Article

Responsible Machine Learning

Interpretable Models, Post-hoc Explanation, and Disparate Impact Testing

Navdeep Gill [1,‡], Patrick Hall [1,‡,*], Kim Montgomery [1,‡], and Nicholas Schmidt [2,‡]

[1] H2O.ai
[2] BLDS, LLC
* Correspondence: phall@h2o.ai; nschmidt@bldsllc.com
‡ These authors contributed equally to this work.

Version November 1, 2019 submitted to Information

Abstract: This text outlines a viable approach for training and evaluating complex machine learning systems for high-stakes, human-centered, or regulated applications using common Python programming tools. The accuracy and intrinsic interpretability of two types of constrained models, monotonic gradient boosting machines (M-GBM) and explainable neural networks (XNN), a deep learning architecture well-suited for structured data, are assessed on simulated datasets with known feature importance and sociological bias characteristics and on realistic, publicly available example datasets. For maximum transparency and the potential generation of personalized adverse action notices, the constrained models are analyzed using post-hoc explanation techniques including plots of individual conditional expectation (ICE) and global and local gradient-based or Shapley feature importance. The constrained model predictions are also tested for disparate impact and other types of sociological bias using straightforward group fairness measures. By combining innovations in interpretable models, post-hoc explanation, and bias testing with accessible software tools, this text aims to provide a template workflow for important machine learning applications that require high accuracy and interpretability and low disparate impact.

Keywords: Machine Learning; Neural Network; Gradient Boosting Machine; Interpretable; Explanation; Fairness; Disparate Impact; Python

0. Introduction

1. Materials and Methods

1.1. Data Description

1.2. Model Description

1.3. Software Resources

2. Results

2.1. Simulated Data Results

2.2. Loan Data Results

3. Discussion

4. Conclusions

Author Contributions: , N.G.; , P.H.; , K.M.; , N.S.

Funding: This research received no external funding.

Acknowledgments:

Conflicts of Interest:

Abbreviations

The following abbreviations are used in this manuscript: