ГАЛИНА ЧОРНОУС
ЯНА ФАРЕНЮК
ІРИНА ДІДЕНКО

ДАТА МАЙНІНГ ДЛЯ ЕКОНОМІСТІВ

DATA MINING FOR ECONOMISTS

НАВЧАЛЬНИЙ ПОСІБНИК

COURSE BOOK

Київ
Видавництво Ліра-К
2023

УДК 339.1:004.89.032.2](075.8)

Approved by the Department of Economic Cybernetics
Minutes No 10, February 14, 2023

Approved by the Academic Council of the Faculty of Economics
Minutes No 9, February 21, 2023

Reviewers: *Natalia Kovtun*, Doctor of Sciences (Economics), Head of the Department of Statistics, Information and Analytical Systems and Demography, Taras Shevchenko National University of Kyiv
Tetiana Polozova, Doctor of Sciences (Economics), Head of the Department of Economic Cybernetics and Management of Economic Security, Kharkiv National University of Radio Electronics

Чорноус Галина, Фаренюк Яна, Діденко Ірина
Ч75 Дата майнінг для економістів : навч. посіб. Київ : Видавництво Ліра-К, 2023. 290 с.
ISBN 978-617-520-555-6

Galyna Chornous, Yana Fareniuk and Iryna Didenko
Data Mining for Economists : Course book. Kyiv : Publishing Lira-K, 2023. 290 p.

Навчальний посібник значною мірою сприятиме засвоєнню основ інтелектуального аналізу даних студентами економічного профілю завдяки його фокусу на моделях, методах та інформаційних технологіях, які використовуються для виявлення знань у базах даних. Різноманітні навчальні вправи та реальні кейси допоможуть студентам здобути глибоке розуміння теорії та набути необхідних практичних навичок.

Для студентів бакалаврату та тих, хто цікавиться інтелектуальним аналізом даних.

The Data Mining for Economists will greatly contribute to mastering the Data Mining Course by the students majoring in Economics due to its focus on the models, methods, and Information Technologies used for Knowledge Discovery in Databases. Various learning activities and real cases will help students to build deep understanding of the theory and to acquire the necessary practical skills.

For undergraduate students and those interested in Data Mining.

УДК 339.1:004.89.032.2](075.8)

ISBN 978-617-520-555-6

CONTENTS

INTRODUCTION

Under globalization and taking into consideration the interdependence of financial and economic processes, as well as strengthening of the role of regulatory mechanisms, the decision-making in business cannot be imagined without the use of modern information technologies, economic and mathematical methods and models.

Now we are witnessing the rapid development of Data Science, the emergence of which is primarily associated with the need to process big data accumulated in modern information systems. Most companies store a huge amount of data in order to use them for effective decision-making in future. How can one use these data to understand what is more beneficial for the company customers, how to allocate resources in the most effective way, or how to minimize losses? In order to answer these questions, the researchers have been developing Data Mining technologies over the past decades These technologies are a part of the Data Science toolkit and are used to search for patterns hidden in the data sets that cannot be determined by classical statistical methods.

Data Mining technologies are ones designed to search for non-obvious, objective and practically useful patterns in the large amounts of data. The scope of Data Mining application is not limited to any certain domain; Data Mining methods are now interesting for business entities that deploy projects based on the latest information technologies. The experience of many companies proves that Data Mining is an important component of successful analytical initiatives and the return on its use can reach 1000%.

Data Mining is a key part of General Data Analytics and one of the core disciplines in the data industry that uses advanced analytical methods to find useful information in data sets. As a scientific discipline, Data Mining has its own philosophy, principles, concepts and multi-faceted tools that are dynamically developing and enriching.

This discipline's object is large sets of structured data accumulated and stored by the companies during their business activities.

The subject of Data Mining is the methodology, concepts and corresponding tools of mathematical modeling, statistics, artificial intelligence, which, using the system paradigm and Computer Information Systems, help analysts to extract knowledge from data, and in this way to build artificial intelligence amplifiers.

The suggested course book contains general information about the essence, building and practical application of Data Mining technologies. It outlines tools for in-depth analysis of problems and phenomena in the economy based on the application of economic and mathematical methods and models to solve problems using appropriate software to support doing quantitative and qualitative economic analysis.

The main objective of the course book is to provide students of economic specialisms with a holistic understanding of the Data Mining process, its stages, its tools, etc., as well as to enable students to master practical skills through the use of Business Intelligence system Microsoft Power BI, analytical platforms: IBM SPSS Modeler, Deductor Academic, Loginom Studio, and open source software Weka.

Each of the 13 chapters of the suggested course book is devoted to the main scientific and applied research results in a particular area of Data Mining. The main stages of development and application of appropriate technologies are accompanied by the use of appropriate software, which is reviewed at the end of each chapter of the course book. Each chapter contains tests, case studies, and a list of recommended literature. The course book also contains a glossary.

To facilitate the presentation and perception of the information in the proposed course book, key terms, definitions and names of the algorithm methods, software tools, etc. are highlighted in bold, whereas, menu items and control button names are highlighted in italics.

This course book will be useful for students majoring in all economic specialisms, postgraduate students, as well as specialists whose interests are related to advanced data analytics.

CHAPTER 1. AN INTRODUCTION TO DATA MINING. BASIC METHODS, MODELS, PROCESSES OF INTELLIGENT CALCULATIONS

1.1. Requirements for modern analytical technologies. Big Data: general characteristics and main trends

Under **analytical technology** we will understand methods, based on certain mathematical models, algorithms, and tools that allow us to estimate the values of unknown characteristics and parameters based on known data.

Analytical technologies in the economy are primarily used by the people who make managerial decisions - managers, analysts, experts, consultants. It should also be noted that there are no clear algorithms for solving real business and production problems. Therefore, managers and experts find solutions to such problems mainly using their personal experience. Analytical technologies allow to use mathematical modeling, which significantly increases the validity of decisions.

Among the classical approaches to data analysis in practice, deterministic and probabilistic techniques turned out to be the most common. Often, classical methods are ineffective for solving many different problems, because it is impossible to accurately describe reality using a small number of model parameters, or calculations based on models are too resource-intensive.

Among the main factors that confirm the need to use new analytical techniques, transferring certain intellectual functions to artificial systems, we can mention increasing of the volume and variety of information and the limited human capabilities for its development and processing.

Now the term **Big Data** is actively used, it describes complex and large sets of different data (both structured and unstructured) that cannot be analyzed using traditional approaches.

Since Big Data definitions are still evolving, it may be simplest to define the concept of Big Data in terms of a set of features. There are the "seven Vs" of Big Data - ***volume, velocity, variety, variability, veracity, volatility, and value*** [5].

According to analysts ' forecasts, the Big Data Market will triple between 2020 and 2028, exceeding the 175 zettabyte mark. Big Data Analytics is called the heart of the digital world, based on analyzing and transforming data into knowledge that provides valuable business insights.

Thus, when supporting the decision-making process, modern analytical technologies should provide:

- processing huge amounts of structured and unstructured data;
- formalization of rational and irrational approaches;

- a combination of two types of intelligence – human and artificial;
- use of formal and informal cognitive analysis techniques;
- proactive production, accumulation, and subsequent use of systematized cognitive information – knowledge-for management.

1.2. Basic model of contemporary analytical work

Analytics is understood as the entire set of principles of methodological, organizational and technological support of individual and collective mental activity, which allows to effectively process information in order to identify the essential and meaningful core in it; improve the quality of existing knowledge and acquire new one; make effective and rational managerial decisions. The basic model of the essence of analytical activity is shown in Figure 1.1.

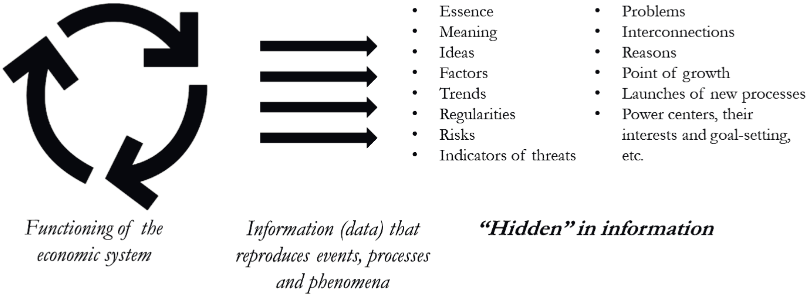
• Essence
• Meaning
• Ideas
• Factors
• Trends
• Regularities
• Risks
• Indicators of threats
• Problems
• Interconnections
• Reasons
• Point of growth
• Launches of new processes
• Power centers, their interests and goal-setting, etc.
Functioning of the economic system
Information (data) that reproduces events, processes and phenomena
"Hidden" in information

Figure 1.1. Basic model of the essence of analytical activity

The first block is the functioning of the economic system, represented through all the diversity of its mental, cultural, institutional and cognitive processes, with all their interconnections and interrelations, full of real problem situations.

The second block is information that reproduces various aspects of the economy functioning with the help of various tools and together creates an information image (model) of the economic system.

Data sources: data from corporate information systems, data from transactional systems, data from web-services, data from information agents, data from social media, data from sensors, etc. [5].

The third block is the latent content of information flows, knowledge got through analytical procedures.

To manage information flows, it is important to understand the levels of the information space, which, according to the degree of increasing complexity, can be arranged in the following sequence: ***data, information, knowledge, understanding, creative thinking***. At the same time, ***Information*** can be understood as processed, meaningful data; but ***knowledge*** in its way - as the information of semantic-level, on the basis of which certain semantic conclusions can be drawn. ***Understanding*** is formed on the basis of available data, information, and knowledge. It provides the possibility of creating new knowledge based on previously acquired one. The highest level of information space is ***creative thinking*** - that allows to use information from other levels in order to create understanding in areas where it is currently missing.

Information and knowledge are the most important factors for creating a competitive advantage in the information economy. Information can be both an economic raw material and a specific economic good. Sources of information as raw materials can be divided into internal, or market ones (data from suppliers, customers, competitors), institutional ones (information from experts, research institutions) and other (scientific, technical publications, etc.).

Knowledge is a product of the highest level; its production is based on knowledge-intensive technologies for information processing. Knowledge differs from traditional goods by its capability of being repeatedly copied and practically used. As an intelligent product, knowledge can have a digital form, be stored on material media and fill the knowledge base, which contains, along with a set of specialized facts, also ways of knowledge presentation that determine how facts are selected for a certain logical sequence. Their creation and accumulation happens due to appropriate IT.

Data analytics is the process that leads from data to management decisions through knowledge.

Large amounts of data do not mean good quality of them by default. Therefore, much attention is paid to data quality issues in modern analytical methods.

In this regard, the modern concept of data analysis should include:

1) data can be inaccurate, incomplete (contain omissions), contradictory, heterogeneous, indirect, and at the same time they may be of a large, almost unlimited volume;

2) data can be quantitative, qualitative, text, multimedia, and so on;

3) all stages of the analysis should be carried out in the shortest possible time to ensure prompt decision making;

4) data analysis algorithms by themselves should have elements of intelligence, namely, be able to draw general conclusions based on partial observations;

5) the processes of processing “raw” data into information, and information in its turn - into knowledge, can no longer be performed manually, and need to be automated;