1. Record Nr. UNINA9910522997303321

Autore	Nandi Anirban
Titolo	Interpreting machine learning models : learn model interpretability and explainability methods / / Anirban Nandi and Aditya Kumar Pal
Pubbl/distr/stampa	New York, New York : , : Apress, , [2022] ©2022
ISBN	1-5231-5100-5 1-4842-7802-X
Descrizione fisica	1 online resource (355 pages)
Disciplina	006.31
Soggetti	Machine learning - Mathematical models
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Nota di bibliografia	Includes bibliographical references and index.
Nota di contenuto	Chapter 1: The Evolution of Machine Learning -- Chapter 2: Introduction to Model interpretability. -- Chapter 3: Machine Learning Interpretability Taxonomy -- Chapter 4: Common Properties of Explanations Generated by Interpretability Methods -- Chapter 5: Human Factors in Model Interpretability -- Chapter 6: Explainability Facts: A Framework for Systematic Assessment of Explainable Approaches -- Chapter 7: Interpretable ML and Explainable ML Differences -- Chapter 8: Framework of Model Explanations -- Chapter 9: Feature Importance methods -- Details and usage examples -- Chapter 10: Detailing rule-based methods -- Chapter 11: Detailing Counterfactual Methods -- Chapter 12: Detailing Image interpretability methods -- Chapter 13: Explaining text classification models -- Chapter 14: Role of Data in Interpretability -- Chapter 15: The 8 pitfalls of explainability methods -- Conclusion. -- References.
Sommario/riassunto	Understand model interpretability methods and apply the most suitable one for your machine learning project. This book details the concepts of machine learning interpretability along with different types of explainability algorithms. You'll begin by reviewing the theoretical aspects of machine learning interpretability. In the first few sections you'll learn what interpretability is, what the common properties of interpretability methods are, the general taxonomy for classifying

methods into different sections, and how the methods should be assessed in terms of human factors and technical requirements. Using a holistic approach featuring detailed examples, this book also includes quotes from actual business leaders and technical experts to showcase how real life users perceive interpretability and its related methods, goals, stages, and properties. Progressing through the book, you'll dive deep into the technical details of the interpretability domain. Starting off with the general frameworks of different types of methods, you'll use a data set to see how each method generates output with actual code and implementations. These methods are divided into different types based on their explanation frameworks, with some common categories listed as feature importance based methods, rule based methods, saliency maps methods, counterfactuals, and concept attribution. The book concludes by showing how data effects interpretability and some of the pitfalls prevalent when using explainability methods. What You'll Learn Understand machine learning model interpretability Explore the different properties and selection requirements of various interpretability methods Review the different types of interpretability methods used in real life by technical experts Interpret the output of various methods and understand the underlying problems Who This Book Is For Machine learning practitioners, data scientists and statisticians interested in making machine learning models interpretable and explainable; academic students pursuing courses of data science and business analytics.