WELLS FARGO

Machine Learning Model Validation

Part 2: Model Diagnostics and Validation

Aijun Zhang, Ph.D.

Corporate Model Risk, Wells Fargo

QU-ML Model Validation Workshop, Session 2, July 6, 2022.

Machine Learning Model Validation

Session 1 (last time)

Machine Learning Interpretability

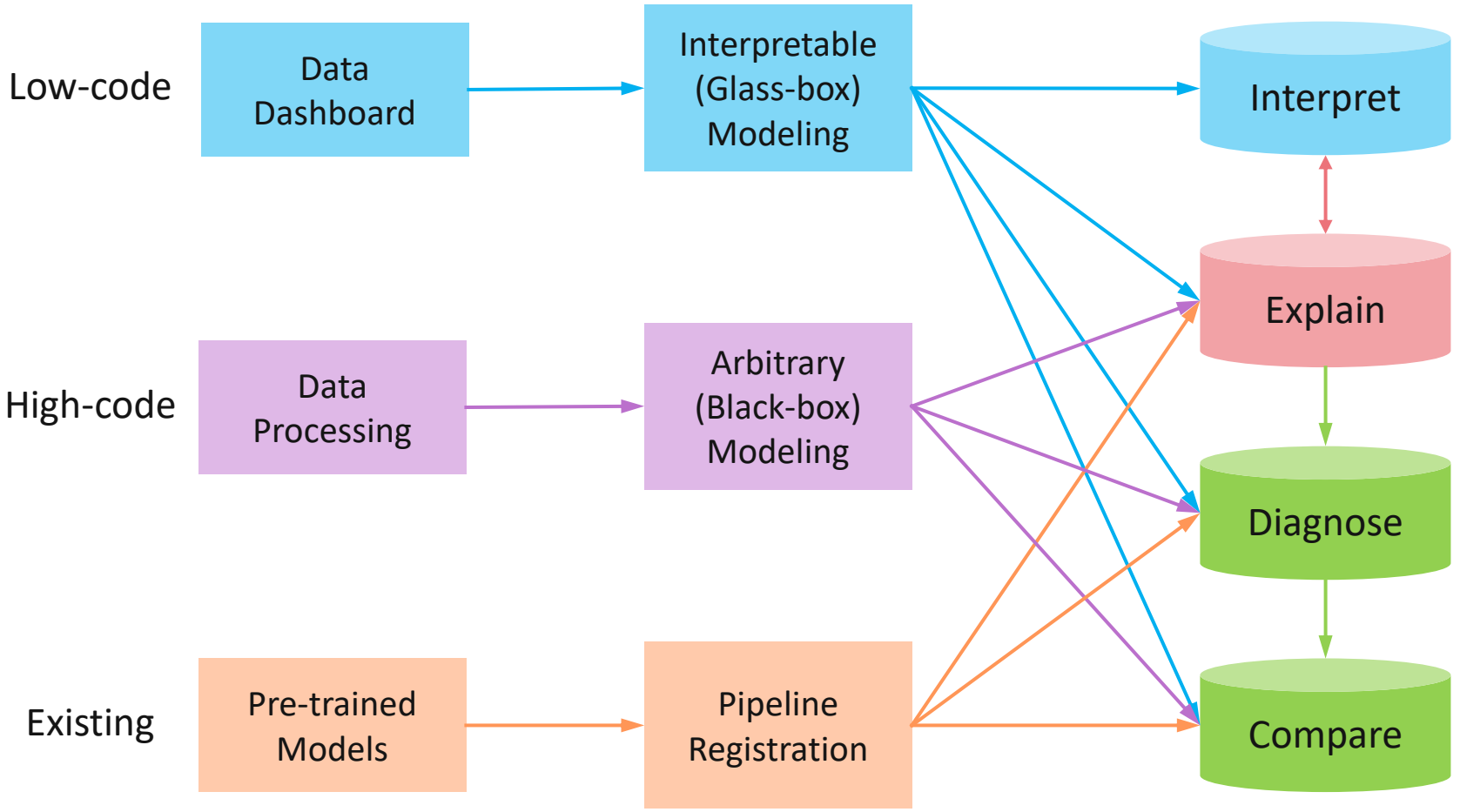
1. Interpretable ML and PiML Toolbox
2. Post-hoc Explainability Tools/Puzzles
3. Designing Interpretable ML Models
4. ReLU Deep Neural Networks
5. FANOVA Models: EBM and GAMI-Net

Session 2 (today)

Model Diagnostics and Validation

1. AI Model Risk and Trustworthiness
2. Accuracy, WeakSpot and Overfit
3. Reliability Testing
4. Robustness and Resilience Testing
5. Model Comparison

PiML Toolbox: Interpret/Explain/Diagnose/Compare

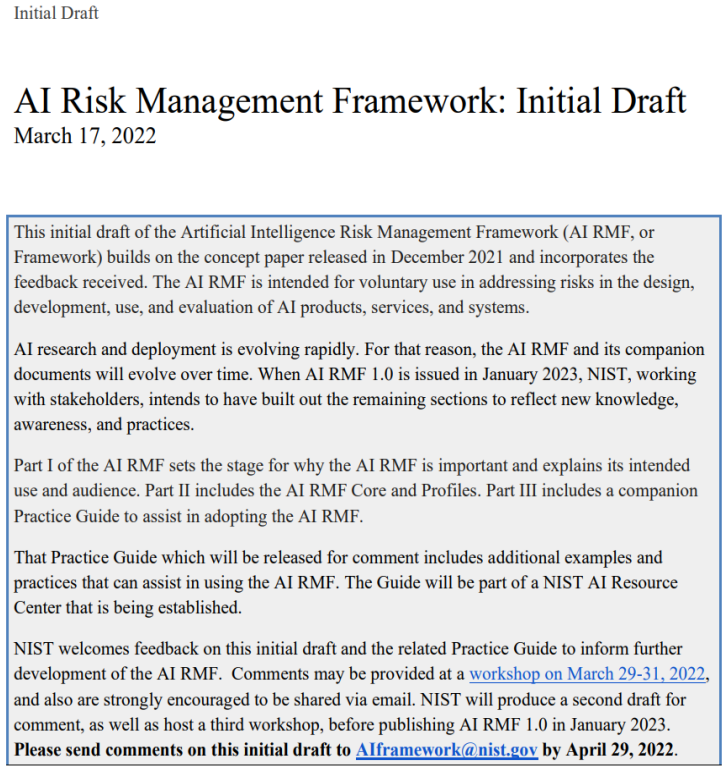


Demo Notebooks: https://github.com/SelfExplainML/PiML-Toolbox/tree/main/docs/Workshop

AI Model Risk and Trustworthiness

NIST's latest release (03/17/2022) of AI Risk Management Framework:

> "**Accuracy** indicates the degree to which ML model is correctly capturing a relationship exists within the data. Accuracy is examined via standard ML metrics as well as assessment of model **underfit** or **overfit**."
>
> "**Reliability** indicates whether a model consistently generates the same results, within the bounds of acceptable statistical error. [It] can be a factor in determining thresholds for acceptable risk."
>
> "**Robustness** is a measure of model sensitivity to variations in uncontrollable factors. It might range from sensitivity of a model's outputs to small changes in its inputs, but might also include error measurements on novel datasets."
>
> "**Resilience** or ML security [is to] withstand adversarial attacks, or more generally, unexpected changes in its environment or use."



Initial Draft

AI Risk Management Framework: Initial Draft
March 17, 2022

This initial draft of the Artificial Intelligence Risk Management Framework (AI RMF, or Framework) builds on the concept paper released in December 2021 and incorporates the feedback received. The AI RMF is intended for voluntary use in addressing risks in the design, development, use, and evaluation of AI products, services, and systems.

AI research and deployment is evolving rapidly. For that reason, the AI RMF and its companion documents will evolve over time. When AI RMF 1.0 is issued in January 2023, NIST, working with stakeholders, intends to have built out the remaining sections to reflect new knowledge, awareness, and practices.

Part I of the AI RMF sets the stage for why the AI RMF is important and explains its intended use and audience. Part II includes the AI RMF Core and Profiles. Part III includes a companion Practice Guide to assist in adopting the AI RMF.

That Practice Guide which will be released for comment includes additional examples and practices that can assist in using the AI RMF. The Guide will be part of a NIST AI Resource Center that is being established.

NIST welcomes feedback on this initial draft and the related Practice Guide to inform further development of the AI RMF. Comments may be provided at a workshop on March 29-31, 2022, and also are strongly encouraged to be shared via email. NIST will produce a second draft for comment, as well as host a third workshop, before publishing AI RMF 1.0 in January 2023. **Please send comments on this initial draft to AIframework@nist.gov by April 29, 2022.**

- These ML aspects can be measured by **outcome testing** using PiML Toolbox.

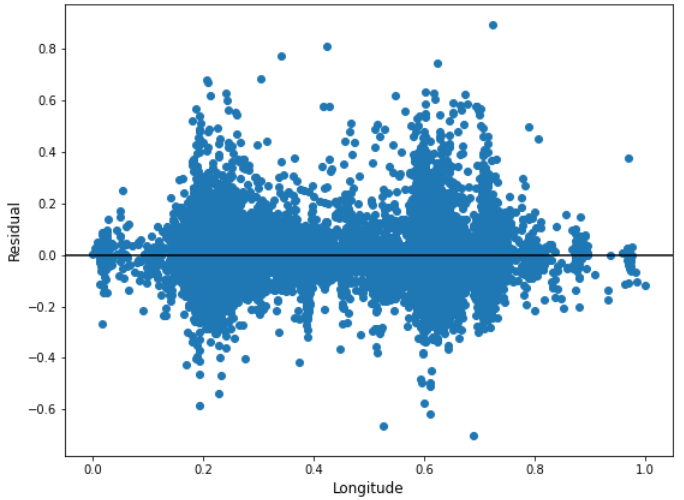
AI Risk Management Framework (Initial Draft). Accessible from nist.gov

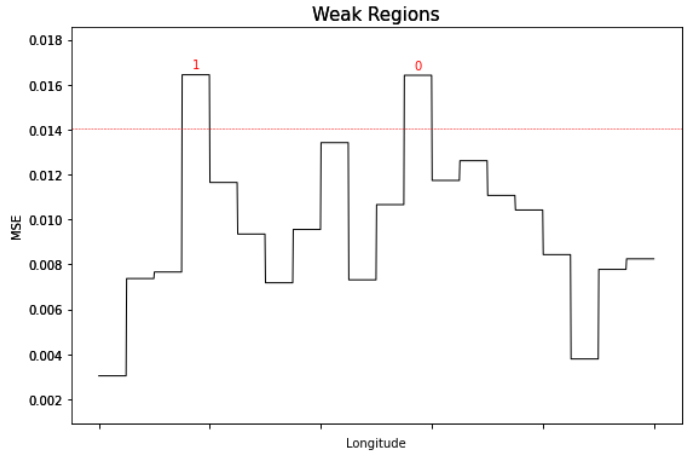
Accuracy, WeakSpot and Overfit

Error Analysis by Slicing Techniques

Accuracy, Residuals and WeakSpot

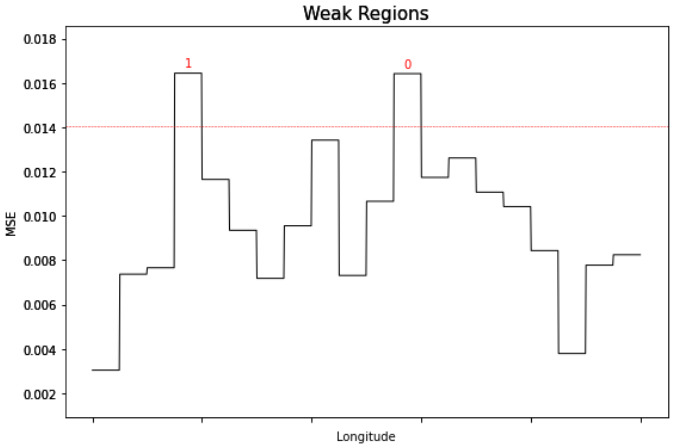
- ML model performance is often measured by **accuracy,** as examined via standard ML metrics (e.g. MSE, MAE, R2, ACC, AUC, F1-score, Precision and Recall).
- However, model assessment by single-valued metrics is insufficient. More granular diagnostics and alternative metrics are needed.
- To check **model underfitting**, perform error analysis based on residuals
  - **Residual plot** marginally for each feature of interest;
  - **WeakSpot** to identify weak regions with high residuals on either training or testing data.
- **PiML toolbox** employs several slicing techniques for WeakSpot.

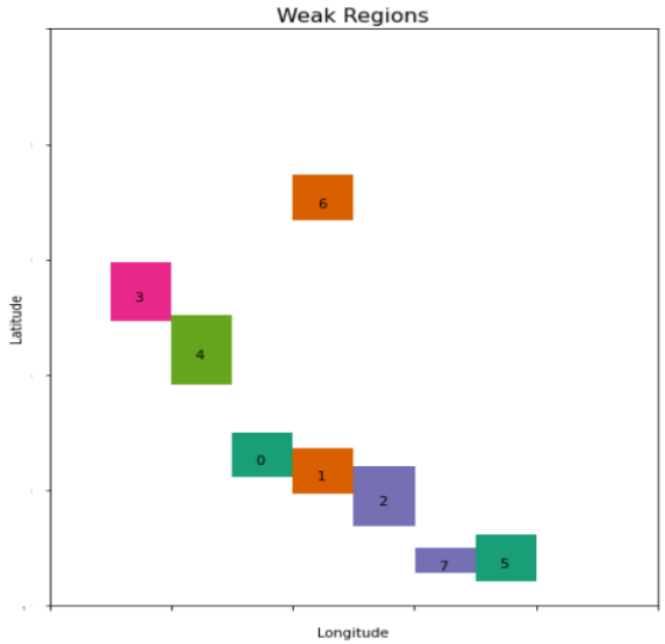




Error Analysis by Slicing Techniques

1. **Specify an appropriate metric** based on individual prediction residuals: e.g., MSE for regression, ACC for classification, train-test performance gap, confidence bandwidth, etc.
2. Specify 1 or 2 slicing features of interest;
3. Evaluate the metric for each sample in the target data (training or testing) as pseudo responses;
4. Segment the target data along the slicing features, by
   a) [Unsupervised] Histogram slicing with equal-space binning, or
   b) [Supervised] fitting a decision tree or tree-ensemble to generate the sub-regions;
5. **Identify the sub-regions** with average metric exceeding the pre-specified threshold, subject to minimum sample condition.





Overfitting: Benign or Malignant?

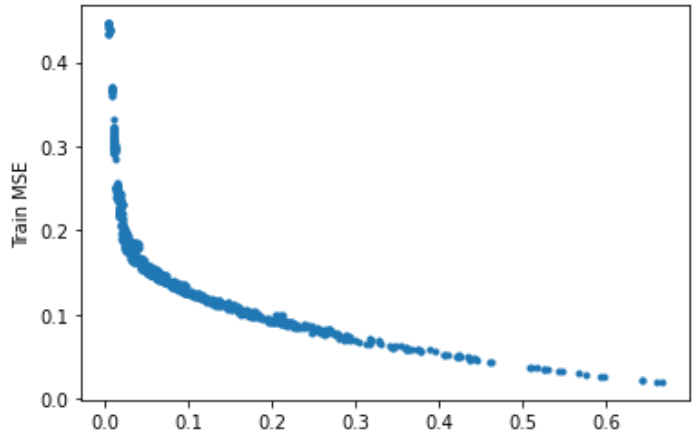
- **Overfitting** is measured by the gap between training and testing errors. The bigger the gap, the more overfitting.
- Consider SURE formula (Stein's Unbiased Risk Estimator) under multivariate normal assumption:

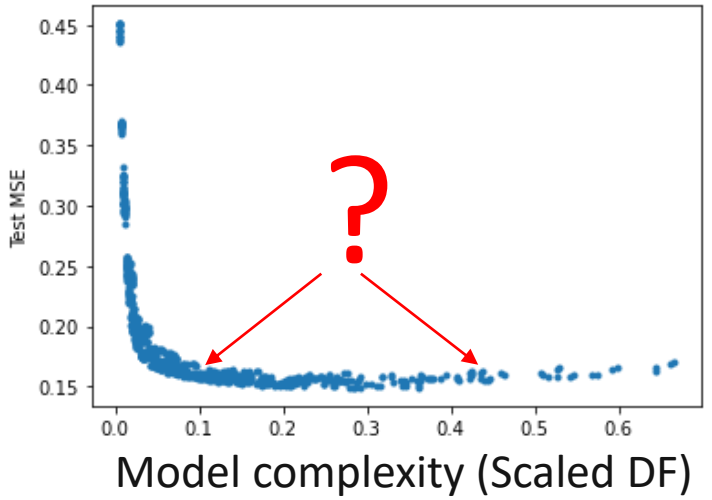
$$\widehat{\mathrm{Err}} = \mathrm{err} + \frac{2\sigma^2}{N}\sum_{i=1}^{N}\frac{\partial \hat{f}_i}{\partial y_i}$$

Testing error — Training error — Degree of freedom

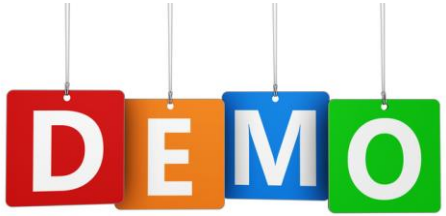
The degree of freedom (DF) can be used to evaluate overfitting. The larger the DF, the more flexible (a.k.a. higher variance) and overfitting the model is.

- **Overparametrized models** tend to be overfitting, which might be benign for a fixed testing data, but **harmful** when distribution shift happens:
  - Data cleaning during model development doesn't reflect data in real use.
  - This links to the next topic of robustness testing under covariate shift.

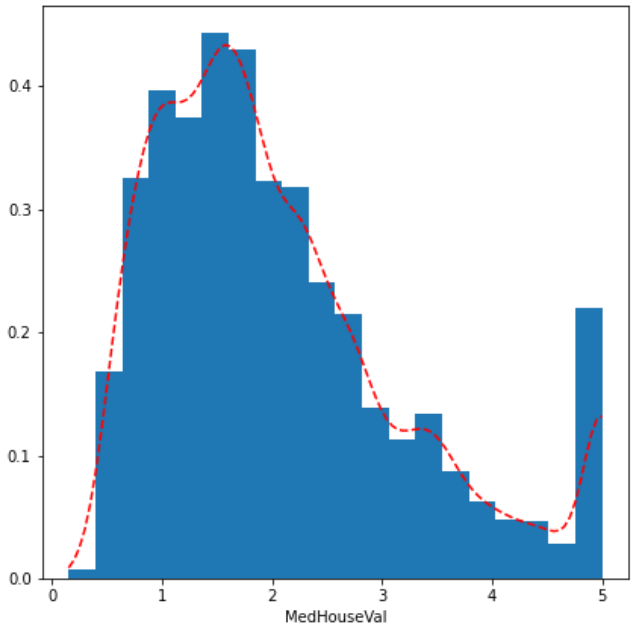


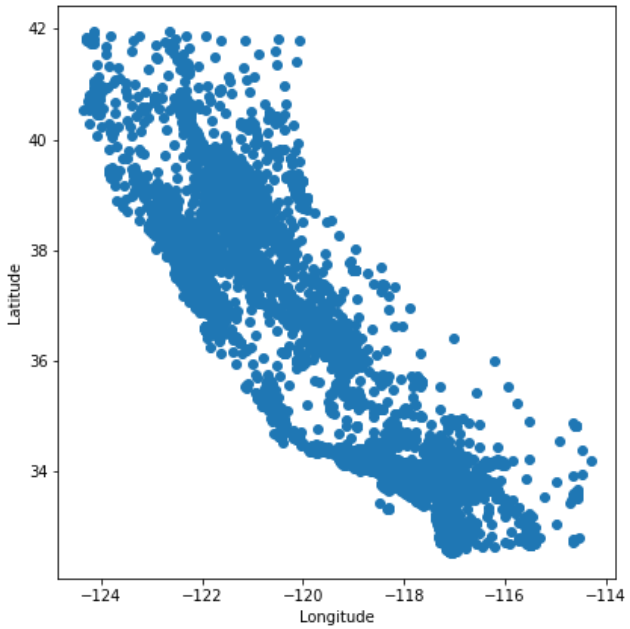


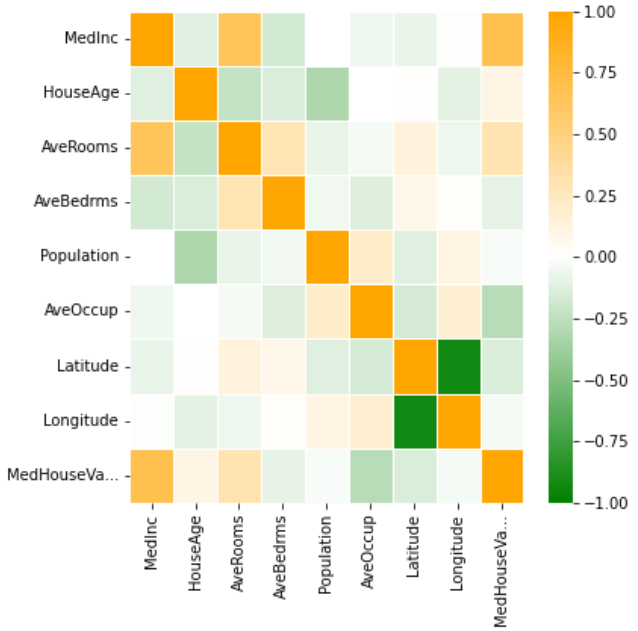
PiML Demo: CaliforniaHousing Dataset



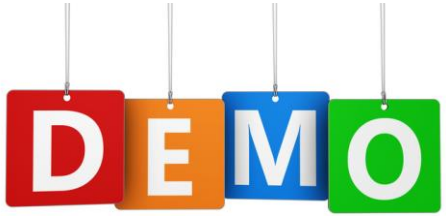
- Data from 1990 Census for California, for housing of geographically compact areas (Total of 20,640 observations)
- Available by `sklearn.datasets.fetch_california_housing`, pre-built in PiML.
- Response: Median house price per area, analyzed as log(median price)
- 8 features: median income, house age, rooms, bedrooms, population, number of households, latitude and longitude.

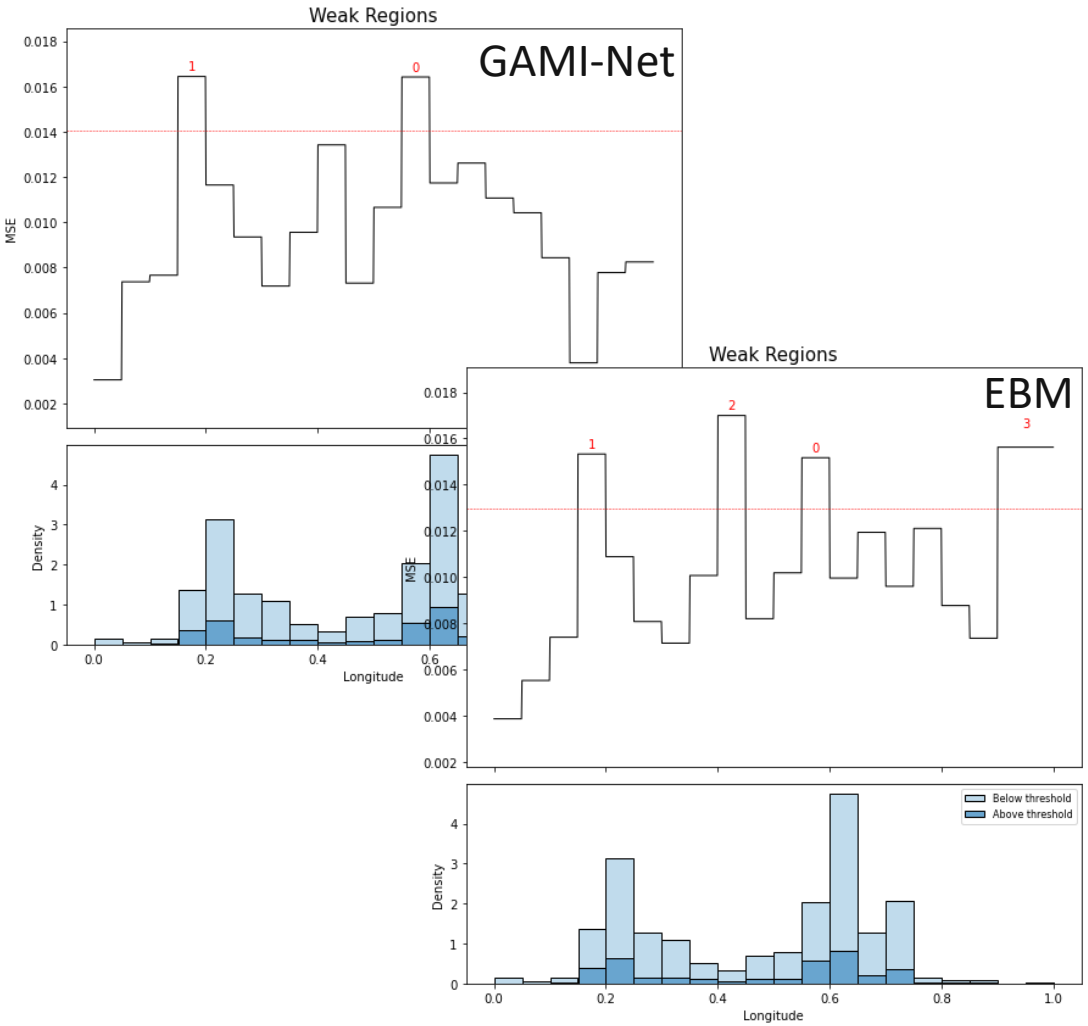


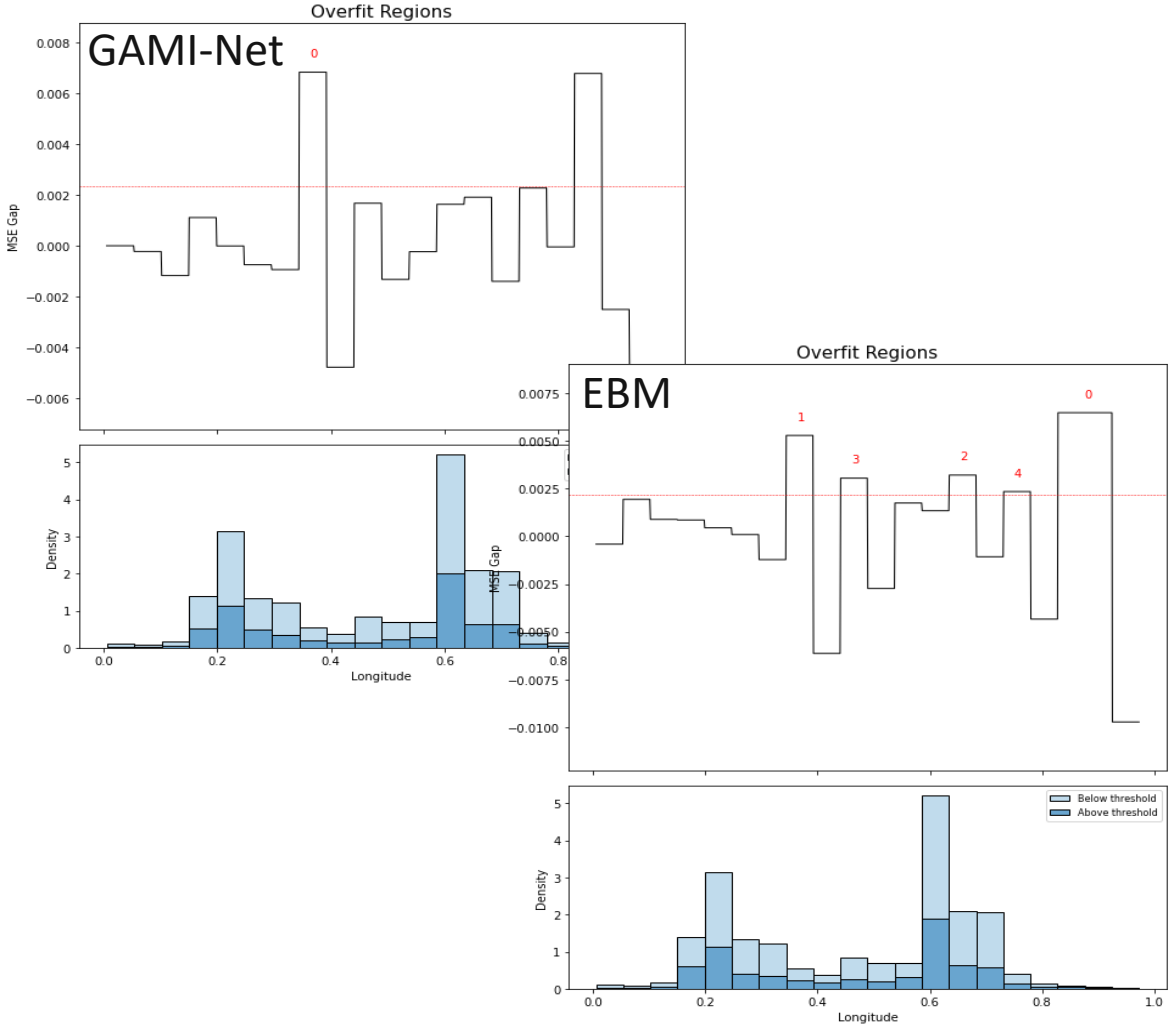




PiML Demo: WeakSpot and Overfit







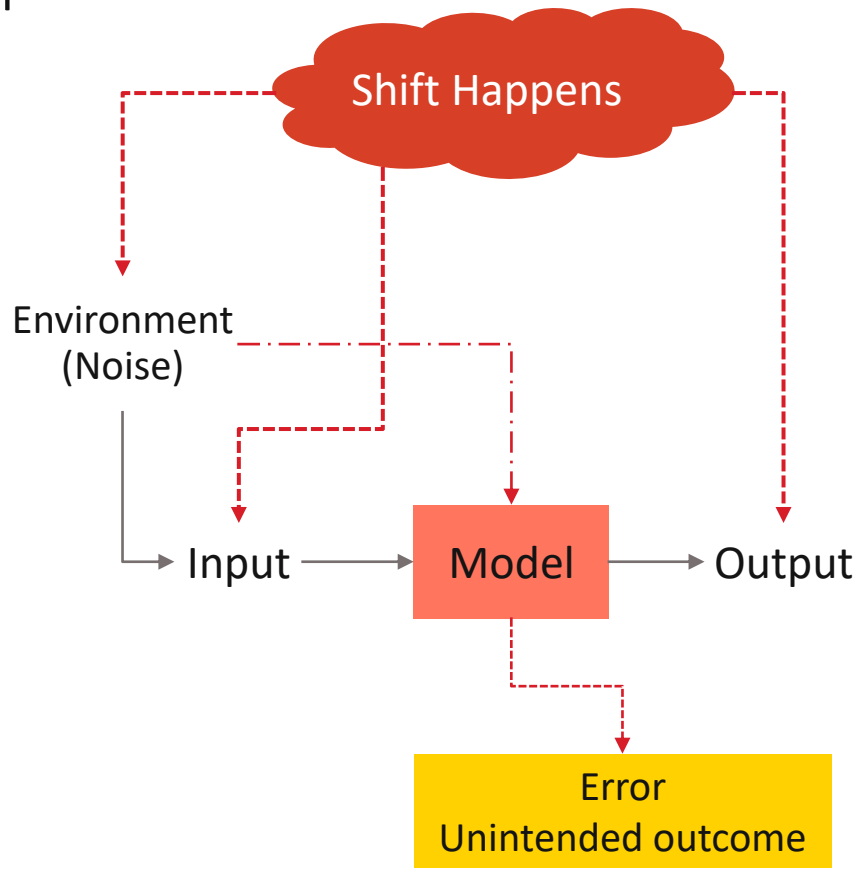
PiML Demo: WeakSpot and Overfit analysis for CaliforniaHousing data fit by GAMI-Net and EBM.

Robustness and Resilience Testing

subject to distribution shift

Robustness and Resilience Testing

- Train-test data split for model development often gives over-optimism of model performance, since model in production will be exposed to data distribution shift.
- **Robustness test**: evaluate the performance degradation under covariate noise perturbation:
  - Perturb testing data covariates with small random noise;
  - Assess model performance of perturbed testing data.
  - Overfitting models often perform poorly in changing environments.
- **Resilience test**: evaluate the performance degradation under worst-case subsampling:
  - Investigate worst-performing samples, conditional on immutable features.
  - Rank covariates using distribution shift measure such as Population Stability Index (PSI);
  - Identify sensitivity and vulnerability due to covariate shift.

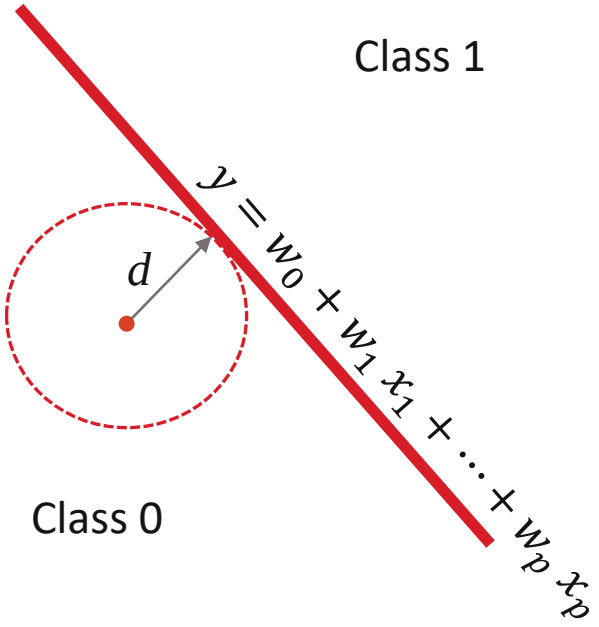


Robustness against Noise Perturbation

- For simplicity, consider the binary classifier with linear decision boundary.
- Distance to decision boundary:

$$d = \frac{|w_0 + w_1 x_1 + \cdots + w_p x_p|}{\|\mathbf{w}\|}$$

  - Perturbation of size d would flip the model output $\rightarrow$ wrong classification
  - The denominator suggests that minimizing $\|\mathbf{w}\|$ is enlarging distance d, thus improving model robustness.
- In general, regularized (i.e., simpler) models are more robust and less performance degradation when inputs are corrupted.
- To the contrary, overparametrized models (with benign overfitting on a fixed testing data) are less robust and harmful when shift happens.



Measuring Distribution Shift

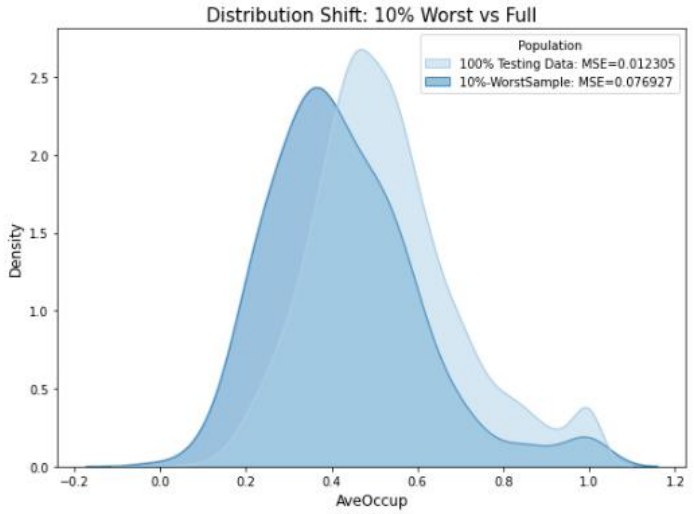
- **Population Stability Index:**

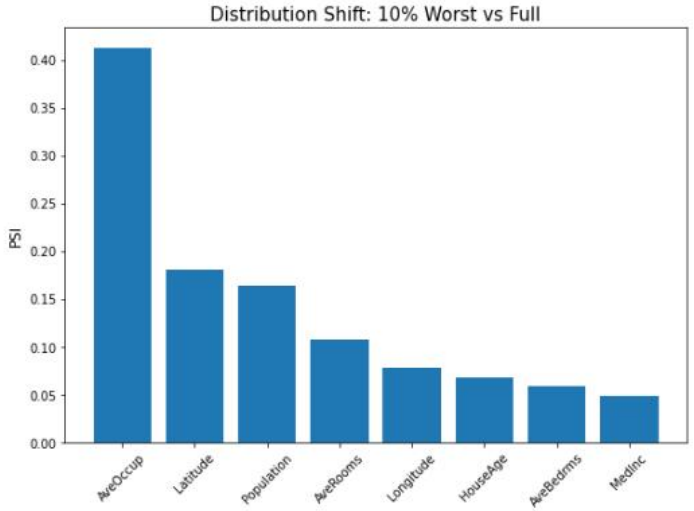
$$PSI = \sum_{i=1}^{B} (\text{Target}_i\% - \text{Base}_i\%) \ln\left(\frac{\text{Target}_i\%}{\text{Base}_i\%}\right)$$

  based on the proportions of samples in each bucket of the target vs. base population. Rule of thumb:
  - PSI < 0.1: no significant distribution change
  - PSI < 0.2: moderate distribution change
  - PSI >= 0.2: significant distribution change

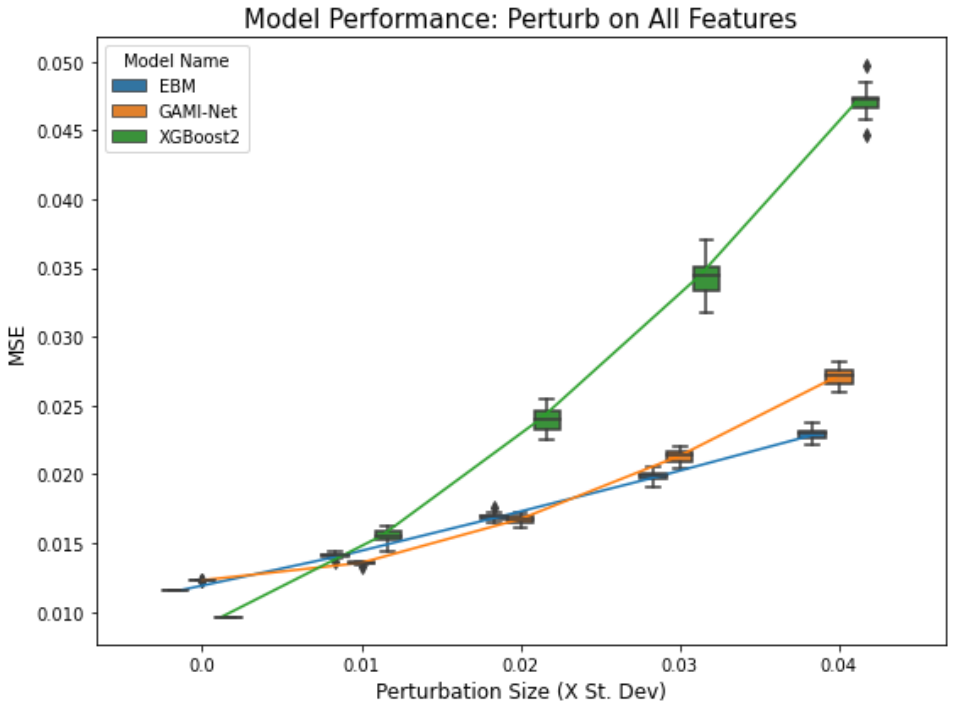
- Other two-sample test: KL divergence, Kolmogorov-Smirnov (KS) and Cramer-von Mises (CM) statistics based on empirical distributions.

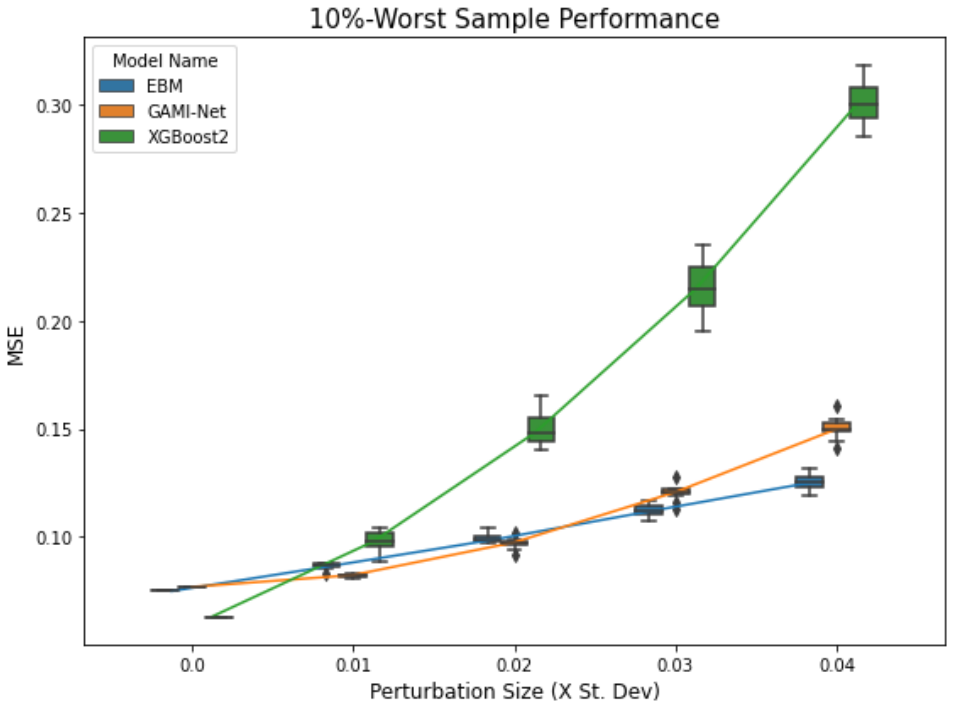
- In resilience testing, PSI measures the distribution shift one-feature-at-a-time. One may further use WeakSpot to perform drill-down analysis on sensitive features.





PiML Demo: Robustness Testing

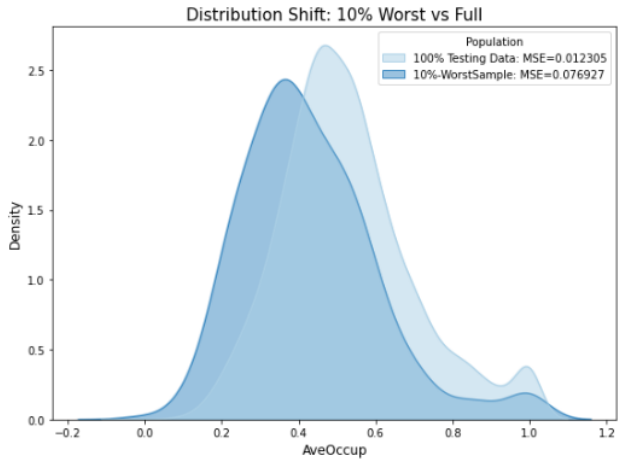


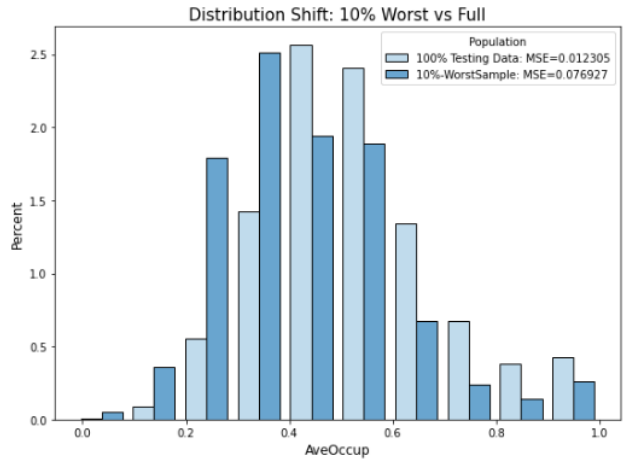


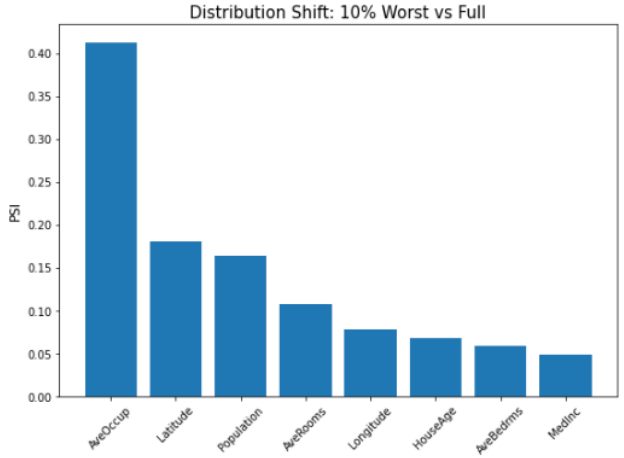
PiML Demo: Robustness Testing for CaliforniaHousing data fit by GAMI-Net, EBM vs XGBoost2.

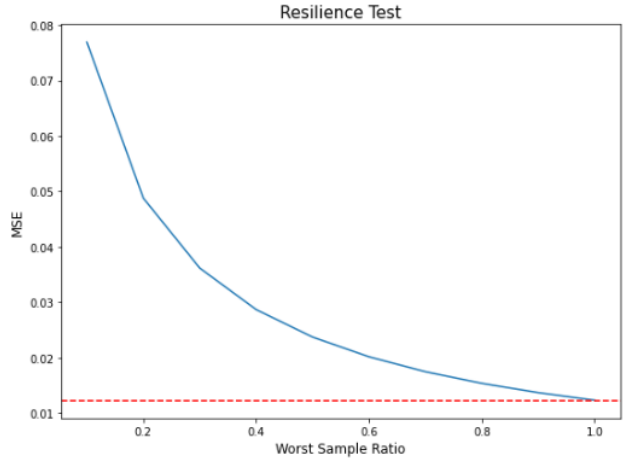
→ See more robustness testing later in model comparison

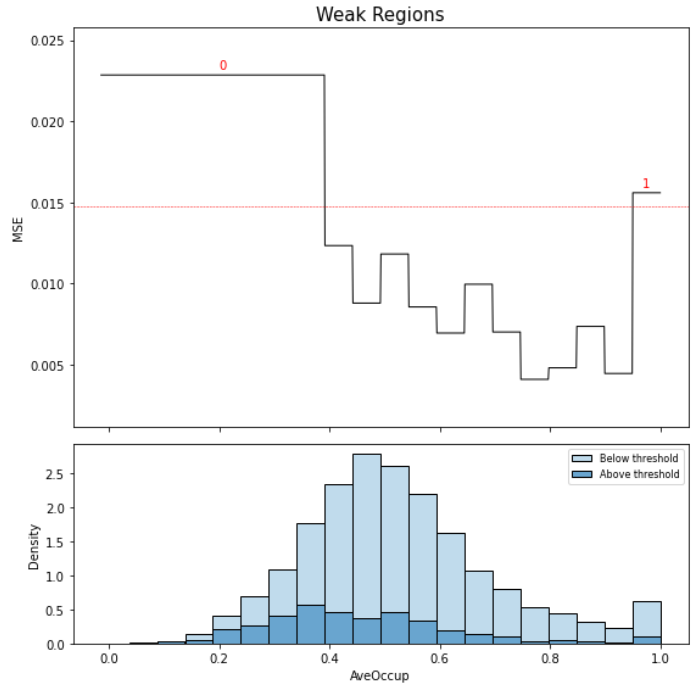
PiML Demo: Resilience Testing











PiML Demo: Resilience Test and WeakSpot for CaliforniaHousing data fit by GAMI-Net

Reliability Testing

via Calibration and Conformal Prediction

Reliability Testing

- Prediction reliability assessment is important to understand where the model produces less reliable prediction:

  Wider prediction interval $\rightarrow$ Less reliable prediction

- Quantification of prediction reliability can be done through **Split Conformal Prediction** under the exchangeability assumption:

Given a pre-trained model $\hat{f}(\boldsymbol{x})$, a hold-out calibration data $\mathcal{X}_{\text{calib}}$, a pre-defined conformal score $S(\boldsymbol{x}, y, \hat{f})$ and the error rate α (say 0.1)

1. Calculate the score $S_i = S(\boldsymbol{x}, y, \hat{f})$ for each sample in $\mathcal{X}_{\text{calib}}$;
2. Compute the calibrated score quantile

$$\hat{q} = \text{Quantile}\left(\{S_1, \dots, S_n\}; \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right);$$

3. Construct the prediction set for the test sample $\boldsymbol{x}_{\text{test}}$ by

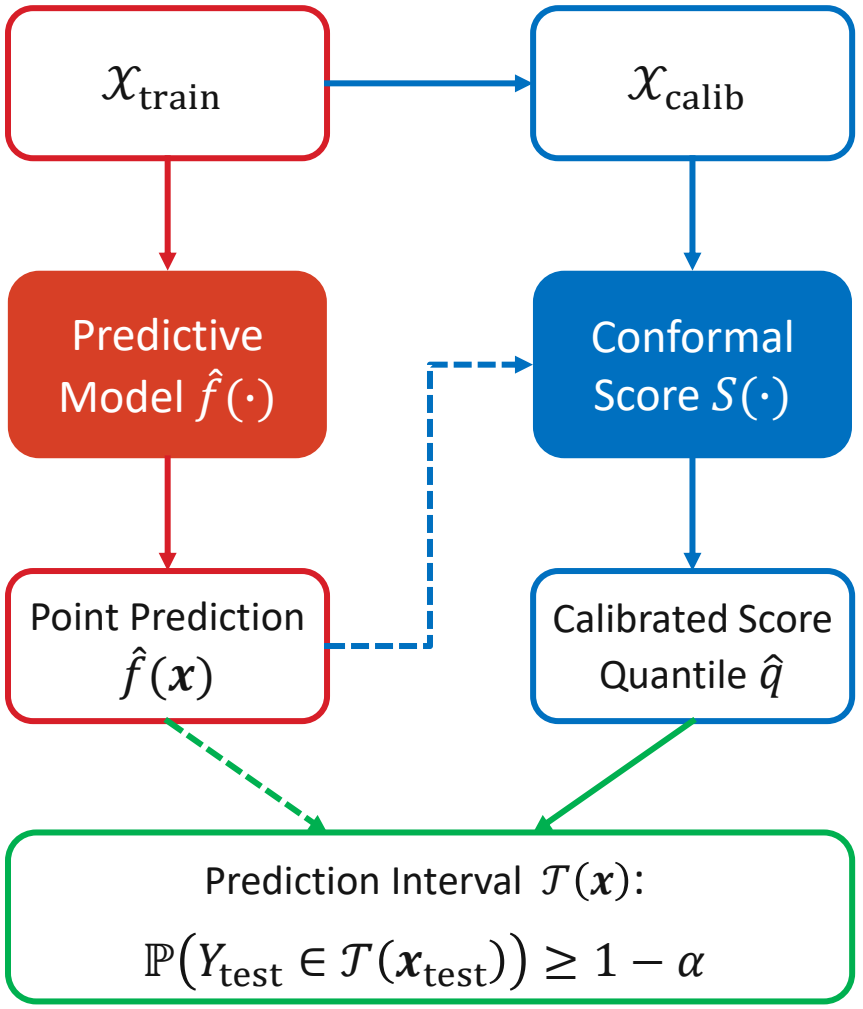
$$\mathcal{T}(\boldsymbol{x}_{\text{test}}) = \left\{y: S\left(\boldsymbol{x}_{\text{test}}, y, \hat{f}(\boldsymbol{x}_{\text{test}})\right) \leq \hat{q}\right\}.$$

Under the exchangeability condition of conformal scores, we have that

$$1 - \alpha \leq \mathbb{P}\left(Y_{\text{test}} \in \mathcal{T}(\boldsymbol{x}_{\text{test}})\right) \leq 1 - \alpha + \frac{1}{n+1}.$$

This provides the prediction bounds with α-level acceptable error.

- For binary classifiers, reliability diagrams are considered using probability calibration w.r.t. observed success frequencies.



Conformalized Residual Quantile Regression

Directly evaluate reliability of a pre-trained regression model $\hat{f}(\boldsymbol{x})$:

1. Obtain residuals $y_i - \hat{f}(\boldsymbol{x}_i)$ for each $i \in \mathcal{X}_{\text{train}}$ or $\mathcal{X}_{\text{split}}$, fit a quantile regressor (e.g. LightGBM with quantile loss) for residuals $\left[\hat{g}_{\alpha/2}(\boldsymbol{x}),\ \hat{g}_{1-\alpha/2}(\boldsymbol{x})\right]$;

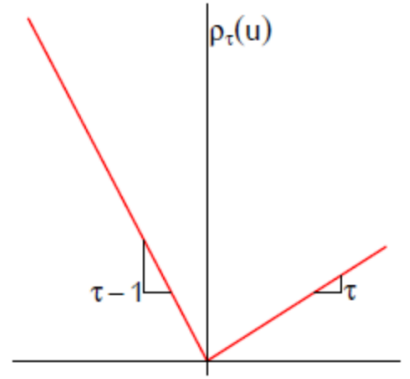
2. Define score $S\left(\boldsymbol{x}, y, \hat{f}\right) = \max\left\{\hat{g}_{\alpha/2}(\boldsymbol{x}) - y + \hat{f}(\boldsymbol{x}),\ y - \hat{f}(\boldsymbol{x}) - \hat{g}_{1-\alpha/2}(\boldsymbol{x})\right\}$

3. Calculate $\hat{q} = \text{Quantile}\left(\{S_1, \ldots, S_n\}; \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right)$, using $S\left(\boldsymbol{x}, y, \hat{f}\right)$ on $\mathcal{X}_{\text{calib}}$

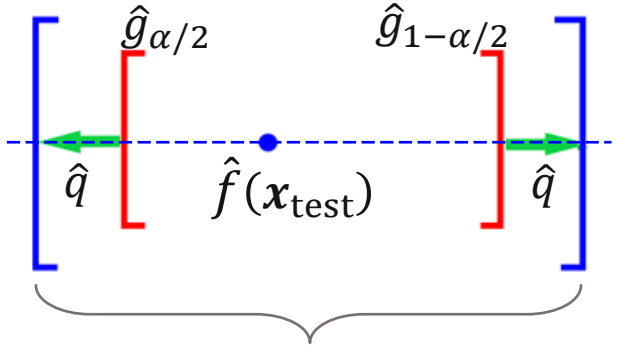
4. Construct the prediction interval for the test sample $\boldsymbol{x}_{\text{test}}$ by

$$\mathcal{T}(\boldsymbol{x}_{\text{test}}) = \left[\hat{f}(\boldsymbol{x}_{\text{test}}) + \hat{g}_{\alpha/2}(\boldsymbol{x}_{\text{test}}) - \hat{q},\ \ \hat{f}(\boldsymbol{x}_{\text{test}}) + \hat{g}_{1-\alpha/2}(\boldsymbol{x}_{\text{test}}) + \hat{q}\right].$$

Interpretation: the final prediction interval is composed of three terms: original prediction, estimated residual quantiles, and calibrated adjustment.

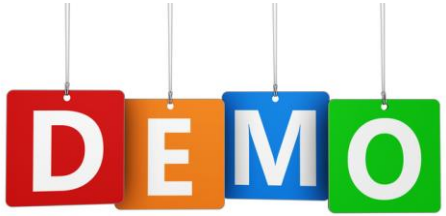


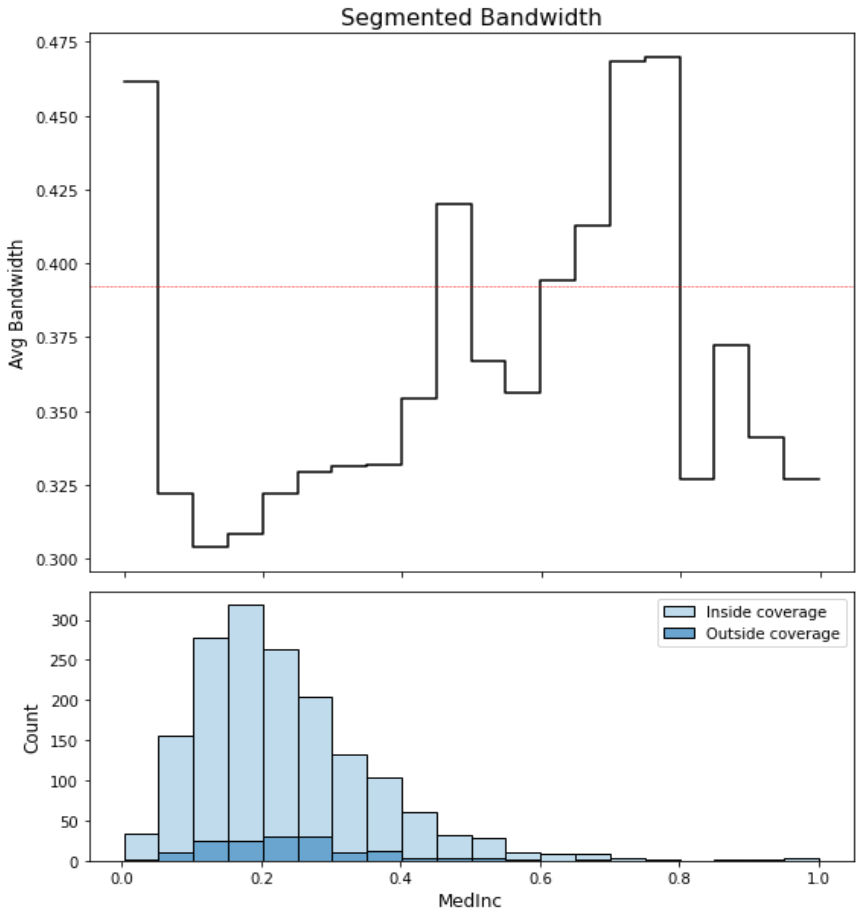
Quantile loss

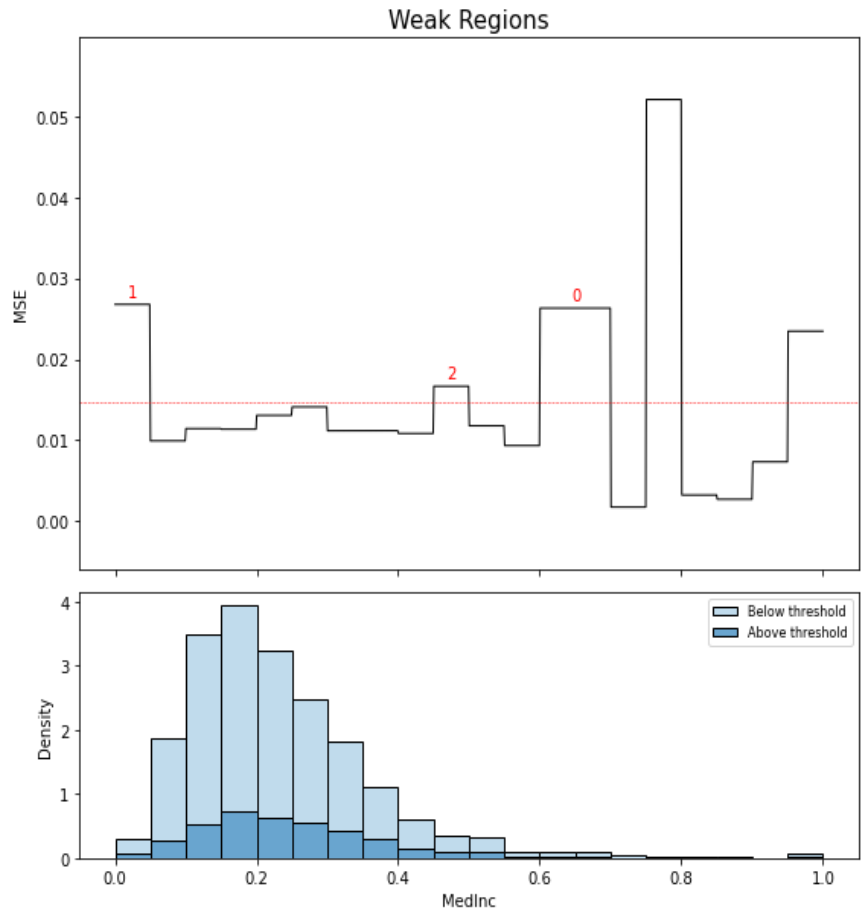


Bandwidth

PiML Demo: Reliability Testing







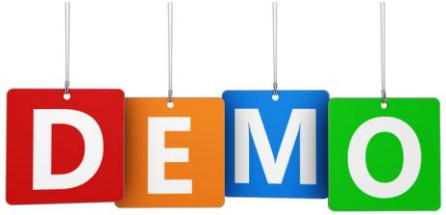
Note that quantile regression makes the interval bandwidth adaptive to heteroscedastic residuals.

PiML Demo: Reliability Testing for CaliforniaHousing data fit by GAMI-Net.

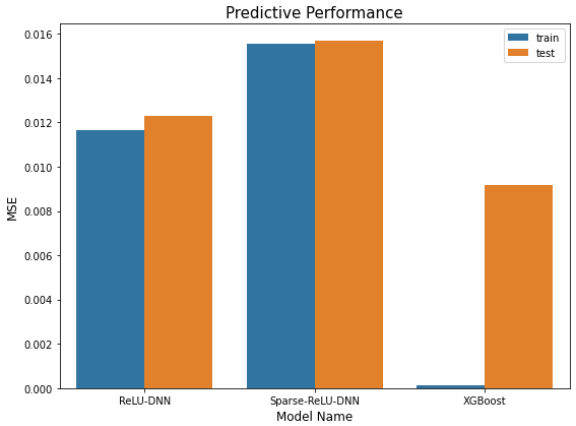
Model Comparison and Benchmarking

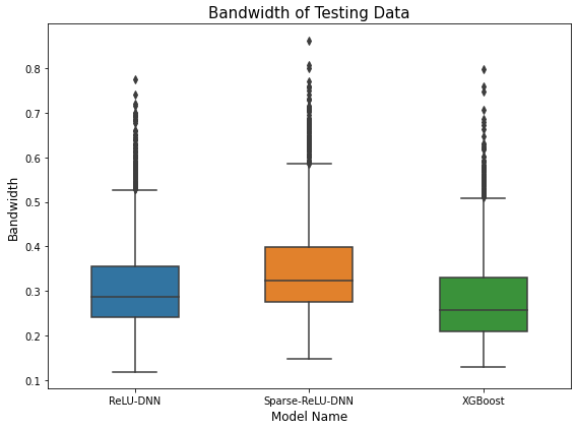
Accuracy, Reliability, Robustness and Resilience

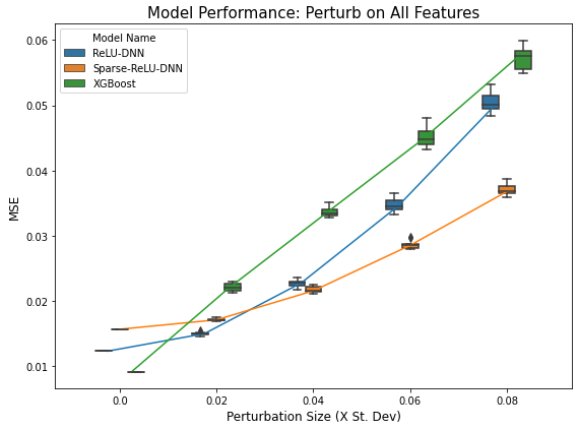
Model Comparison and Benchmarking

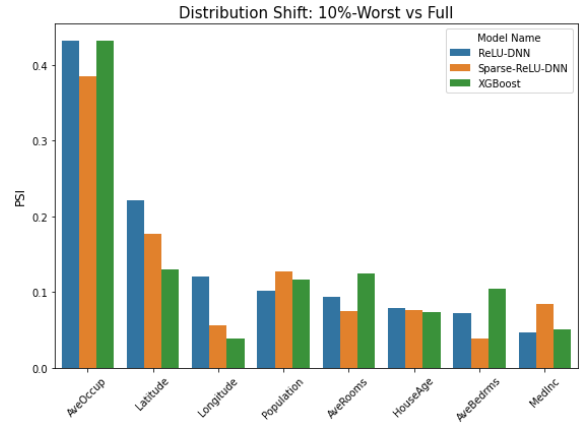


- Previous example of California Housing data fit by GAMI-Net and EBM shows that the simpler FANOVA-interpretable models are more robust and resilient than complex models.
- Similar comparative results are observed in many other cases, e.g., Sparse ReLU-DNN vs. Dense ReLU-DNN vs. XGBoost (depth-7) for CaliforniaHousing data modeling, or other models for other datasets.









PiML Demo: Model comparison for Sparse and Dense ReLU DNNs vs. XGBoost (CaliforniaHousing data)

WELLS FARGO

Thank you

Aijun Zhang, Ph.D.
Senior Lead, Advanced Technologies for Modeling (AToM), CMoR

Agus Sudjianto, Ph.D.
EVP, Head of Corporate Model Risk (CMoR)