

Interpretable Machine Learning

A Guide for Making Black Box Models Explainable

Christoph Molnar

Contents

Preface	1
Introduction	3
Story Time	5
What Is Machine Learning?	13
Terminology	15
Interpretability	19
Importance of Interpretability	19
Taxonomy of Interpretability Methods	26
Scope of Interpretability	28
Evaluation of Interpretability	31
Properties of Explanations	32
Human-friendly Explanations	36
Datasets	43
Bike Rentals (Regression)	43
YouTube Spam Comments (Text Classification)	45
Risk Factors for Cervical Cancer (Classification)	47
Interpretable Models	49
Linear Regression	51
Logistic Regression	71
GLM, GAM and more	79
Decision Tree	102
Decision Rules	110
RuleFit	130
Other Interpretable Models	140

Model-Agnostic Methods	143
Partial Dependence Plot (PDP)	147
Individual Conditional Expectation (ICE)	155
Accumulated Local Effects (ALE) Plot	161
Feature Interaction	184
Feature Importance	193
Global Surrogate	203
Local Surrogate (LIME)	209
Shapley Values	221
Example-Based Explanations	237
Counterfactual Explanations	240
Adversarial Examples	251
Prototypes and Criticisms	263
Influential Instances	275
A Look into the Crystal Ball	295
The Future of Machine Learning	297
The Future of Interpretability	299
Contribute to the Book	303
Citing this Book	305
Acknowledgements	307
References	309
R Packages Used for Examples	313