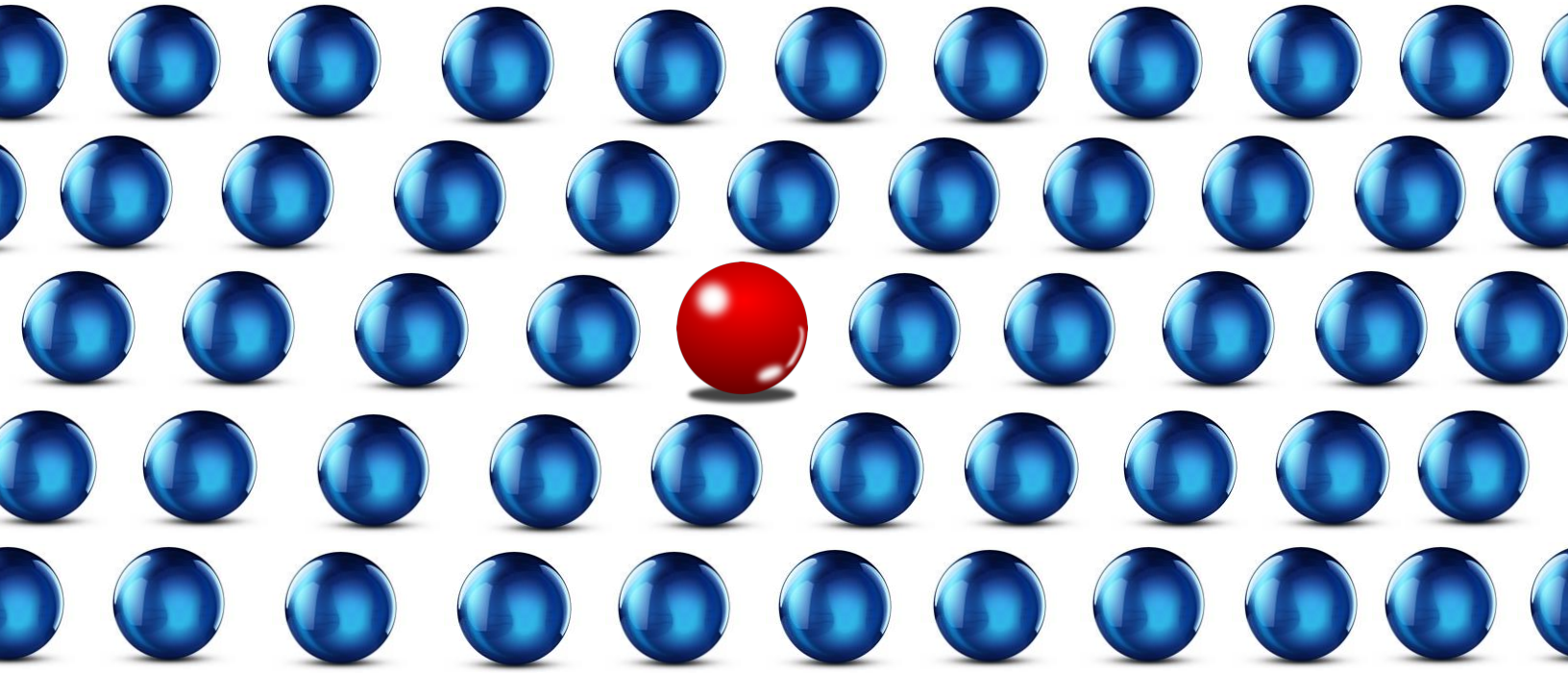
Handbook of Anomaly Detection

Cutting-edge Methods and Hands-On Code Examples

2nd Edition

With 200+ Questions & Answers

Q&A



CHRIS KUO

INNOVATION PRESS

Paperback: ISBN 979-8-9907810-1-6
Digital: ISBN 979-8-9907810-2-3

First edition: February 2023
Second edition: November 2024

This book was typeset using LaTeX software. The TeX template acknowledgement:
The *Beautybook* [https://ctan.org/pkg/beautybook?lang=en]
Comprehensive TEX Archive Network (CTAN)

The Innovation Press,
Ann Arbor, MI

INNOVATION
PRESS

About the author

Chris Kuo is a data scientist and an adjunct professor with 23+ years of experience. He led various data science solutions including customer analytics, health data science, fraud detection, and litigation analytics. He is also an inventor of a U.S. patent. He has worked at several Fortune 500 companies in the insurance and retail industries. In addition to teaching at Columbia University, he has taught courses in time series forecasting, mathematical finance, economics, and management at Boston University, University of New Hampshire, and Liberty University. Chris Kuo received his Ph.D. in Economics from SUNY at Stony Brook and his B.S. in Nuclear Engineering from National Tsing-Hua University, Taiwan. He and his wife live in New York, New York.

Chris Kuo is a writer on Medium.com, also known as 'Dr. Dataman' in the field of data science. He has published in several economic and management journals and is the author of the following data science books (by year):

2024 *The Handbook of Anomaly Detection (2nd Edition): Cutting-edge Methods and Hands-On Code Examples*
2024 *Modern Time Series Forecasting Techniques For Predictive Analytics and Anomaly Detection: From Classical Foundations to Cutting-Edge Applications*
2023 *The Handbook of NLP with Gensim: Leverage topic modeling to uncover hidden patterns, themes, and valuable insights within textual data*
2023 *The Handbook of Anomaly Detection (1st Edition): Build and modernize your anomaly detection models with examples*
2022 *Transfer Learning for Image Classification: With Python examples*
2022 *The eXplainable A.I.: With Python Examples*
2021 *Modern Time Series Anomaly Detection: With Python and R examples*

To our beloved daughter, Eileen,

You are the blossom that brightens us with every smile,
and fills our lives with beauty, joy and love.

Mom and Dad

Surely goodness and mercy shall follow me
All the days of my life;
And I will dwell in the house of the Lord Forever.
Psalm 23:6 (NKJV)

PREFACE

Welcome to the second edition of the *Handbook of Anomaly Detection.* The second edition is an updated and expanded edition in response to the demand and positive feedback from readers of the first edition.

Sec Why read the 2nd edition of the book?

Anomaly detection has become an indispensable skill in data science. Anomalies, or outliers, often signal critical insights, from detecting fraudulent activities to identifying system failures. This second edition is meticulously designed to equip you with the knowledge and skills necessary to excel in this crucial area. Here's why this edition is a must-read for both aspiring and experienced data scientists:

Comprehensive and in-depth coverage: The second edition offers a wide range of techniques for anomaly detection. It ensures a well-rounded understanding of the foundational concepts, making it suitable for beginners and seasoned professionals alike. Each chapter builds on the previous one, creating a cohesive and comprehensive learning experience.

Enhanced explanations and visual presentations: We've enriched this edition with a wealth of visualizations. These enhancements are designed to clarify complex concepts. You'll learn not just the 'how' but also the 'why' behind each method.

Supervised and unsupervised learning techniques: We expand the second edition to cover the advanced machine learning techniques for anomaly detection. On the unsupervised learning part, you will be equipped with a wide range of modern techniques to discover new patterns. On the supervised learning part, you will gain a deeper understanding into grid-search, hyperparameter tuning, regularization, under-sampling and over-sampling techniques. You will use the code examples to build a suite of high-power machine learning models.

200+ data science Q&A: One of the standout features of the second edition is the inclusion of the Q&A sections in each chapter. By working through these questions, you'll be well-prepared to showcase your expertise and secure your next job opportunity.

Sec Who this book is for

This book is designed to serve a diverse audience, catering to both beginners and experienced professionals in the field of anomaly detection. Here's who will benefit most from this book:

Data science professionals: Data science professionals seeking to deepen their understanding of anomaly detection methods will find this book invaluable. Professionals from various industries, including finance, healthcare, cybersecurity, and manufacturing, where anomaly detection is critical, will benefit from this book. It demonstrates how anomaly detection techniques can be applied to implement effective solutions. Whether you are a data analyst, machine learning engineer, or data scientist, this book will enhance your toolkit with practical and advanced techniques.

Students and academics: Advanced undergraduate and graduate-level students will find this book valuable. Instructors can also adopt it as a textbook for a semester-long course.

Researchers and developers: Researchers looking to explore new frontiers in anomaly detection will appreciate the thorough coverage of recent advancements and innovative methods. This book provides a solid foundation for academic research, offering references to additional readings and advanced topics. Developers working on implementing anomaly detection systems will also find the techniques in this book helpful.

Sec Recommended prerequisites and background knowledge

To get the most out of this book, it is helpful to have a foundational understanding of certain topics and skills.

Statistical concepts: A foundational understanding of statistical concepts is required, including Descriptive Statistics, Probability Distributions such as the Normal distribution, Linear Regression, and basic Decision Trees.

Python skills: A working knowledge of Python is essential for following the examples and exercises in this book. It includes some familiarity with Python Numpy and scikit-learn libraries.

Sec How to Use This Book

Each chapter builds progressively from fundamental concepts to advanced topics. If you're new to anomaly detection or specific techniques, begin with the earlier chapters before moving on to more complex material. This approach ensures a solid understanding of the core principles before tackling advanced models. Each chapter in this book follows a consistent format:

- Introduction: Overview of the topic with context and real-world applications.
- Concepts & Techniques: Detailed explanations of the algorithms, models, and methods.
- Reading Time Estimate: At the start of each chapter, an estimated reading time is provided to help with time management. The reading time estimate does not include your implementation of the code examples.
- Software Requirements
- Code examples & descriptions

You are advised to leverage the Q&A sections to gauge your grasp of key topics and improve your ability to explain complex concepts in simple terms.

For educators, a dedicated instructors' guide is available that provides additional tips, teaching strategies, and deeper explanations to facilitate classroom discussions.

Sec Motivation for writing this book

The success and positive feedback from the first edition of the "Handbook of Anomaly Detection" have been truly gratifying. As the field of anomaly detection continues to grow and evolve, so does the need for up-to-date resources. This motivated me to write the second edition, with the goal of providing an even more valuable and relevant guide for both new and seasoned practitioners.

Chris Kuo,
New York City, New York

CONTENTS

Part 1 INTRODUCTION

Reading time: 30 minutes

An **anomaly** is something that deviates from the usual or expected behavior or pattern. It stands out as rare, different, abnormal, or peculiar. Anomalies are also called outliers, irregularities, unexpected incidences, noise, novelties, exceptions, or rare events. They are typically the extreme values in a statistical analysis. Rare events can detrimentally impact business operations and result in significant losses. Therefore, companies all aim to apply detection techniques to find them accurately and quickly. If every operation in every industry has rare events, then every field needs **anomaly detection techniques** to raise red flags when something doesn't align with expected behaviors. This book presents the process and techniques of anomaly detection to identify irregular patterns.

Sec 1.1 Anomaly detection applications

Let's understand a few applications of anomaly detection in mechanical failure detection, insurance fraud detection, and cybersecurity.

1.1.1 Mechanical Failure Detection

In high-precision manufacturing industries, mechanical failure, though rare, can lead to costly downtime and extensive repairs. Manufacturers employ anomaly detection systems to alert potential mechanical failures before they happen. The U.S. Office of Nuclear Energy has shown in this report [1] how anomaly detection techniques are applied as a **proactive approach** to nuclear power plants. The proactive approach to detect abnormal situations can give engineers the precious lead time to react and prevent crises from happening. Although a proactive approach to guarding against potential issues is widely supported, **too many false alerts** are problematic because they can lead to time-consuming inspections and delay necessary investigations. The trade-off between a proactive approach and too many false alerts highlights an important criterion for an anomaly detection model: we need a model to **deliver a high level of accuracy and a low level of false positives.** The challenge for data scientists is to develop models with accurate predictions and very low levels of false positives.

Anomaly detection is also applied in the insurance industry. Let's see how it is applied.

1.1.2 Insurance Fraud

Insurance fraud can lead to substantial financial losses to an insurance company. Insurance fraud not only imposes extra costs on insurance companies but also financially impacts consumers and businesses. The growth of fraud has become unprecedented in recent years. The **Coalition Against Insurance Fraud** reports that fraud has cost businesses and consumers $308.6 billion a year [2]. The FBI estimates that fraud costs the average family between $400 and $700 a year in premiums [3].

How does insurance fraud happen? In **auto insurance**, fraudsters might stage a car accident and use fake witnesses to falsely accuse someone of causing the crash. Or a car owner may collaborate with someone to steal their car, sell it for parts, and then file a false theft claim. In **fire insurance**, a claimant may exaggerate their losses or even intentionally set fire to claim insurance money. In **theft and burglary insurance**, a claimant may stage a theft to profit from the insurance payout. The National Insurance Crime Bureau [4] has documented many insurance fraud that appeared in headline news, educating the public to prevent insurance fraud.

By implementing anomaly detection techniques, insurance companies can:

- **Reduce False Claims:** Automatically flagging suspicious claims for review helps in preventing payouts for fraudulent activities.
- **Enhance Efficiency:** Streamlining the claims review process by focusing on the most suspicious cases allows for more efficient use of resources.
- **Protect Policyholders:** Ensuring that legitimate claims are processed quickly while fraudulent ones are caught helps maintain trust and satisfaction among policyholders.

Anomaly detection is also important in cybersecurity. Let's see how it helps.

1.1.3 Cybersecurity

Cyber-attacks can occur due to malicious intent, negligent operations, or external malware. Any single attack can destroy the financial assets of enterprises and result in unrecoverable losses. Intrusion detection systems help monitor traffic through firewalls, web gateways, and all networks. Machine learning algorithms are built into these systems to detect any malicious activity. By employing anomaly detection techniques, organizations can:

- **Detect Intrusions:** Identify potential breaches by spotting unusual activity that may indicate unauthorized access.
- **Prevent Data Loss:** Quickly detecting and responding to anomalies can prevent sensitive data from being exfiltrated or compromised.
- **Enhance Threat Response:** Continuously monitoring for anomalies enables faster and more effective responses to potential cyber threats, minimizing damage and recovery time.

Let's understand the characteristics of anomalies themselves.

Sec 1.2 What are the characteristics of anomalies?

Anomalies have three distinct properties: **rarity**, **heterogeneity**, and **evolving**. The rarity of anomalies means the events are so rare that sometimes it is hard to detect one. If an anomaly detection problem is formulated as a standard supervised learning problem, the rare events are the minority class of the target variable. The target variable is extremely imbalanced in such cases and requires a variety of under-sampling or over-sampling techniques to increase the predictability. To address the challenges of rarity, this book includes the unsupervised learning techniques, such as One-Class Support Vector Machine (OC-SVM) in Chapter 6, and the sampling techniques in Chapter 14.

Anomalies are heterogeneous. They can arise from different underlying causes. Let's again take auto insurance fraud as an example. First, anomalies can be staged accidents. Fraudsters might intentionally cause a collision with another vehicle to file a claim. They might even plant witnesses to corroborate their story, making it appear as though the insured driver was at fault. This type of anomaly is characterized by unusual patterns in accident reports and witness statements. Second, anomalies can be exaggerated injuries. A claimant might exaggerate the extent of their injuries to receive a larger payout. Third, anomalies can be false theft claims. A car owner might falsely report their vehicle as stolen to receive an insurance payout. These different types of anomalies show the heterogeneity of anomalies.

Anomalies are evolving. Fraudsters learn or invent new methods to attack systems. Let's use cyber phishing attacks as an example. A phishing attacker can deceive individuals into providing sensitive information, such as usernames, passwords, credit card numbers, or other personal details. As consumers became more aware of these tactics, fraudsters evolved their methods to include sophisticated account takeover schemes. Fraudsters gain access to a user's entire account by compromising login credentials through methods such as social engineering or malware. Once inside, they can change account details and make fraudulent transactions that appear legitimate. Therefore, data scientists continue to innovate new features to capture unseen patterns. Besides manual creation of new features based on domain knowledge, this book introduces **representation learning** in Chapter 15 that discovers new data patterns and automatically creates new features. Representation learning is also used in modeling tasks like image recognition and natural language processing. It can reduce the need for manual feature engineering, and its learned representations often lead to better performance on downstream tasks.

It is noteworthy to mention that every variable has outliers, but not all outliers are fraudulent. For instance, the typical number of gallons of milk purchased in a retail trip is one or two, while a large family might buy more than five gallons. The latter case is an outlier but not fraudulent. That being said, fraudulent activities often fall within the outliers, so our focus will be on these outliers to detect potential fraud. Let's see why is anomaly detection challenging.

Sec 1.3 Why is anomaly detection challenging?

Anomaly detection is about **prediction accuracy** and **model stability**. On prediction accuracy, if a model incorrectly classifies an anomaly as normal and it goes unnoticed, the business could incur significant, unrecoverable losses. Since we don't want to miss a potential fraudulent case,

why don't we include all possible candidates? This will cause a model to **generate too many false alarms**, which will disrupt regular operations and lead to wasted resources.

On model stability, a model may overfit a particular training dataset and fail to generalize to future, unseen data. If a rare event is not just identified by one single model but **multiple models**, the chance that it is a true anomaly is greatly increased. Thus, the use of multiple algorithms, or averaging out the predictions of multiple models can lead to more reliable detection outcomes.

You may raise an important question: Since there are many machine learning techniques, how can we best use the techniques to manage an anomaly detection data science project?

Sec 1.4 When to employ supervised and unsupervised learning techniques?

Is supervised learning better than unsupervised learning for detecting anomalies? The choice depends on the goals of business applications. Because anomalies typically exhibit three properties: rarity, heterogeneity, and evolving, these properties have direct implications for selecting the appropriate modeling algorithms.

Anomalies are evolving and new patterns continue to emerge. Unsupervised learning algorithms are naturally suitable for **detecting new types of outliers** due to their ability to identify patterns without relying on labeled data. On the other hand, supervised learning is highly effective in detecting known types of outliers. The drawback of supervised learning is its inability to detect novel types of anomalies if there are no labeled examples.

Given these considerations, unsupervised learning algorithms are recommended to explore new types of anomalies. These algorithms excel at identifying novel and unexpected patterns without relying on labeled data. Over time, as number of anomalies are collected and verified, supervised learning algorithms can be developed to target these specific types more precisely. Let's see how we can apply to a real case. In the fire claim scenarios we mentioned earlier, there are heterogenous types of fraud: One type might involve staged arson, while another might involve exaggeration of fire loss. A data scientist may develop a supervised learning model to identify any of the two types. These heterogenous types of fraud can all be labeled as the target in a supervised learning model. This labeling allows the supervised learning model to identify these two types of fraud.

Sec 1.5 Manage a data science project with unsupervised- and supervised-learning techniques

Unsupervised learning techniques are employed to discover new data patterns or when there is no target variable to apply supervised learning techniques. The results of an unsupervised learning model typically are a refined subset of data that contains a higher percentage of anomalies compared to the entire dataset. While finding anomalies is often like finding a needle in a haystack, this subset of data is like a refined haystack that has more needles, thus substantially reducing the time and effort needed to discover the true anomalies. After this review process, the discovered anomalies are thoroughly **examined with human feedback**. From a data scientist's perspective, they can be labeled as the target for preparing a supervised learning model.

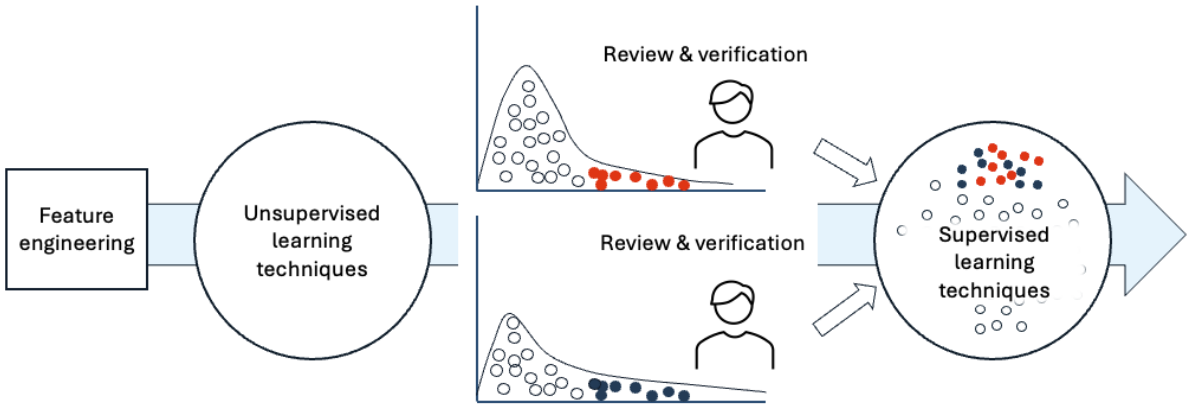


Figure 1.1: Managing a data science anomaly detection project

As more anomalies are verified over time, a labeled dataset of anomalies and non-anomalies can be established, facilitating the development of supervised learning models. These models aim to detect similar cases in future unseen data. The transition from unsupervised learning to supervised learning is illustrated in Figure 1.1. Multiple unsupervised learning techniques can be employed to discover new patterns and generate refined subsets for review. The verified anomalies can then serve as targets for the supervised learning techniques, which seek to enhance the accuracy of detecting similar anomalies.

In summary, we can apply both unsupervised learning techniques and supervised learning techniques for their relative advantages. For this appealing reason, this book include the core techniques of unsupervised and supervised learning techniques. Also, given the rarity property of anomalies, the under-sampling and over-sampling techniques are presented in this book to equip you to build effective models.

Sec 1.6 What are the selection criteria for techniques in this book?

With the wealth of anomaly detection techniques, this book applies three selection criteria to select the unsupervised learning techniques in the book. Figure 1.2 summarizes the techniques according to the three criteria. The first criterion is fast algorithm speed. Fast algorithms can facilitate the iterative processes of model training and assessment. These efficient algorithms include Empirical Cumulative Outlier Detection (ECOD), Histogram-Based Outlier Score (HBOS), and Isolation Forest (IForest).

The second criterion for including an algorithm in this book is popularity. Algorithms like Principal Component Analysis (PCA), K-nearest Neighbors (KNN), Gaussian Mixture Models (GMM), and One-Class Support Vector Machine (OCSVM) are well-known and widely used. Many readers may have encountered these algorithms in their statistics courses. Their popularity not only makes them more accessible and easier to understand but also enhances their adoption in real-world applications.

The third criterion for including algorithms in this book is their connection with other techniques. By learning these algorithms, you can apply the underlying principles to other variations.

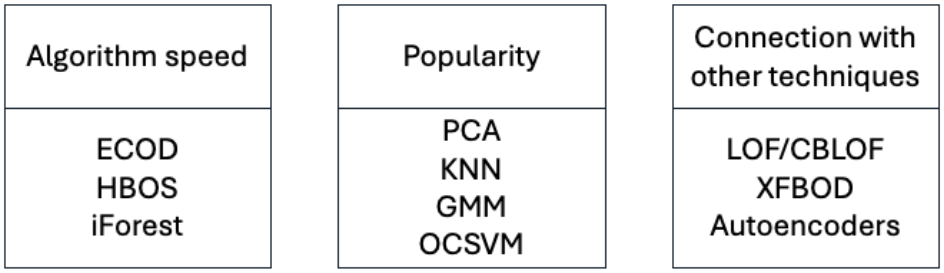


Figure 1.2: Selection criteria for the unsupervised learning techniques in this book

For example, Local Outlier Factor (LOF) and Clustering-Based Local Outlier Factor (CBLOF) utilize the KNN algorithm. Extreme Gradient Boosting-Based Outlier Detection (XGBOD) uses features through representation learning in a supervised learning framework. You can be inspired by its framework in other supervised learning models. Additionally, autoencoders illustrate key concepts that are transferable to many neural network-based models.

Further, anomaly detection models can be classified into three main categories: **proximity-based**, **distribution-based**, and **ensemble-based algorithms**.

First, proximity-based algorithms measure the nearness or closeness between data points, operating under the intuition that anomalies are rare events that are distant from others. Examples of proximity-based algorithms include K-nearest Neighbors (KNN), Isolation Forest (IForest), One-Class Support Vector Machine (OCSVM), Local Outlier Factor (LOF), and Clustering-Based Local Outlier Factor (CBLOF).

Second, distribution-based algorithms fit the distribution of a variable to identify outliers. These algorithms leverage the concept of data distribution to detect anomalies. Examples include Histogram-Based Outlier Score (HBOS), Empirical Cumulative Outlier Detection (ECOD), and Gaussian Mixture Models (GMM).

Finally, ensemble-based algorithms combine multiple models to enhance accuracy. These algorithms, such as IForest and Extreme Gradient Boosting-Based Outlier Detection (XGBOD), use ensemble techniques to achieve better performance. Figure 1.3 groups these proximity-based, distribution-based, and ensemble-based techniques.

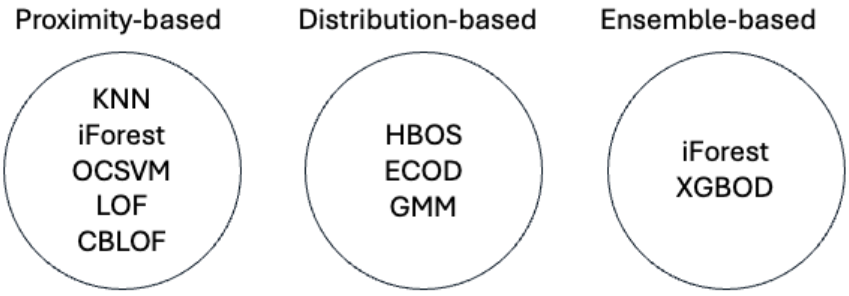


Figure 1.3: Types of unsupervised learning algorithms

Sec 1.7 The algorithms in this book are for multivariate data

While there are many different data types, here we highlight three common types of data: multivariate data, serial data, and image data.

Multivariate Data: This is perhaps the most common type of data and is represented as a data matrix where each row corresponds to an individual observation, and each column represents a different variable. For example, a dataset of customer transactions might include columns for age, income, transaction amount, and purchase frequency, with each row representing a different customer. Anomaly detection in this context might involve identifying unusual patterns in spending behavior or income distributions.

Serial Data: This category includes any single-variable data that is ordered or follows a sequence. Examples are time series data, which tracks changes over time like stock prices, weather patterns, or sensor readings; text sentences, where each word or phrase is analyzed in order; and voice stream data, which involves analyzing audio signals over time. Anomaly detection in time series data, for example, might involve spotting sudden spikes or drops in stock prices that deviate from expected trends.

Image Data: Image data consists of digital pictures or visual information captured by electronic devices such as cameras or scanners. This type of data includes photographs, scanned documents, or medical imaging. Anomaly detection in image data might involve identifying defects in manufacturing processes through visual inspection or detecting unusual patterns in medical scans that could indicate abnormalities.

The algorithms discussed in this book are designed for multivariate data. These algorithms excel in handling complex datasets with multiple variables and are well-suited for identifying anomalies in such structured data.

Sec 1.8 How Is This Book Organized?

This book covers the major algorithms in supervised learning and unsupervised learning techniques for anomaly detection. Each part equips you with the knowledge and skills to manage anomaly detection projects:

Part I: Unsupervised learning techniques for anomaly detection

Part II: Supervised learning techniques for anomaly detection

In Part I, this book selects algorithms from proximity-based, distribution-based, and ensemble-based algorithms. These algorithms will enable readers to expand to other related algorithms. In Part II, this book covers ensemble methods, and regularization techniques, then introduces the over-sampling and under-sampling techniques. The goal is to guide you to build and fine-tune a suite of supervised learning models including Random Forests, Gradient Boosting, XGB, and Deep Learning all together.

We build the contents to be cumulative **from basic concepts to more complex treatments**. In Part I, we embed the ensemble concept and practice in almost every chapter for model stability. This leads to the ensemble learning in Part II. For example:

- The concept of "Forest" in Chapter 4: *Isolation Forest* is bridged to *Random Forest* in Chapter 12.
- The concept of neural networks in Chapter 11: *Autoencoders* prepares a deeper discussion in Chapter 12: *Supervised Learning Primer* of Part II.
- The KNN algorithm in Chapter 8 is used extensively in other models such as LOF in Chapter 9 and CBLOF in Chapter 10. It is also used in under-sampling techniques in Chapter 14.
- The eXtreme Gradient Boosting (XGB) model covered in Chapter 13 are used in representation learning and XGBOD in Chapter 15.
- The outlier scores in Chapter 2 to 10 are used as the features in representation learning and XGBOD in Chapter 15.

Let me introduce each chapter.

Chapter 1: *Introduction* discusses the characteristics of anomalies. We outline the strengths and limitations of supervised and unsupervised learning, and how to manage a data science anomaly detection project. After completing this chapter, you will have a comprehensive understanding for the choice of techniques.

Chapter 2: *Histogram-Based Outlier Score (HBOS)* presents this intuitive and efficient unsupervised method — HBOS. This chapter also details the danger of overfitting and the use of aggregation methods to mitigate overfitting.

Chapter 3: *Empirical Cumulative Distribution-based Outlier Detection (ECOD)* explains how the ECOD algorithm works. You will find ECOD is very approachable like HBOS.

Chapter 4: *Isolated Forest* explains how IForest works, and how IForest differs from other modeling approaches. IForest has a similar approach to the data sampling strategy to Random Forests to build multiple models.

Chapter 5: *Principal Component Analysis (PCA)* covers this well-known unsupervised learning technique and its application for anomaly detection. you shall get a renewed understanding of this familiar method.

Chapter 6: *One-Class Support Vector Machine (OC-SVM)* first explains what SVM is, then its extension to OC-SVM. OC-SVM is especially valuable when outliers are too rare to find. For example, mechanical failure of equipment may only happen once or twice a year. OC-SVM can be a good modeling technique.

Chapter 7: *Gaussian Mixture Model (GMM)* details the fundamental principles of GMM, how it relates to the K-means clustering algorithm, and how to solve GMM using the Expectation-Maximization algorithm. You will gain sound knowledge through the visualization in this chapter.

Chapter 8: *K-nearest Neighbors (KNNs)* starts with an explanation for why KNNs can be used in a supervised or unsupervised learning setting. Then it explains how KNN is applied for anomaly detection. This chapter will give you a fresh look at KNN.

Chapter 9: *Local Outlier Factor (LOF)* explores the details of LOF, its motivation, and its practical application. Its idea is that an outlier will have a significantly lower local density compared to its neighbors. So by observing the low density around a data point, we can identify it as a potential outlier.

Chapter 10: *Cluster-Based Local Outlier Factor (CBLOF)* incorporates the cluster concept with LOF of Chapter 9. Sometimes it is more intuitive to consider data in clusters rather than individual points. You shall understand the strengths and limitations of LOF, and how CBLOF overcomes the limitations.

Chapter 11: *Autoencoders* gives a comprehensive survey of the feedforward neural neural framework, which is the backbone of autoencoders. This chapter bridges from a regression perspective to neural network frameworks. The neural network knowledge prepares you to navigate to the neural network models in supervised learning in Chapter 12. Again, it is the design of this book to build data science principles step by step for an efficient learning outcome.

Chapter 12: *Supervised Learning Primer* presents a comprehensive set of supervised learning models, including Random Forests, Gradient Boosting, and Deep Learning. This chapter explains hyperparameters and uses grid search for hyperparameter tuning.

Chapter 13: *Regularization* explains the problem of overfitting with regression examples, the bias-variance tradeoff, and the $L1$ and $L2$ regularization in GLMs and XGB. Together with Chapter 12, you will be able to build a suite of supervised learning models with complex regularization effectively.

Chapter 14: *Sampling Techniques for Extremely Imbalanced Data* explores the principles and practices of under-sampling and over-sampling techniques, including their benefits and limitations. Because the rare class in the target variable of a supervised learning model is typically extremely imbalanced and posts modeling challenges, you will be better equipped to develop robust machine learning models that perform well even in the presence of highly imbalanced datasets.

Chapter 15: *Representation Learning for Outlier Detection* introduces Representation Learning. Representation learning focuses on discovering systematic representations for raw data autonomously, without human intervention. All the unsupervised learning techniques in the previous chapters, including HBOS, ECOD, iForest, PCA, OCSVM, GMM, KNN, LOF, CBLOF, and Autoencoders, can be used for representation learning to produce new scores as data features. The new features are used as the inputs in the supervised learning model to detect anomalies. This chapter introduces a supervised learning technique called XGBOD. (Extreme Gradient Boosting Outlier Detection). This chapter is especially helpful for data scientists who are looking for automatic exploration of new patterns through representation learning.

Sec 1.9 Data science Q&A

This book presents more than 200 data science questions and answers, each chapter is followed by a Q&A chapter for the content covered in that chapter. By practicing these questions, you can review what you have learned, or even prepare for interviews. They can also be a basis for discussions in study groups. For professionals, the Q&A section serves as a continuous learning tool.

Questions marked with an asterisk (*) can be used as general interview questions. Questions marked with a plus sign (+) can be used as specific technical questions. Because data science has processed rapidly in recent years, many mid-level questions can be considered fundamental data science questions. For example, the preceding intermediate questions can be considered general data science questions and are labeled with an asterisk (*) :

(*) Q. What is the purpose of a loss function in machine learning?

(*) Q. Explain how Stochastic Gradient Descent (SGD) works and why it is widely used.

(*) Q. How can you identify if a model is overfitting?

(*) Q. What is a Precision-Recall (PR) curve, and how does it differ from an ROC curve?

If you use this book as your resource for interview preparation, you may appreciate the intermediate level that this book aims at.

Sec 1.10 Download the code examples

The Python notebooks for all chapters are available for download at:
https://github.com/dataman-git/Handbook-of-anomaly-detection
If there's an update to the code, it will be updated in the GitHub repository.

Sec 1.11 More publicly available outlier detection datasets

If you are looking for public datasets for model training, a well-organized one is the ODDS (Outlier Detection Datasets) (http://odds.cs.stonybrook.edu) [7]. It contains many datasets that represent various domains and types of data, such as multivariate data, time series data, time series graph data, video data, and cybersecurity data. Some of the datasets have labeled targets. Therefore the datasets in ODDS enable supervised or unsupervised learning model development. An example in ODDS is the Wine dataset in ODDS. It contains 13 wine attributes and 3 classes. This dataset has the target label so is suitable for supervised learning. Interested readers can reference Sathe and Aggarwal [8] for their target creation. Another one is the MNIST dataset of handwritten digits in ODDS (http://odds.cs.stonybrook.edu/mnist-dataset/) with an additional variable. The MNIST dataset is a well-known dataset in the data science community. It contains handwritten digits originally provided by the U.S. Postal Service. Different from the regular MNIST dataset, this dataset has the target labels "1" for outliers and "0" for inliers. It is suitable for supervised learning. Interested readers can reference Bandaragoda, Tharindu R., et al. [6] for its target creation.

Sec 1.12 Software requirements

As the tagline in its GitHub (https://github.com/yzhao062/pyod) says, "the Python Outlier Detection Module is a Comprehensive and Scalable Python Library for Outlier Detection (Anomaly Detection)". **PyOD** provides unified APIs for 55+ algorithms and continues to expand. Since its introduction in 2017, it has received wide attention and more than 8 million downloads. The main contributors are Yue Zhao, Zain Nasrullah, and Zheng Li [5], [6].

This book uses the **Combo** library that combines machine learning model scores. Model combination has been widely used in real-world applications and is a subset of ensemble learning.

We introduce the ensemble learning concept for model stability and apply the combo library in many of the supervised learning techniques in the book. Autoencoders are neural-network-based algorithms and require **TensorFlow**. To prevent unintended consequences, the PyOD does not install TensorFlow automatically. This is explained on PyOD's installation page [9]. The installation of tensorflow is included in Chapter 12: *Autoencoders*.

We use the **H2O** platform to demonstrate supervised learning techniques in Chapters 12 and 12. H2O is an open-source, in-memory, distributed machine learning and predictive analytics platform designed for building models on big data. It supports a wide range of machine learning algorithms, including deep learning, generalized linear models, and gradient boosted machines. H2O is known for its scalability and speed, making it suitable for enterprise environments. Its almost one-line of code style makes it a good choice for us, while we can spend more time on the theories in this book.

For an effective learning process, we recommend **Google Colab** to execute the Python Notebooks in this book. Google Colab, short for Colaboratory, is a platform that enables users to write and execute Python code through a web browser. We recommend Google Colab mainly because of its pre-installed libraries, and compatibility of most library versions for the execution of complex code. It offers free access to powerful computational resources including GPUs and TPUs. Colab supports collaboration by allowing multiple users to work on the same notebook simultaneously, share code easily, and integrate with Google Drive for seamless file management.

References

1. "Light Water Reactor Sustainability Program Process Anomaly Detection for Sparsely Labeled Events in Nuclear Power Plants," PRS/EXT-21–01409 (inl.gov)
2. Hallo, Steve (2022). "$308.6 billion: The real annual cost of insurance fraud in the U.S.", the PropertyCasualty360, https://www.propertycasualty360.com/2022/08/25/308-6-billion-the-real-annual-cost-of-insurance-fraud-in-the-u-s/?slreturn=20220817092131
3. Reports and Publications: Insurance Fraud. the Federal Bureau of Investigation. https://www.fbi.gov services/publications/insurance-fraud
4. The National Insurance Crime Bureau, https: www.nicb.org
5. Zhao, Y., Nasrullah, Z. and Li, Z., 2019. PyOD: A Python Toolbox for Scalable Outlier Detection. Journal of machine learning research (JMLR), 20(96), pp.1–7.
6. https://github.com/yzhao062/pyod
7. Shebuti Rayana (2016). ODDS Library [http://odds.cs.stonybrook.edu]. Stony Brook, NY: Stony Brook University, Department of Computer Science.
8. Saket Sathe and Charu C. Aggarwal. LODES: Local Density meets Spectral Outlier Detection. SIAM Conference on Data Mining, 2016.
9. https://pyod.readthedocs.io/en/latest/install.html

INTRODUCTION - Q&A

Reading time: 15 minutes

Questions marked with an asterisk (*) can be used as general interview questions.
Questions marked with a plus sign (+) can be used as specific technical questions.

(*) Question 1: What is an anomaly in data analysis?

Answer: An anomaly, also known as an outlier, refers to a data point or pattern that deviates significantly from the expected behavior or norm. It can represent something rare, unusual, or unexpected in a dataset. Anomalies are critical to identify because they may indicate errors, rare events, or novel insights. For example, in financial transactions, anomalies could signal fraudulent activity, while in manufacturing, they might indicate potential equipment failure.

(*) Question 2: What are the three main properties of anomalies?

Answer: Anomalies typically possess three distinct properties: rarity, heterogeneity, and evolving nature.

- **Rarity:** Anomalies are rare events in a dataset, making them challenging to detect.
- **Heterogeneity:** Anomalies can arise from diverse underlying causes, presenting different patterns and characteristics. For example, fraud in insurance claims can manifest in various forms, such as staged accidents or exaggerated injuries.
- **Evolving:** Anomalies can evolve as new methods or technologies emerge, such as more sophisticated cyber-attack techniques. This evolving nature requires continuous updating of detection models to identify new types of anomalies effectively.

(*) Question 3: What is the difference between an outlier and a fraudulent event?

Answer: An outlier is a data point that deviates significantly from the norm but isn't necessarily indicative of fraud. For example, a large purchase of milk by a big family is an outlier but not fraudulent. A fraudulent event, on the other hand, specifically involves intentional deception for financial gain, often falling within the range of outliers. In data science, it is crucial to distinguish between outliers and fraudulent activities to avoid false accusations and ensure accurate detection of actual fraud.

(*) Question 4: What makes anomaly detection challenging?

Answer: Anomaly detection is challenging due to the need for high prediction accuracy and model stability. Misclassifying an anomaly as normal can lead to significant losses, while over-detecting anomalies can result in too many false positives, disrupting operations. Additionally, anomalies are often rare and diverse, making it difficult to gather enough labeled data for supervised learning. The evolving nature of anomalies, such as new fraud techniques, further complicates the task, requiring continuous model updates.

(*) Question 5: How can multiple algorithms improve the stability of anomaly detection models?

Answer: Using multiple algorithms or ensemble methods can improve the stability of anomaly detection models by reducing the likelihood of false positives and enhancing the accuracy of predictions. By combining the outputs of several models, it becomes more likely that true anomalies are accurately identified, as different models may capture different aspects of the data. This approach also helps to mitigate the impact of noise and random fluctuations, leading to more reliable detection outcomes.

(*) Question 6: When can supervised learning techniques be employed in anomaly detection?

Answer: Supervised learning techniques should be employed in anomaly detection when there is sufficient labeled data representing known anomalies. These techniques excel at detecting specific types of anomalies with high accuracy by learning from labeled examples.

(*) Question 7: When is unsupervised learning more suitable for anomaly detection?

Answer: Unsupervised learning is more suitable for anomaly detection when there is a lack of labeled data or when the anomalies are novel and previously unseen. Unsupervised algorithms can identify unusual patterns and deviations without relying on labeled examples, making them effective for discovering new types of outliers. This approach is particularly useful in dynamic environments where the nature of anomalies can evolve, such as emerging fraud techniques or new cybersecurity threats.

(*) Question 8: What benefits do unsupervised learning techniques provide in the identification of anomalies within a dataset?

Answer: Unsupervised learning techniques are used to discover new data patterns, particularly when there is no target variable for supervised learning. The results of these techniques typically yield a refined subset of data that contains a higher percentage of anomalies compared to the entire dataset. This refined subset is likened to a "refined haystack" with more "needles," making it significantly easier and quicker to identify true anomalies.

(*) Question 9: How can transitioning from unsupervised to supervised learning benefit anomaly detection projects?

Answer: Transitioning from unsupervised to supervised learning in anomaly detection projects can enhance the accuracy and specificity of identifying known types of anomalies. Unsupervised learning helps discover a wide range of potential anomalies without needing labeled data. As these anomalies are verified and labeled, a dataset is built, allowing for the development of supervised models. These models can then target specific types of anomalies with greater precision, leveraging the labeled data for more accurate detection in future cases.

(*) Question 10: What is supervised learning?

Answer: Supervised learning is a machine learning approach where a model is trained using a labeled dataset, which means each input is paired with the correct output. The goal is for the model to learn the mapping from inputs to outputs so it can predict the output for new, unseen inputs accurately. Common supervised learning tasks include classification, where the model predicts discrete labels, and regression, where it predicts continuous values. Examples of supervised learning algorithms include linear regression, logistic regression, decision trees, support vector machines, and neural networks.

(*) Question 11: What is unsupervised learning?

Answer: Unsupervised learning is a type of machine learning where the model is trained on data without labeled responses. Instead of being given explicit instructions on what to predict, the model identifies patterns and structures within the data. Common tasks in unsupervised learning include clustering, where the model groups similar data points together, and association, where the model identifies relationships between variables. This approach is often used for exploratory data analysis, market basket analysis, customer segmentation, and anomaly detection.

(*) Question 12: Why is it important to use multiple algorithms for anomaly detection?

Answer: Using multiple algorithms can improve detection accuracy and reduce false positives. If multiple models identify the same anomaly, it is more likely to be a true anomaly. This approach helps uncover hidden data patterns more reliably.

(*) Question 13: How does the rarity of anomalies affect supervised learning models?

Answer: The rarity of anomalies leads to class imbalance in supervised learning settings. This imbalance can cause bias toward the majority class, reducing classification performance and increasing false negatives. Techniques like over-sampling and under-sampling are necessary to address this issue.

(+) Question 14: What does it mean "proximity-based" in data science algorithm?

Answer: In data science, "proximity-based" refers to algorithms that measure the closeness or distance between data points to identify patterns, clusters, or anomalies. These algorithms operate under the assumption that similar data points will be closer together, while dissimilar points will be farther apart. Proximity-based methods use distance metrics such as Euclidean distance, Manhattan distance, or cosine similarity to quantify this closeness. Common examples include k-nearest Neighbors (KNN) or Local Outlier Factor (LOF).

(+) Question 15: What does it mean "distribution-based" in data science algorithm?

Answer: Distribution-based algorithms fit the probability distribution of a variable to identify outliers. Examples of distribution-based algorithms include Histogram-Based Outlier Score (HBOS), Empirical Cumulative Outlier Detection (ECOD), and Gaussian Mixture Models (GMM).

(*) Question 16: How can companies benefit from early detection of anomalies?

Answer: Companies can benefit from early detection of anomalies by preventing potential issues before they escalate into significant problems, thereby saving costs and improving operational efficiency. Early anomaly detection allows for timely maintenance and repair, reducing unexpected downtime and extending the lifespan of equipment. It enhances safety by identifying risks before they lead to accidents, protecting employees and assets. In fields like cybersecurity, early detection of anomalies can prevent data breaches and mitigate the impact of cyber-attacks. Additionally, in financial sectors, identifying fraudulent activities early helps in minimizing financial losses and maintaining customer trust.

(*) Question 17: How does anomaly detection apply to mechanical failure detection?

Answer: Anomaly detection is crucial in business operations because it helps identify rare and unusual events that can lead to significant financial losses or operational disruptions. By detecting anomalies early, companies can take preventive measures to mitigate risks, such as fraud, equipment failure, or cybersecurity threats. This proactive approach not only protects financial assets but also enhances efficiency and maintains customer trust by ensuring the timely resolution of issues.

(*) Question 18: What is a common challenge in anomaly detection for mechanical failures?

Answer: A common challenge is achieving a balance between accuracy and minimizing false positives. While it's crucial to detect potential issues early, generating too many false alerts can be problematic. False positives can lead to unnecessary inspections, wasted resources, and delays in addressing actual problems. Therefore, developing models that accurately identify true anomalies while maintaining a low rate of false positives is vital for efficient operations.

(*) Question 19: How does anomaly detection help in detecting insurance fraud?

Answer: Anomaly detection assists in identifying unusual patterns in insurance claims that may indicate fraudulent activities. For example, in auto insurance, anomaly detection can flag staged accidents or exaggerated injury claims. By automatically identifying suspicious claims, insurance companies can prevent fraudulent payouts, protect policyholders, and enhance the efficiency of the claims review process. This technology helps companies focus on the most suspicious cases, ensuring that legitimate claims are processed quickly.

Part I

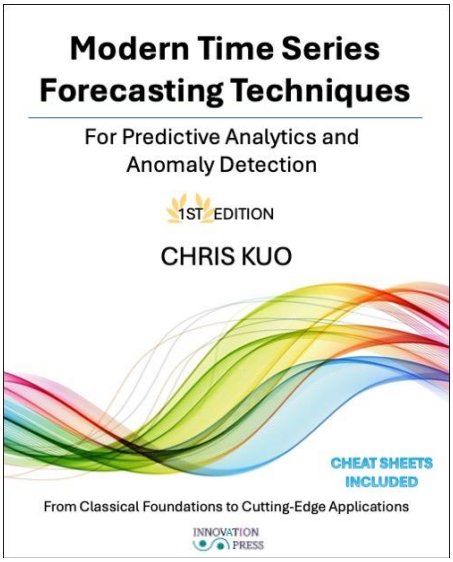
PART I: Unsupervised Learning Techniques

INDEX

Other Books by Chris Kuo

Modern Time Series Forecasting: For Predictive Analytics and Anomaly Detection. https://a.co/d/4L8Za4r

2024



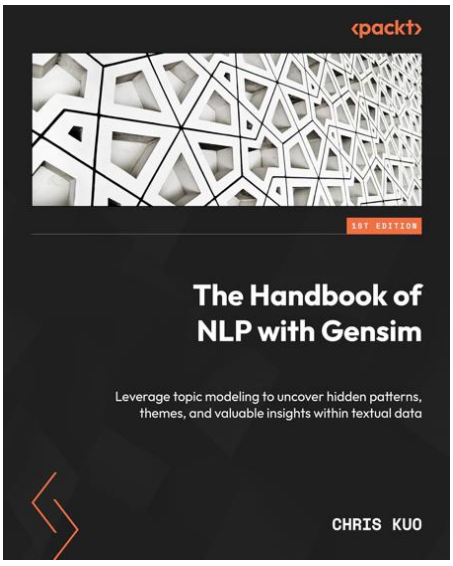
Key features

This book offers a comprehensive exploration of methodologies, algorithms, and practical applications, catering to both seasoned data scientists looking to refine their skills and newcomers eager to begin their analytical journey. Equipped with this resource, you'll confidently decipher intricate temporal patterns and harness predictive capabilities across various domains.

- Part 1: From Prophet to NeuralProphet
- Part 2: Getting probabilistic forecasts
- Part 3: Autoregressive-based Time Series Techniques
- Part 4: Tree-based Time Series Techniques
- Part 5: Deep into Deep learning-based Time Series Techniques
- Part 6: Transformer-based Time Series Techniques

The Handbook of NLP with Gensim: Leverage topic modeling to uncover hidden patterns. https://a.co/d/ifO2LxR

2023



Key features

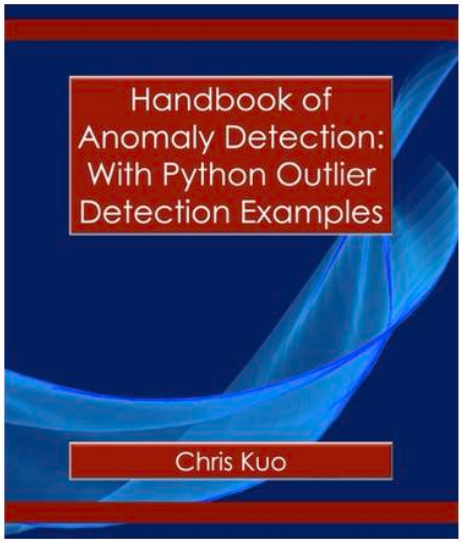
This book demystifies NLP and equips you with hands-on strategies spanning healthcare, e-commerce, finance, and more to enable you to leverage Gensim in real-world scenarios. You'll begin by exploring motives and techniques for extracting text information like bag-of-words, TF-IDF, and word embeddings. This book guides you on topic modeling using methods such as Latent Semantic Analysis (LSA) for dimensionality reduction and discovering latent semantic relationships in text data, Latent Dirichlet Allocation (LDA) for probabilistic topic modeling, and Ensemble LDA to enhance topic modeling stability and accuracy. By the end of this book, you'll have mastered the techniques essential to create applications with Gensim and integrate NLP into your business processes.

Other Books by Chris Kuo

Handbook of Anomaly Detection: With Python Outlier Detection: Build and modernize your anomaly detection with examples

https://a.co/d/hHuHJEV

2023



Key features

Learn a wide range of proven techniques that cover proximity-based, distribution-based, and ensemble-based algorithms. Enjoy systematic learning of modern techniques through appealing visualization of complex concepts. Become an experienced data scientist in outlier detection by hands-on code examples

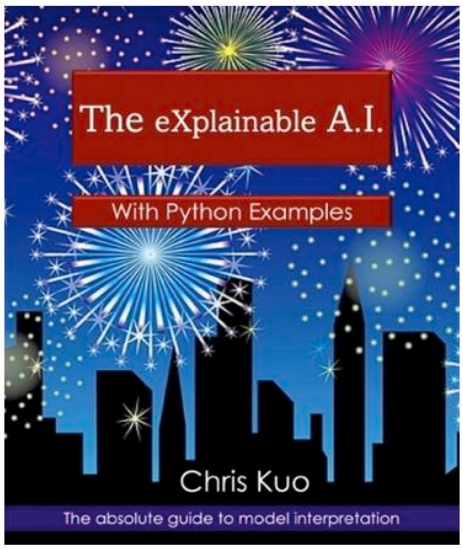
Developer review

"Anomaly detection is a critical technique to identify rare items in risk modeling, security, and healthcare. Dr. Kuo's book provides hands-on guidance on how to use existing tools, such as PyOD, to leverage this technique in your daily data science work. More importantly, this book reviews more than 10 leading detection algorithms, with both algorithm descriptions and detailed code examples." ~ Yue Zhao, Principal Developer of PyOD

The eXplainable A.I.: With Python examples

https://a.co/d/i8htOES

2022



Key features

- Learn algorithms with easy-to-understand examples.
- Learn what explainability A.I. is
- Learn SHAP to explain your machine learning models
- Learn SHAP with H2O Models. Learn SHAP Kernel Explainer
- Learn Microsoft's interpretML
- Learn how to use LIME to explain your models.

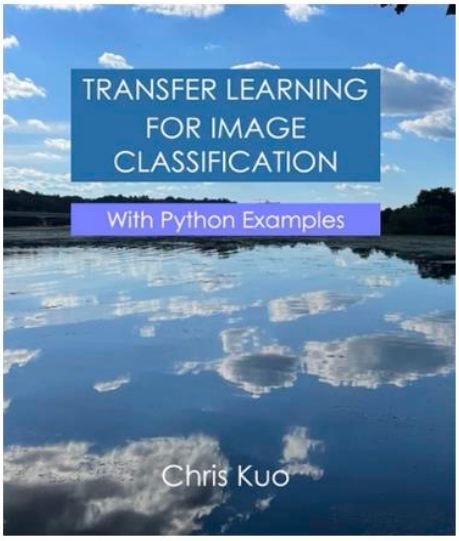
Who this book is for

This book is for anyone who wants to understand model explainability and get your complex machine-learning models adopted. Although the content is extensive, data science professionals who do not have much machine learning background will find this book accessible

Other Books by Chris Kuo

Transfer Learning for Image Classification: With Python examples. https://a.co/d/e3MuX0e

2022



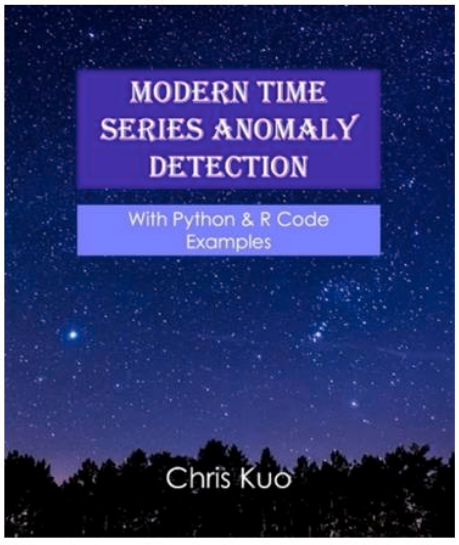
What you will learn

- Learn how deep learning models treat image data.
- Learn what a convolutional neural network (CNN) is.
- Learn each layer of a CNN by visualizing what it sees in an image layer-by-layer
- Learn the development of pre-trained image models
- Learn how to annotate images programmatically and pre-process images for modeling
- Follow Step 1,2,3 in the book to build deep learning models with Keras
- Build a repeatable pipeline to apply transfer learning for any image projects

Book Description

The best way of learning is to understand the motives of a technique and then practice with code examples. Thus, this book explains the reasons, describes a method and then shows the code examples. This book presents large-scale pre-trained image models and how to apply transfer learning techniques to build your models.

2021



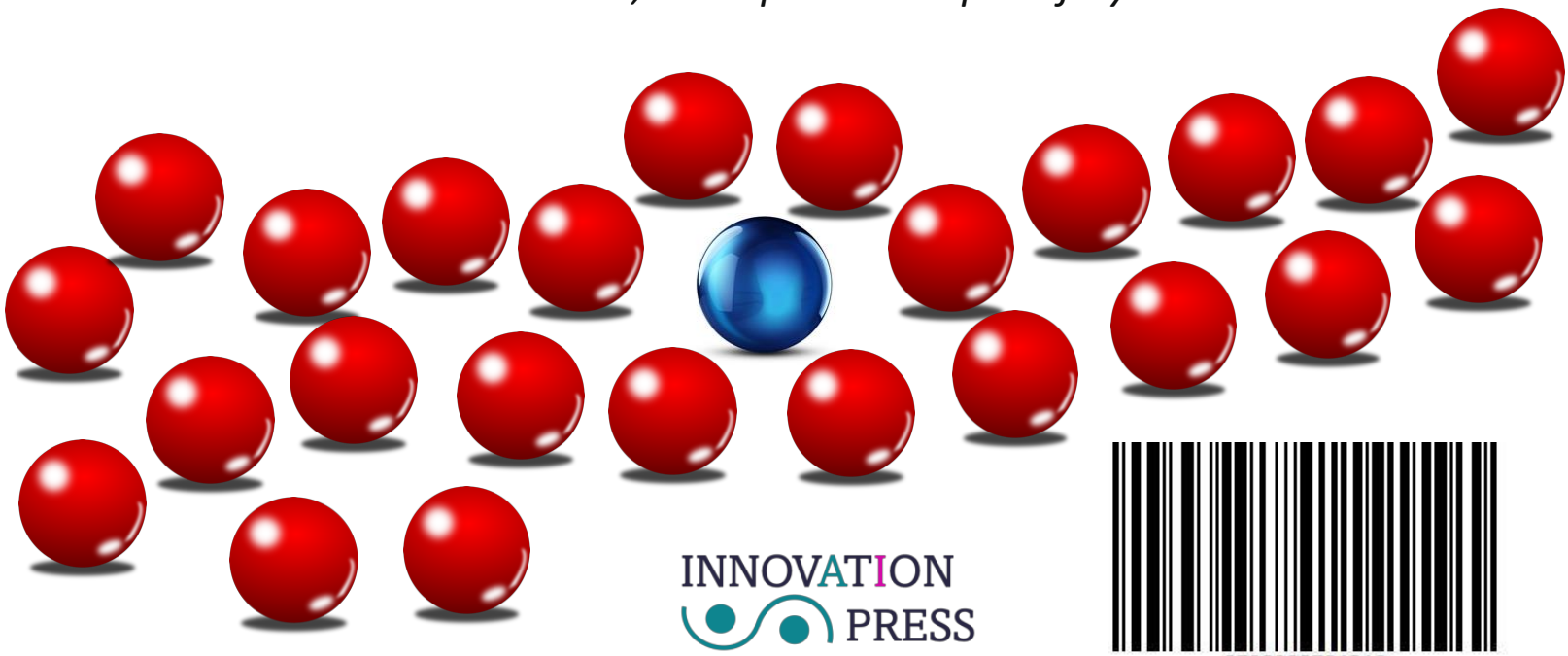
Modern Time Series Anomaly Detection: With Python and R examples (Out of stock)

Handbook of Anomaly Detection

The second edition offers a comprehensive and up-to-date guide that addresses both foundational and advanced topics in anomaly detection. It expands upon the first edition by including a deep dive into modern machine learning techniques, covering both supervised and unsupervised methods with detailed Python code examples. Readers will find enhanced visual presentations that simplify complex concepts, making it easier to grasp the intricacies of algorithms. This edition also provides over 200 questions and answers, offering a unique and practical tool for professionals preparing for data science roles. The step-by-step explanations help bridge the gap between understanding algorithms and implementing them. The second edition stands as a vital resource for students, professionals, and instructors alike. This book will provide the tools and knowledge you need.

"Anomaly detection is a critical technique to identify rare items in risk modeling, security, and healthcare. Dr. Kuo's book provides hands-on guidance on how to use existing tools, such as PyOD, to leverage this technique in your daily data science work. More importantly, this book reviews more than 10 leading detection algorithms, with both algorithm descriptions and detailed code examples."

~ Yue Zhao, Principal Developer of PyOD



INNOVATION PRESS

9798990781016