Modern Data Mining with Python

A risk-managed approach to developing and deploying explainable and efficient algorithms using ModelOps

Dushyant Singh Sengar

Vikash Chandra

www.bpbonline.com

First Edition 2024

ISBN: 978-93-55519-146

LIMITS OF LIABILITY AND DISCLAIMER OF WARRANTY

The information contained in this book is true to correct and the best of author's and publisher's knowledge. The author has made every effort to ensure the accuracy of these publications, but publisher cannot be held responsible for any loss or damage arising from any information in this book.

All trademarks referred to in the book are acknowledged as properties of their respective owners but BPB Publications cannot guarantee the accuracy of this information.

www.bpbonline.com

Dedicated to

My sources of joy: daughters ***June*** *and* ***Marchie***

and

My pillars of strengths: wife ***Amrita*** *and*
parents ***Usha*** *and* ***Narendra Singh Sengar***

- Dushyant Singh Sengar

Foreword

Over the last 20 years, my passion for Data Mining, Machine Learning, and Deep Learning has been a cornerstone of my career. My evolving interest in these fields has enabled me to drive forward the Banking, Payments, and Retail industries by emphasizing the crucial role of AI/ML in business advancement. As a Thought Leader and Coach for over 15 years, I have trained leaders in strategic AI/ML initiatives and fostered a culture of continuous learning.

In this data-centric era, the book, *Modern Data Mining with Python: A risk-managed approach to developing and deploying explainable and efficient algorithms using ModelOps* emerges as a vital resource. It is a comprehensive guide aimed at demystifying data mining complexities, especially focusing on explainability and traceability. Spanning 13 well-crafted chapters, the book is a beacon for understanding and implementing data mining techniques in the banking sector and beyond.

Educational Journey and Ethical Implications

The book starts from the basics of statistics and exploratory data analysis and then ventures into advanced deep learning techniques. It emphasizes ethical machine learning model development, tackling biases, ensuring algorithmic transparency, and adhering to responsible AI principles. This approach is not only about learning techniques but also about becoming a responsible decision-maker in the data-driven business world.

Methodology and Case Studies

Adopting a first-principles teaching method, the book starts with fundamental business questions and then applies data mining concepts, particularly in the regulated banking industry. It includes case studies reflecting best practices in model building and maintenance, applicable across various industries. This approach simplifies complex topics, making them accessible without requiring deep knowledge of algebra, probability, or calculus.

Welcome to a journey of mastering data mining with a conscience.

– Dr. Rajendra Prasad Gangavarapu

About the Authors

- **Dushyant Singh Sengar** is a passionate leader in AI and Risk management with experience building high-performing teams and leading organizations to become data-driven. His extensive 18 years of experience on both sides of the Atlantic spans various roles, including model development, risk assessment, and driving AI product development initiatives. He is a seasoned professional in the banking and consulting industry with experience modernizing retail, credit risk, and marketing platforms leveraging AI/ML techniques on modern-day cloud and MLOps infrastructure.

  Dushyant Sengar also brings a wealth of experience in the model risk management (MRM) arena that enables him to provide a holistic and efficient road map for the AI-based product development lifecycle.

  He holds an M.S. in data science from Northwestern University, Chicago, and a B.E. in Information technology from M.I.T.S. Gwalior, India. He has also been a freelance data science and analytics trainer during his India-based professional experience and is now a tenured speaker in the AI, Innovation, and Risk management-driven conference scene in the United States. Dushyant thrives on bringing cross-functional knowledge and teams together to ensure alignment between technology, risk, and business goals. He has a proven track record of mentoring students from various backgrounds and nurturing their data science potential.

  When he is not working, you can often find him reading, writing, or exploring new places and cultures. He is passionate about using technology for social good, driven by a mission to spread technology and AI knowledge among enthusiasts for positive change.

- **Vikash Chandra** is a data scientist and software developer having industry experience in executing and implementing projects in the area of predictive analytics and machine learning across multiple business domains. He has experience in handling and modifying large quantities of both structured and unstructured data leveraging SAS, R, Python, and other big data technologies.

  He is an alumni of prestigious institutions like Jawaharlal Nehru University, and Shri Ram College of commerce, University of Delhi, India.

About the Reviewers

- **Dishant Banga** is currently working as a Senior Data Analyst at Bridgetree, USA. He received his master's degree in systems Engineering and Engineering Management with a specialization in Data Analytics/Data Science from the University of North Carolina, Charlotte, in 2018. His interests include developing statistical and machine learning models, artificial intelligence, Machine Learning, and Data Science applications to solve complex business problems. He has participated in various national and international competitions and is recognized for his work.

- **Vivek Chaudhary**, Founder of Dyota AI, pioneers advancements in artificial intelligence, propelling the company to the forefront of AI technology. With pivotal roles at startups like Vitra.ai, Neurodynamic.ai, and CodeVector. Vivek, an accomplished mentor, has guided 5000+ AI professionals through collaborations with Upgrad, Board Infinity, EC Council, Packt, AlmaBetter, and freelance projects. As a Thought Leader in Computer Vision at Global AI, he authored Amazon's bestselling "Data Investigation-EDA the Right Way." Vivek's 5+ years of expertise in EDA, Machine Learning, and Computer Vision shape the AI education landscape as a mentor for CODERED E-Learning.

- **Swagata Ashwani** is a seasoned data scientist with a rich background in analytics and big data. Currently serving as the Principal Data Scientist at Boomi, Swagata plays a crucial role in harnessing the power of data to drive innovation and efficiency. In her role, she plays a crucial role in leading generative AI initiatives for the company. She is also Chapter Lead at SF Women in Data, where she fosters building a rich community for women to celebrate women in varied data roles.

Acknowledgement

Writing a book is a bit like trying to build an aircraft in the hope that it would take readers on a fulfilling journey. I started this endeavor with my professional experiences as the raw materials and a dream as the duct tape. However, undoubtedly, the unconditional support of loved ones, who sacrificed a lot of fun family times, also served as the bedrock of this illuminating experience.

First and foremost, I would like to thank my wife and daughters for being patient with me and continuously encouraging me to write the book. My parents have always taught me to chase the North Star and this life-long learning helped me sail through the twists and turns of book writing odyssey as well.

I am grateful to my co-author, Vikash, for helping me with his Python expertise, which adds immense value to the application focus of this book. I gratefully acknowledge Mr. Dishant Banga, Ms. Swagata Ashwini, and Mr. Vivek Chadhury for their technical scrutiny of this book.

I would also like to extend my heartfelt thanks to Dr. Rajendra Prasad Gangavarapu for his time, wisdom, and the genuine passion he brought to the foreword. His willingness to lend a voice to this project is an honor that I cherish deeply. My sincere gratitude also goes to the team at BPB Publication for being supportive and understanding enough to smoothen and expedite the book publishing process.

Last but not least, to the reader who is holding this book—a big, heartfelt thank you for joining me on this AI journey of infinite possibilities. Your curiosity and willingness make all the late nights, early mornings, and weekends entirely worth it.

Preface

In an era where data is the new gold, the ability to mine this vast resource for insights and decisions is not just an advantage but a necessity. *Modern Data Mining with Python: A risk-managed approach to developing and deploying explainable and efficient algorithms using ModelOps* is a comprehensive guide tailored to unravel the complexities of data mining, particularly with an emphasis on black-box model explainability and traceability.

Through its 13 meticulously crafted chapters, this book describes responsible AI approaches to improving various AI/ML techniques adoption and business processes by demystifying best practices in model risk management and operations. Author Dushyant Sengar created this guide for budding data scientists and experienced business leaders looking to improve real-world AI/ML system outcomes in the banking sector and beyond.

This book establishes a solid foundation of data mining, starting with exploratory data analysis and inferential statistics and progressing through advanced techniques like XGBoost, and deep learning. Each chapter, with a focus on a specific data mining technique and complete with its applications and nuances, serves as a building block of a comprehensive data mining knowledge base. You will explore industry best practices for making fair and efficient decisions while learning technical approaches for responsible AI across model validation, explainability, bias management, and AI-based product development using MLOps.

In a world where business decisions have far-reaching consequences on consumers' lives, the ethical deployment of data mining techniques is paramount. As of December 2023, there have been numerous public reports of model mismanagement and algorithmic discrimination, leading to unexplainable outcomes in highly regulated industries. This book addresses this need head-on, emphasizing the importance of developing machine learning models that are efficient, ethical, and explainable. It guides you through the intricacies of mitigating biases, ensuring algorithmic transparency, and upholding the principles of responsible AI. By doing so, it aims to foster an environment where the benefits of data-driven decision-making are equitably distributed.

The unique approach of this book lies in its method of teaching. It adopts a first-principles approach, starting with fundamental business questions. Subsequently, these questions are broken down into smaller manageable chunks and framed as data, model, and optimization problems. The case studies provided offer a window into the best practices and critical considerations in model building and maintenance specifically in the highly regulated banking industry, and further extend their relevance across other industries. The beauty of this approach is that it simplifies complex topics without requiring a deep background in algebra, probability, or calculus.

As you turn the pages of this book, you are not just learning data mining techniques; you are embarking on a journey to become a responsible and effective decision-maker in the data-driven world. This book is not just a learning tool; it is a commitment to ethical and fair data practices that will shape the future of business and technology. Welcome to the journey of mastering data mining with a conscience. The chapter-wise details are listed below.

Chapter 1: Understanding Data Mining in a Nutshell - introduces data mining, explaining the data mining growth journey, biases, and risks while discussing how humans have been pushing the boundaries over the decades to train machines using approaches similar to various human learning models. This chapter provides a foundation for understanding the development of modern-day data mining models that touch various aspects of human life, from agriculture to space exploration. It discusses various risks and challenges associated with machines trying to imitate human learning and making decisions before wrapping up with potential solutions to these modern-day data mining challenges.

Chapter 2: Basic Statistics and Exploratory Data Analysis - discusses the art of exploring datasets to gain an understanding of their features' strengths and discrepancies to prove real-world hypotheses. Before diving into inferential statistics, the chapter teaches the use of non-graphical and graphical data exploration methods using various descriptive techniques. The chapter leverages Python for data exploration and inferential statistics and will continue to do so for all case study implementations throughout the book.

Chapter 3: Digging into Linear Regression - covers the depth of the regression technique with an explanation of fitting a line to the data to the advanced twists and turns of regularization, including the need and complexities of Lasso and Ridge techniques. This statistical deep dive will be supplemented by implementation details on MLflow, a model reproducibility and experimentation tracking tool to manage model-development related risks.

Chapter 4: Exploring Logistic Regression - covers logistic regression with the intricacies of its data requirements, the art of estimating probabilities, and the movements of loss functions. This will include a discussion on the challenges and assumptions while decoding the results and mastering the interpretation of logistic regression models. This chapter will also lay the foundation for model explainability and interpretability, which is a cornerstone of modern-day data mining.

Chapter 5: Decision Trees with Bagging and Boosting - uncovers decision trees – a versatile, non-parametric method that is foundational for powerful algorithms like Random Forest and Gradient Boosting. This chapter explains the ensemble modeling concepts of bagging and boosting by navigating through four case studies using MLFlow and SHAP libraries. These Python tools emphasize the importance of model reproducibility and interpretability in machine learning model development.

Chapter 6: Support Vector Machines and K-Nearest Neighbors - expands the discussion from chapters 4 and 5 on supervised methods by introducing unique concepts of support and look-alike neighbors. The chapter builds the conceptual foundations theoretically, following it up with a detailed case study that highlights the effectiveness of SVM and KNN in solving a specific business problem in the financial services industry.

Chapter 7: Putting Dimensionality Reduction into Action - explores a crucial data preparation step in a model development exercise. The dimensionality reduction techniques allow practitioners to distill complexity into simplicity and uncover the essence of the data. This chapter will discuss best practices for transforming high-dimensional data into lower-dimensional representations to gain a deeper understanding, facilitate visualization, and enhance subsequent modeling and analysis tasks using a case study.

Chapter 8: Beginning with Unsupervised Models - explains what unsupervised models are and how they differ from supervised ones. It covers three widely used methodologies, including DBSCAN, that fit different use cases well. This chapter aims to provide a holistic approach to developing real-world unsupervised solutions using a case study approach in the financial services industry.

Chapter 9: Structured Data Classification using Artificial Neural Networks - provides a profound explanation of the Artificial Neural Network (ANN). It will also discuss the practical examples based on a real-life scenario and explain how to apply this modeling technique to solve a medium-complexity problem.

Chapter 10: Language Modeling with Recurrent Neural Networks - describes the solid principles of Recurrent Neural Networks (RNN) using real-life examples. It will provide a comparison between popular practices behind the principles of text data handling, preparation and RNN modeling techniques suitable for language processing and understanding.

Chapter 11: Image Processing with Convolutional Neural Networks - focuses on bridging the gap between raw visual information and actionable insights. Convolutional Neural Networks (CNNs), are revolutionizing industries by tackling image processing challenges such as extracting hierarchical features for recognizing patterns, shapes, and objects within images, regardless of their scale, orientation, or position. This chapter will discuss CNNs capabilities with their high-dimensional data requirements and architectural design to ultimately present a case study to extract meaningful business information from PDF documents.

Chapter 12: Understanding Model Risk Management for Data Mining Models - explains the importance of understanding, identifying, and handling the risks associated with the use of data mining models during development and in a production environment, making inferences on previously unseen data. This chapter will shed light on various guidelines, regulations, and technology-driven approaches to risk mitigation. We will conclude with an industry-focused case study to make clear the role of a model developer in risk management efforts within an organization.

Chapter 13: Adopting ModelOps to Manage Model Risk - brings together the knowledge from the previous 12 chapters on model development, operational efficiency, and risk management. It builds a foundation for developing a responsible AI solution by first describing the intricacies of the financial services regulatory landscape, specifically in the fair credit lending space. It then goes on to explore the significance of MRM as it relates to fair lending. Eventually, the readers will gain from a compelling case study: the development of a Fair Lending risk assessment web application leveraging modern ModelOps tools.

Code Bundle and Coloured Images

Please follow the link to download the
Code Bundle and the ***Coloured Images*** of the book:

https://rebrand.ly/9b9424

The code bundle for the book is also hosted on GitHub at
https://github.com/bpbpublications/Modern-Data-Mining-with-Python.
In case there's an update to the code, it will be updated on the existing GitHub repository.

We have code bundles from our rich catalogue of books and videos available at **https://github.com/bpbpublications**. Check them out!

Errata

We take immense pride in our work at BPB Publications and follow best practices to ensure the accuracy of our content to provide with an indulging reading experience to our subscribers. Our readers are our mirrors, and we use their inputs to reflect and improve upon human errors, if any, that may have occurred during the publishing processes involved. To let us maintain the quality and help us reach out to any readers who might be having difficulties due to any unforeseen errors, please write to us at :

errata@bpbonline.com

Your support, suggestions and feedbacks are highly appreciated by the BPB Publications' Family.

Piracy

If you come across any illegal copies of our works in any form on the internet, we would be grateful if you would provide us with the location address or website name. Please contact us at **business@bpbonline.com** with a link to the material.

If you are interested in becoming an author

If there is a topic that you have expertise in, and you are interested in either writing or contributing to a book, please visit **www.bpbonline.com**. We have worked with thousands of developers and tech professionals, just like you, to help them share their insights with the global tech community. You can make a general application, apply for a specific hot topic that we are recruiting an author for, or submit your own idea.

Reviews

Please leave a review. Once you have read and used this book, why not leave a review on the site that you purchased it from? Potential readers can then see and use your unbiased opinion to make purchase decisions. We at BPB can understand what you think about our products, and our authors can see your feedback on their book. Thank you!

For more information about BPB, please visit **www.bpbonline.com**.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

Table of Contents

Chapter 1
Understanding Data Mining in a Nutshell

Introduction

Human ingenuity, innovation, and the quest for knowledge have kept us moving from the ancient trading days of the Silk Route in 130 B.C. to the modern days of global, electronic, algorithmic trading and banking platforms. Notably, the growth trajectory of statistics, followed by data mining, and related algorithms has had a major role in the advancements of various financial services disciplines over the centuries. This potentially stems from the fact that financial services and data mining are both driven by the same human principles of observation, critical thinking, problem solving and community sharing. Today, these two disciplines have evolved to arrive at the promising juncture of the **financial technology (FinTech)** revolution, but not without multiplying the risks of the two components together.

Specifically, the recent rapid advancements in data mining technologies have come with challenges and risks that warrant the development of ever-new regulations for industries such as financial services. This ensures the financial safety of globally connected trading markets and other customers of FinTech services in various local markets.

In this regard, this chapter will explore the various aspects of the data mining process, and trace its similarities with the human learning process, to track its growth journey, biases, and risks while discussing how humans have been pushing the boundaries to train machines using approaches similar to their learning model. This has led to the development of modern-day data mining models touching various aspects of human life ranging from agriculture to financial services to space exploration. We will discuss various risks and challenges associated with machines trying to imitate human learning and making decisions before wrapping them up with potential solutions to these modern-day data mining challenges.

Structure

This chapter will cover the following topics:

- What defines modern data mining
- The lifecycle: Data to insights consumption
- Understanding pattern recognition
  - Significance of the human learning process
  - The human learning process and mental models
  - Data: The key ingredient for meaningful patterns and relationships
- How machines leverage data to build models
  - Machine learning process
    - Two dominant strategies: classification and regression
    - Biases and learning shortfalls
    - Measuring learning accuracy and balancing trade-offs
    - Can data size and sample impact learning
  - How do humans benefit from data and learning
  - Modern-data data mining challenges, risks, and remediations frameworks

Objectives

By the end of this chapter, you will be able to describe how various types of learning needs drive the learning approaches supported by the available data. You will be able to identify the right learning technique in a given scenario and outline the numerous pitfalls associated with the modern-day data mining landscape.

What defines modern data mining

We are drowning in information but starved for knowledge

– John Naisbitt, 1982 Megatrends

Data mining originated in the 1970s, flourished in the 1980s, and continues strong today. This is thanks to many new data formats and storage systems that the world is getting exposed to. In simple terms, data mining can be defined as the extraction or mining of knowledge from large and varied amounts of data. This typically involves characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis, or evolution analysis. Myriad data mining tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research. However, the growing gap between newer data and essential knowledge recently calls for innovative data mining tools and techniques to turn the new-age data wastelands into orchards of knowledge.

Historically, many manual processes that relied on sophisticated database querying, statistical techniques, and human intervention were leveraged to detect patterns and relationships in cleaned and stored databases. One such industry was retail which transitioned from manual processes to sophisticated **customer relationship management** (**CRM**) systems that involved the use of advanced database querying, statistical techniques, and human expertise to extract valuable patterns and relationships from cleaned and stored databases. The main types of such relationships are classes (partitioned into predefined groups), clusters (partitioned into logically related groups), associations, behavior patterns, and trends. This transformation significantly enhanced the ability of retail businesses to understand and respond to customer behavior in a more data-driven and effective manner.

However, this had to change with the advent of big data which required the process of knowledge extraction and learning to be automated and more intelligent on its own. In recent years, we have also seen the rise of machine learning and **artificial intelligence** (**AI**) techniques that have redefined the data mining landscape by enabling computers to learn a bit as humans do. These ML algorithms can imitate human learning by learning from experiences and grow better in decision-making when exposed to more diverse datasets.

Modern data mining systems hence involve computers that are supplied with high-quality data, to discover novel data patterns driven by sophisticated machine learning algorithms for knowledge discovery on a fair and continuous basis.

The lifecycle: Data to insights consumption

Over the decades, different process frameworks have been attached to the field of data mining to derive the best return on investment of time and effort. These framework proposals have been driven primarily by the data quality, data size, business needs, and the level of algorithmic sophistication prevalent in the era. In today's world, the most consistent framework that is used across a range of organizations across the globe can be described in the following six steps:

1. **Defining the problem**: In a bank, this could mean improving the credit risk assessment process to minimize default rates and enhance lending decisions. This is the most critical step that needs to be done right.

2. **Data identification and pre-processing**: This would mean collecting historical data on loan applicants, including their financial information, credit history, employment status, and other relevant variables, and cleaning up all this information to resolve inconsistencies and anomalies.
3. **Defining the data mining requirement**: Such as characterization, discrimination, association or correlation analysis, classification, prediction, clustering, anomaly or outlier analysis to group loan applicants into *low risk* and *high risk*.
4. **Domain knowledge and measurement indicators for pattern evaluation**: This helps in developing and maintaining a healthy credit risk program over a period based on business and industry guidelines.
5. **Data visualization for easy consumption**: These visualizations help non-technical stakeholders understand the model's performance and interpret results for better adoption of algorithm results.
6. **Deployment and decision monitoring**: Lastly, the integration of the predictive model into the bank's loan approval system and automation of the credit assessment process for continuous monitoring of any potential deviations. This allows for regular updates of the model using new data to adapt to changing patterns.

These six steps correlate heavily with the five steps of the **Knowledge Discovery in Databases** (**KDD**) that was coined by *Gregory Piatetsky-Shapiro* in 1989. Another industry standard framework known as the **Cross-Industry Standard Process for Data Mining** (**CRISP-DM**) is perfectly consistent with the current six steps of the data mining process. It is an iterative process; many tasks backtrack to previous tasks and repeat certain actions to bring more clarity. These six major phases of the CRISP-DM process are idealized sequences of activities, but they relate well with the six-step industry-standard framework discussed above. *Figure 1.1* illustrates the CRISP-DM process steps (colored boxes) with the industry-standard steps (in text) mapped alongside it:

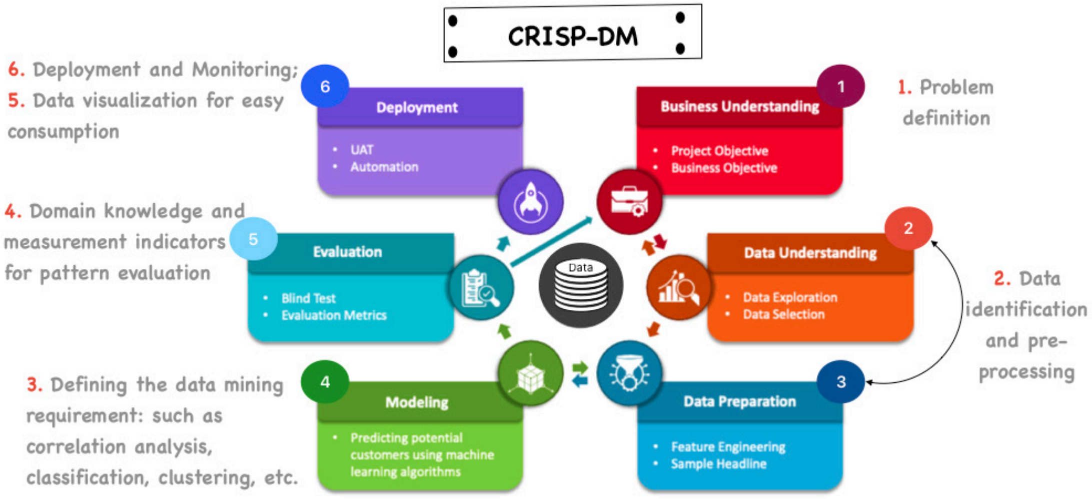
CRISP-DM
6. Deployment and Monitoring;
5. Data visualization for easy consumption
6
Deployment
UAT
Automation
1
Business Understanding
Project Objective
Business Objective
1. Problem definition
4. Domain knowledge and measurement indicators for pattern evaluation
5
Evaluation
Blind Test
Evaluation Metrics
Data
2
Data Understanding
Data Exploration
Data Selection
2. Data identification and pre-processing
3. Defining the data mining requirement: such as correlation analysis, classification, clustering, etc.
4
Modeling
Predicting potential customers using machine learning algorithms
3
Data Preparation
Feature Engineering
Sample Headline

Figure 1.1: *CRISP-DM mapped to six standard industry practices*

In summary, the outcome of a data mining process is either an exploration exercise to understand the data patterns or build predictive models also known as mathematical equations based on classification or regression techniques. However, in today's world, it is not sufficient to explore the data and build a model, but it is equally important to derive a long-term return on investment by evaluating and deploying the model. Steps 5 and 6 need careful consideration to ward off any potential risks from continuous usage of the developed models. Hence model monitoring becomes a significant critical step in the process. Before moving ahead with data mining, let us understand more about models and how they learn from data.

Understanding pattern recognition

Let us start by understanding human mental models, which are made up of interrelated memories, conceptual knowledge, and causal beliefs that create an understanding of how some things work in the real world and form expectations about future events. For example, many individuals in the world have a mental model of a bank's operation and money-making process.

Primarily, banks take deposits from individuals and compensate depositors with certain interest rates and then lend that money to other borrowers at a higher interest rate and profit from the interest rate spread. The interest rates vary depending on many external factors and could be unique to individuals. Humans use this particular set of conceptual knowledge, expectations, and causal beliefs to form a pattern or model around the bank's profit-making. This is formed primarily by human historical personal experiences and might also be formed by the transfer of theoretical knowledge through literature and experimentation.