Machine learning has great potential for improving products, processes and research. But computers usually do not explain their predictions which is a barrier to the adoption of machine learning. This book is about making machine learning models and their decisions interpretable.

After exploring the concepts of interpretability, you will learn about simple, interpretable models such as decision trees, decision rules and linear regression. The focus of the book is on model-agnostic methods for interpreting black box models such as feature importance and accumulated local effects, and explaining individual predictions with Shapley values and LIME. In addition, the book presents methods specific to deep neural networks.

All interpretation methods are explained in depth and discussed critically. How do they work under the hood? What are their strengths and weaknesses? How can their outputs be interpreted? This book will enable you to select and correctly apply the interpretation method that is most suitable for your machine learning project. Reading the book is recommended for machine learning practitioners, data scientists, statisticians, and anyone else interested in making machine learning models interpretable.

Christoph Molnar is a Machine Learning expert and independent author.

You can learn more about his work on Twitter (@ChristophMolnar) and his website (christophmolnar.com).

Interpretable Machine Learning

Christoph Molnar

Second Edition

Interpretable Machine Learning

A Guide for Making Black Box Models Explainable

Christoph Molnar