Intepretable Machine Learning:
Applications, Techniques, and Themes

Patrick Hall

1 Popular Intepretable Modeling Approaches

To build a truly understandable model, best to use an interpretable approach from the beginning of your project.

- Advanced Regression
  - Elastic Net
  - GAMs
  - Quantile Regression
- Decision Trees
- Monotonic Gradient Boosting Machines (w/ XGBoost)
- Rule-based Models

2 Major Interpretability Techniques

These techniques are used to increase understanding of data and enhance understanding and trust in models. Ideally they are paired with an accurate modeling approach selected from Section 1, but most be can be used for any type of standard machine learning or statistical approach (i.e. most are model-agnostic.)

- Data Visualizations
  - 2-D Projections
  - Correlation Graphs
- Model Visualizations
  - Individual Conditional Expectation Plots
  - Partial Dependence Plots
  - Residual Plots
- Reason Codes

- Residual Plots
- Sensitivity Analysis
- Surrogate Models
  - Decision Tree
  - Local Interpretable Model-Agnostic Explanations
  - Rule-based (Anchors?)
- Tree Interpreter
- Variable Importance
  - Global Variable Importance
    - Gradient Boosting Machines
    - Neural Networks?
    - Random Forest
  - Local Variable Importance: Leave-One-Covariate-Out (LOCO)

3 Major Interpretability Applications

These are the three major classes of applications for the techniques in section 2.

- Debugging Models for *Accuracy*
- Deriving Insights from Models and Explanations for Models for Increased *Accountability*, *Fairness*, and *Transparency*
- Testing for Models *Stability* and *Trustworthiness*

4 Major Interpretability Themes

These are the major themes of discussion and writing pertaining to intepretable machine learning.

- Social Motivation for ML and ML Interpretability
- Commercial Motivation for ML and ML Interpretability
- Exact Models with Approximate Explanations
- The Multiplicity of Good Models, Model Locality, and Explanation Instability
- Risk from Local Error, Unwanted Bias, and Prediction Variance

- A Scale for Interpretability
  - High: Linear, Monotonic
  - Medium: Nonlinear, Monotonic
  - Low: Nonlinear, Non-monotonic
- Instability of Machine Learning Results
- Trust and Understanding
- Global Interpretability and Local Interpretability
- Model-Agnostic Interpretability and Model-Specific Interpretability
- Defining Interpretability
- Testing Interpretability

5 Matching Applications to Techniques

Application	Techniques
Debugging Models for *Accuracy*	• Model Visualizations Partial Dependence Plots Residual Plots
Deriving Insights and Explanations for *Accountability*, *Fairness*, and *Transparency*	• Model Visualizations Individual Conditional Expectation Plots Partial Dependence Plots • Reason Codes • Surrogate Models Decision Tree LIME Rule-based (Anchors?) • Tree Interpreter • Variable Importance Global Variable Importance LOCO
Testing for *Stability* and *Trustworthiness*	• Data Visualizations 2-D Projections Correlation Graphs • Surrogate Models: Decision Tree • Sensitivity Analysis • Testing Explanations

6 Matching Themes to Applications and Techniques

6.1 Social Motivation of ML and ML Interpretability

ML promises unbiased, data-driven decisions in touchy socio-economic scenarios. ML Intepretability helps gaurantee the promise, by assuring fairness, transparency, and trustworthiness of ML decisions.

6.2 Commerical Motivation of ML and ML Interpretability

ML promises more accurate predictions than linear models, but the lack of transparency in ML models limits their use in regulated industries and makes many practicioners and decision-makers suspicious of ML predictions. ML Intepretability can help with both the regulatory and trustworthiness concerns, opening the door for more accurate predictions and better financial margins.

6.3 The Multiplicity of Good Models, Model Locality, and Explanation Instability

Preliminary consideration. ML model predictions can be unstable and many good models exist at many local minima of error for a given model and data set. These are the primary, but not only reasons, explanations are approximate. This problem is solved emperically by proving stability and increasing trust, and maybe mathematically, by regularization and ensembles. This is an open problem.

6.4 Exact Models with Approximate Explanations

Preliminary consideration. Assuming approximate explanations have value, this idea *intellectually* validates all techniques.

6.5 Risk from Local Error, Unwanted Bias, and Variance

The accuracy, accountability, fairness, and transparency applications and associated techniques seek to rememedy these problems. These problems are remedied by greater understanding of model mechanisms.

6.6 Instability of Machine Learning Results

The sensitivity analysis and data visualization techniques seek to rememdy this problem. This problem is solved by emperically proving stability and increasing trust, and mathematically by regularization and ensembles.

6.7 A Scale for Interpretability

Preliminary consideration. Some models are too difficult to explain, some models are too appoximate to be useful. You should consider this fact before starting a project, because you may not be able to back-fit an explanation onto a very complex model.

6.8 Trust and Understanding

Sensitivity analysis and data visulaization are about increasing trust. Other techniques are more about increasing understanding.

6.9 Global Interpretability and Local Interpretability

Some techniques are global, some techniques are local. Both are important and have strengths and weaknesses.

6.10 Model-Agnostic Interpretability and Model-Specific Interpretability

Tangential, but important consideration. Some interpretability techniques work for all models, some only work for certain types of models.

6.11 Defining Interpretability

Preliminary consideration. This is an open problem.

6.12 Testing Interpretability

This is an open problem. This idea *emperically* validates all techniques. Also falls under *trust*.

- Human Assessment
- Reason Code Stability with Increased Prediction Accuracy
- Reason Code Stability Under Data Perturbation