- Preface

    1. Who Should Read This Book
    2. What Readers Will Learn
    3. Alignment with the NIST AI Risk Management Framework
    4. Book Outline
        1. Part I
        2. Part II
        3. Part III
    5. Example Datasets
        1. Taiwan Credit Data
        2. Kaggle Chest X-Ray Data
    6. Conventions Used in This Book
    7. Online Figures
    8. Using Code Examples
    9. O'Reilly Online Learning
    10. How to Contact Us
    11. Acknowledgments
        1. Patrick Hall
        2. James Curtis
        3. Parul Pandey

- I. Theories and Practical Applications of AI Risk Management
- 1. Contemporary Machine Learning Risk Management

    1. A Snapshot of the Legal and Regulatory Landscape
        1. The Proposed EU AI Act
        2. US Federal Laws and Regulations
        3. State and Municipal Laws
        4. Basic Product Liability
        5. Federal Trade Commission Enforcement
    2. Authoritative Best Practices
    3. AI Incidents
    4. Cultural Competencies for Machine Learning Risk Management
        1. Organizational Accountability
        2. Culture of Effective Challenge
        3. Diverse and Experienced Teams
        4. Drinking Our Own Champagne
        5. Moving Fast and Breaking Things
    5. Organizational Processes for Machine Learning Risk Management
        1. Forecasting Failure Modes
        2. Model Risk Management Processes
        3. Beyond Model Risk Management
    6. Case Study: The Rise and Fall of Zillow's iBuying
        1. Fallout
        2. Lessons Learned
    7. Resources

- 2. Interpretable and Explainable Machine Learning

    1. Important Ideas for Interpretability and Explainability
    2. Explainable Models
        1. Additive Models
        2. Decision Trees
        3. An Ecosystem of Explainable Machine Learning Models
    3. Post Hoc Explanation
        1. Feature Attribution and Importance
        2. Surrogate Models
        3. Plots of Model Performance
        4. Cluster Profiling
    4. Stubborn Difficulties of Post Hoc Explanation in Practice
    5. Pairing Explainable Models and Post Hoc Explanation
    6. Case Study: Graded by Algorithm
    7. Resources

- 3. Debugging Machine Learning Systems for Safety and Performance

    1. Training
        1. Reproducibility
        2. Data Quality
        3. Model Specification for Real-World Outcomes
    2. Model Debugging
        1. Software Testing
        2. Traditional Model Assessment
        3. Common Machine Learning Bugs
        4. Residual Analysis
        5. Sensitivity Analysis
        6. Benchmark Models
        7. Remediation: Fixing Bugs
    3. Deployment
        1. Domain Safety
        2. Model Monitoring
    4. Case Study: Death by Autonomous Vehicle
        1. Fallout
        2. An Unprepared Legal System
        3. Lessons Learned
    5. Resources

- 4. Managing Bias in Machine Learning

    1. ISO and NIST Definitions for Bias
        1. Systemic Bias
        2. Statistical Bias
        3. Human Biases and Data Science Culture
    2. Legal Notions of ML Bias in the United States

3. Who Tends to Experience Bias from ML Systems
4. Harms That People Experience
5. Testing for Bias
    1. Testing Data
    2. Traditional Approaches: Testing for Equivalent Outcomes
    3. A New Mindset: Testing for Equivalent Performance Quality
    4. On the Horizon: Tests for the Broader ML Ecosystem
    5. Summary Test Plan
6. Mitigating Bias
    1. Technical Factors in Mitigating Bias
    2. The Scientific Method and Experimental Design
    3. Bias Mitigation Approaches
    4. Human Factors in Mitigating Bias
7. Case Study: The Bias Bug Bounty
8. Resources

- 5. Security for Machine Learning

1. Security Basics
    1. The Adversarial Mindset
    2. CIA Triad
    3. Best Practices for Data Scientists
2. Machine Learning Attacks
    1. Integrity Attacks: Manipulated Machine Learning Outputs
    2. Confidentiality Attacks: Extracted Information
3. General ML Security Concerns
4. Countermeasures
    1. Model Debugging for Security
    2. Model Monitoring for Security
    3. Privacy-Enhancing Technologies
    4. Robust Machine Learning
    5. General Countermeasures
5. Case Study: Real-World Evasion Attacks
    1. Evasion Attacks
    2. Lessons Learned
6. Resources

- II. Putting AI Risk Management into Action
- 6. Explainable Boosting Machines and Explaining XGBoost

1. Concept Refresher: Machine Learning Transparency
    1. Additivity Versus Interactions
    2. Steps Toward Causality with Constraints
    3. Partial Dependence and Individual Conditional Expectation
    4. Shapley Values
    5. Model Documentation
2. The GAM Family of Explainable Models

        1. Elastic Net–Penalized GLM with Alpha and Lambda Search
        2. Generalized Additive Models
        3. GA2M and Explainable Boosting Machines
    3. XGBoost with Constraints and Post Hoc Explanation
        1. Constrained and Unconstrained XGBoost
        2. Explaining Model Behavior with Partial Dependence and ICE
        3. Decision Tree Surrogate Models as an Explanation Technique
        4. Shapley Value Explanations
        5. Problems with Shapley values
        6. Better-Informed Model Selection
    4. Resources

- 7. Explaining a PyTorch Image Classifier

    1. Explaining Chest X-Ray Classification
    2. Concept Refresher: Explainable Models and Post Hoc Explanation Techniques
        1. Explainable Models Overview
        2. Occlusion Methods
        3. Gradient-Based Methods
        4. Explainable AI for Model Debugging
    3. Explainable Models
        1. ProtoPNet and Variants
        2. Other Explainable Deep Learning Models
    4. Training and Explaining a PyTorch Image Classifier
        1. Training Data
        2. Addressing the Dataset Imbalance Problem
        3. Data Augmentation and Image Cropping
        4. Model Training
        5. Evaluation and Metrics
        6. Generating Post Hoc Explanations Using Captum
        7. Evaluating Model Explanations
        8. The Robustness of Post Hoc Explanations
    5. Conclusion
    6. Resources

- 8. Selecting and Debugging XGBoost Models

    1. Concept Refresher: Debugging ML
        1. Model Selection
        2. Sensitivity Analysis
        3. Residual Analysis
        4. Remediation
    2. Selecting a Better XGBoost Model
    3. Sensitivity Analysis for XGBoost
        1. Stress Testing XGBoost
        2. Stress Testing Methodology
        3. Altering Data to Simulate Recession Conditions

        4. Adversarial Example Search
    4. Residual Analysis for XGBoost
        1. Analysis and Visualizations of Residuals
        2. Segmented Error Analysis
        3. Modeling Residuals
    5. Remediating the Selected Model
        1. Overemphasis of PAY_0
        2. Miscellaneous Bugs
    6. Conclusion
    7. Resources

- 9. Debugging a PyTorch Image Classifier

    1. Concept Refresher: Debugging Deep Learning
    2. Debugging a PyTorch Image Classifier
        1. Data Quality and Leaks
        2. Software Testing for Deep Learning
        3. Sensitivity Analysis for Deep Learning
        4. Remediation
        5. Sensitivity Fixes
    3. Conclusion
    4. Resources

- 10. Testing and Remediating Bias with XGBoost

    1. Concept Refresher: Managing ML Bias
    2. Model Training
    3. Evaluating Models for Bias
        1. Testing Approaches for Groups
        2. Individual Fairness
        3. Proxy Bias
    4. Remediating Bias
        1. Preprocessing
        2. In-processing
        3. Postprocessing
        4. Model Selection
    5. Conclusion
    6. Resources

- 11. Red-Teaming XGBoost

    1. Concept Refresher
        1. CIA Triad
        2. Attacks
        3. Countermeasures
    2. Model Training
    3. Attacks for Red-Teaming

        1. Model Extraction Attacks
        2. Adversarial Example Attacks
        3. Membership Attacks
        4. Data Poisoning
        5. Backdoors
    4. Conclusion
    5. Resources

- III. Conclusion
- 12. How to Succeed in High-Risk Machine Learning

    1. Who Is in the Room?
    2. Science Versus Engineering
        1. The Data-Scientific Method
        2. The Scientific Method
    3. Evaluation of Published Results and Claims
    4. Apply External Standards
    5. Commonsense Risk Mitigation
    6. Conclusion
    7. Resources

- Index