments with a sufficient probability [8].

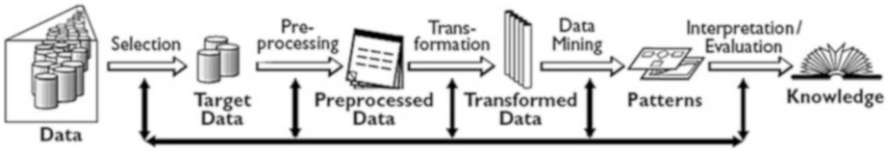

Fig. 1 KDD process model [9]

3.1 Knowledge Discovery in Databases Process Model

Independently of the procedural model discussed in the literature, individual comparable process steps can be identified within the KDD, which are shown in Fig. 1.

Data Mining, which is often used synonymously for KDD, is describing the pattern recognition in large data sets. It is usually arranged as automatic or semi-automatic process for solving problems like sales predictions [21]. It is noteworthy that the computer bound analytic methods used therein are becoming a matter of course. Thus, the current discussions turn away from explicit Data Mining procedures to more abstract terms like Predictive Analytics and due to this more into an application focus. Data Mining as such does not concentrate by definition on historical data (which is also not excluded). Earlier definitions state that Data Mining is a concept, which is able to anticipate future developments. But this is not reflected in its application; it is determined based on historical data, what happened historically. This history is applied to new data. Predictive Analytics extends this by future oriented descriptive elements explicitly, for example with data about demographic developments in context of customer analysis.

3.2 KDD Enhancements in Context of a Predictive Analytics

Analytics, also known as Business Analytics, describes the analysis and evaluation of data in context of a future oriented enterprise control [6]. The described approach of the common KDD as method of a data analysis and to the forecast implicates that there are no changes within manifested market processes and due to this in future markets. This assumption, typical for historically based approaches [12], has to be considered critically. Rather sudden shocks, as crisis charged features on the financial markets, or long term shifts of inquire relevant fundamental data (for example a demographic development), can endanger the forecast content of historical data. This development process can be e.g. clarified with age pyramids. They show substantial differences for Germany both in the intertemporal and in the interregional comparison. Essential accordingly straight for a regionally oriented enterprise is a region typical and long term anticipation of these structural changes, in order to be prepared for accompanying changes of the inventory and potential customers as well as their needs. Is generally valid: The past and due to this prognosticated values form the basis for strategic decision making, whereby data

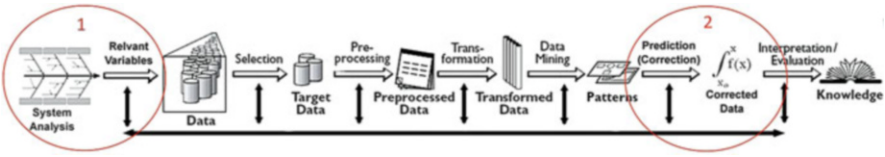

Fig. 2 Enhanced KDD process model in sense of predictive analytics

or analysis errors and/or wrong forecasts can lead to false decisions with strategic consequence. In order to reduce this risk, prognoses are to be supplemented by means of the KDD with information about intermediate as well as in the future foreseeable changes. This extension is the central thought of Predictive Analytics. An improvement in relation to the values of the classical KDD is obtained in particular by the extension and correction of the forecast model around future oriented data. In order to illustrate this correction calculation in context of an extended KDD processing concept, different possibilities exist.

Figure 2 shows that before the actual pattern analysis starts, it has to be analyzed whether the operational question, which can be answered by KDD, or if there are relevant changes of variables in relation to their initial value for the time of the emergence of the historical data (supplemental step system analysis (→ 1 in the figure)). As prerequisite step to the actual model building, all parts of a system have to be determined with their relations and their cooperation. This is a difference to the common Data Mining step. This means not just to identify unknown patterns in data, but rather the inclusion of possible well known patterns into the later forecast model (→ 2 in the figure). For this purpose, System Dynamics cause/effect diagrams [19] or Ishikawa diagrams [3] can be helpful. If there are no identified dependencies or relevant values of the subject matter period, the KDD process can proceed in its common steps. However, if changes in the period have to be considered, which can be prognosticated, a correction calculation must be done (supplemental step application of samples and correction), in order to achieve an as high prognosis quality as possible for the predicted data. To support this, the KDD Markup Language (KDDML) is used in order to realize the supplement of the model by appropriate meta data. Further probabilities can be put into the model to be able to describe certain events by more parameters, in order to describe future situations consciously.

4 Case Study: Customer Relationship Management in a German Saving and Loans

The case is based on anonymous customer data of a German savings and loans bank. The data are retrieved from their Data Warehouse and preprocessed as well as transformed according to the KDD process. During preprocessing incorrect data

records are removed or data adjustments are implemented, if necessary. During the transformation, product instances are aggregated and converted from individuals to households or numerical values such as account balances are transformed into binary values, in order to receive the information about an account usage. On the basis of the transformed data, characteristic of current customers are identified, who already decided for an investment in the financial product Deka funds. Different Data Mining algorithms such as decision trees and artificial neural networks are applied.

A decision tree consists of a root, leaves, and branches. The root contains all data records; each data record contains all data which can be evaluated. Leaves are divided in accordance with the evaluation of an attribute with each individual data record [21]. Goal of this procedure is the selection of a dividing attribute to produce leaves with homogeneous data records [4]. The application of a decision tree shows that just 7 % of the overall private customers use the Deka funds, while the ratio among the customers, who possess also a savings agreement for building purposes, is already 22 %. Even if this sequence is included, sample statements can be derived toward an increased probability of sale from Deka funds at building savers. The temporal accumulation of the product conclusions is likewise examined. It is shown that all customers, who possessed both products, bought the savings agreement for building purposes first and thereafter the purchase of the Deka funds happened.

In contrast to it, an artificial neural network works as black box to a certain extent. Are procedure calculates a result, which can be used to forecast the purchase behavior of the Deka funds [4]. Both decision trees and artificial neural networks supply appropriate results, whereby the valuation criteria remain unsettled with the latter. Both can be used parallel and be chosen according to their forecast quality. The found samples can be applied to the customer data set in order to identify customers who do not possess Deka funds so far and, based on the historical data set, have a high relevant probability of to purchase such funds.

If it is identified for each product and each customer, whether and when he/she is a buyer (or not), the Customer Lifetime Value (CLV) of the customer can be calculated. Appropriate products are offered according to the specific interests and points in time to be able to manage the relationship to the customer during his/her lifetime (in the sense of the CLV). In relation to the principle of a uniform distribution this procedure offers the advantage that an improved allocation of resources can take place, because not all customers cause the same expenditure. The first trial of the described procedure is done without a purposeful and for the relationship of the savings bank valid basic correction of the forecast data. At this point, the principle of the Predictive Analytics sets.

The examined enhancement of common KDD to a Predictive Analytics can be compared on the basis of a CLV computation. With its accuracy, the necessity decreases of uncertainty conditioned adjustments and increases in response to the reliability of a CLV as decision basis for the Customer Relations Management. Depending upon demarcation of the forecast the received values can also affect the strategic business development. On this basis, a former campaign of the example

Table 2 Result comparison of the common and the modified KDD process

Customer ID	E	F	L	Prog_ Procukt_ A	Prog_ Dat_ A	Prog_ Procukt_ B	Prog_ Dat_ B	CLV	CLV-corrected
1399XXX	E	M	L	0		0		234.34	107.89
1428XXX	E	F	N	1	2012	0		5,043.34	5,043.34
1445XXX	E	M	N	1	2015	1	2011	10,003.23	8,763.99
1450XXX	E	F	W	0		0		−645.00	−623.45
1451XXX	G		N	0		0		0	0
1477XXX	E	F	L	1	2010	1	2012	1,534.20	1,534.20
1494XXX	E	M	N	0		1	2013	8,674.5	8,455.40
202576XXX	E	F	N	0		1	2010	375.44	375.44
1968711XXX	G		N	0		0		198.45	234.48
1517XXX	E	M	V	0		0		0	0
1525XXX	E	M	N	0		0		0	0
157XXX	E	M	N	1	2015	0		9,700.56	8,951.67

bank was done with the common and the modified KDD process to be able to compare the results.

Table 2 shows the CLV result of the common KDD process (column CLV) and the result of the modified KDD process (column CLV corrected). Including the majority of the test data, a correction of the CLV forecast took place. Since at several times customer data was extracted from the Data Warehouse, a comparison with actual results was possible based on the documented customer history. This comparison confirmed the more precise forecasts of the modified KDD process. Thus the descriptive modification of the KDD process makes a contribution for the increase of the efficiency of a value oriented banking business.

5 Conclusions

An implemented forecast model offers a systematic support, which is essential for successful enterprises. With the estimation of the necessary input data, Predictive Analytics offers the support and thus, decision support improvement potentials. This research project aims on a confrontation of the results of different forecast models and the comparison of their results.

This chapter contributes to the discussion of Business Intelligence supported forecast processes. The accuracy increase of the CLV, which is shown by comparing results of two approaches of prognosticating a CLV in the specified example, confirms the improvement potential by the modification of the KDD process. The processing concept is enhanced with two further steps: The first step contains the analysis of the source systems and their data. If this step shows the fact that a stable system is present, it will not be necessary to implement further adjustment.

Otherwise, corrective measures are necessary after the computation of the samples. The integration of additional data like demographic prognoses can be realized by the use of semantic approaches like KDDML.

Supplementing research is necessary regarding further application fields of the modified KDD process in the sense of a Predictive Analytics. Valid basic conditions and adjustments have to be tested in order to prove the usefulness and thus decision support potentials for applications. Besides, also basic conditions for the economically meaningful employment of the method have to be analyzed as the expenditure for the production of the prognosis rises.

References

1. Anandarajan M, Anandarajan A, Srinivasan CA (2004) Business intelligence techniques. A perspective from accounting and finance. Springer, Berlin
2. Armstrong JS (2001) Introduction. In: Armstrong JS (ed) Principles of forecasting: a handbook for researchers and practitioners. Springer, Berlin, pp 1–12
3. Bauer K (2005) KPI identification with fishbone enlightenment. DM Rev 15(3):12
4. Chakrabarti S, Witten IH, Cox E (2009) Data mining: know it all. Morgan Kaufmann, Burlington
5. Chamoni P, Gluchowski P (2006) Analytische Informationssysteme—Einordnung und Überblick. In: Chamoni P, Gluchowski P (eds) Analytische Informationssysteme: Business Intelligence-Technologien und -Anwendungen, 3rd edn. Springer, Berlin, pp 3–22
6. Davenport TH, Harris JG (2007) Competing on analytics—the new science of winning. Harvard Business School Press, Boston
7. Ebert F, Isnik Z (2010) Qualitative und quantitative Szenarioplanung als Grundlage der unternehmerischen Nachhaltigkeit. http://www.haufe.de/controllerwissen/controllermagazin/magazineItemDetail?editionID=1260365185.13&articleID=6. Accessed 31 July 2011
8. Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) Knowledge discovery and data mining: towards a unifying framework. In: KDD proceedings, AAAI
9. Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) The KDD process for extracting useful knowledge from volumes of data. Comm ACM 39(11):27–34
10. Hall DJ (2008) Decision makers and their need for support. In: Burstein F, Holsapple CW (eds) International handbooks on information systems. Springer, Berlin, pp 83–102
11. Hevner AR, March ST, Park J (2004) Design science in information systems research. MIS Q 28(1):75–105
12. Horsch A, Schulte M (2010) Wertorientierte Banksteuerung II: Risikomanagement, 4th edn. Frankfurt School, Frankfurt/M
13. Kemper H-G, Mehanna W, Unger C (2006) Business Intelligence—Grundlagen und praktische Anwendungen: Eine Einführung in die IT-basierte Managementunterstützung, 2nd edn. dpunkt, Wiesbaden
14. Luhn HP (1958) A business intelligence system. IBM J Oct(1958):314–319
15. Nissen N (2009) Fehlinvestition Konjunkturprognose? http://www.heise.de/tp/r4/artikel/30/30933/1.html. Accessed 7 July 2011
16. Plickert P, Bernau P (2010) Nur ein bisschen besser als Würfeln. http://www.faz.net/s/RubB8DFB31915A443D98590B0D538FC0BEC/Doc~E7ADDAB15913F436BAF82FB662016E18D~ATpl~Ecommon~Scontent.html. Accessed 7 July 2011
17. Sabherwal R, Becerra-Fernandez I (2010) Business intelligence. Wiley, Hoboken
18. Simon HA (1955) A behavioral model of rational choice. Q J Econ 69(1):99–118

19. Sterman JD (2000) Business dynamics: systems thinking and modeling for a complex world. McGraw Hill, Boston
20. Vogler-Ludwig K (2009) Prognosen ohne Zukunft. http://www.economix.org/PrognosenohneZukunft.pdf. Accessed 7 July 2011
21. Witten IH, Frank E (2005) Data mining. Elsevier, San Francisco