Interpretable Machine Learning for Criminal Recidivism
Bhrij Patel, Bin Han, Caroline Wang, Cynthia Rudin
Duke University

Abstract
• Some states use black-box models to predict criminal recidivism
• These algorithms impact many lives and transparency is important to understand how risk-scores are computed
• In criminal justice, simple, interpretable, transparent models perform as well as black-box models

Objectives
• Recidivism - re-offense; committing a crime again
• Research interpretable models for criminal recidivism
• Compare performance between black-box models and interpretable models
• Show that simple, interpretable and transparent models perform just as well as black-box models

Methods
• Obtained criminal data from Broward County, FL (public, reproducible) and Kentucky from Department of Shared Services Research and Statistics
• Processed the data to create features of prior criminal history and age
• Created labels for two-year recidivism generally and for specific crimes
• Ran interpretable and black-box models predicting recidivism

Data-Processing
Features Created
• Prior arrest for specific crimes
• Priors times on probation
• Prior times in jail/prison
• Demographic features
  • Age
  • Gender

Crime Labels Created
• Drug
• Traffic
• Property
• Trespass
• DUI
• Sex Offense
• Murder
• Misdemeanor
• Felony
• Violent

Performance
• Performance (measured by AUC) for a given prediction problem stayed roughly the same between all models, black-box and interpretable
• Below is performance for KY two-year recidivism in general

Conclusion
• No real need to use proprietary, black-box models for predicting criminal recidivism
• Simple, interpretable, transparent algorithms can perform just as well as black-box
• Between prediction tests, performance differed, but stayed roughly the same between models

References
• C. Wang, B. Coker, C. Rudin, The age of secrecy and unfairness in recidivism prediction, Manuscript in Preparation (2019)