WELLS FARGO

Model Diagnostics: WeakSpot, UQ, and Robustness

Aijun Zhang, PhD
Corporate Model Risk, Wells Fargo

Machine Learning Model Validation Course, June 21-23 | Risk.net

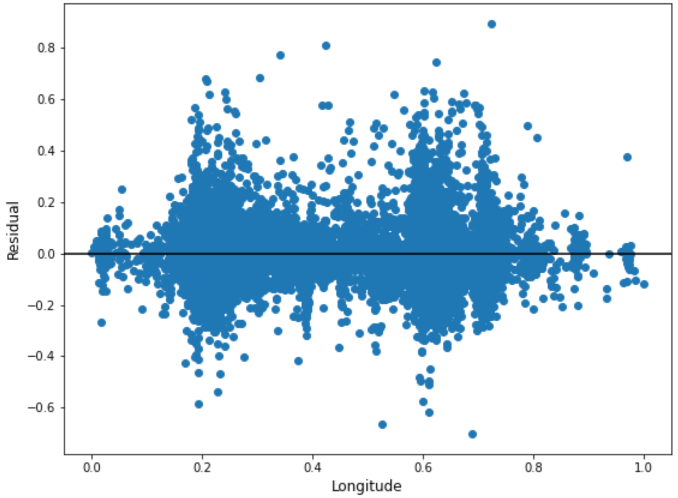
Accuracy, Residuals and WeakSpot

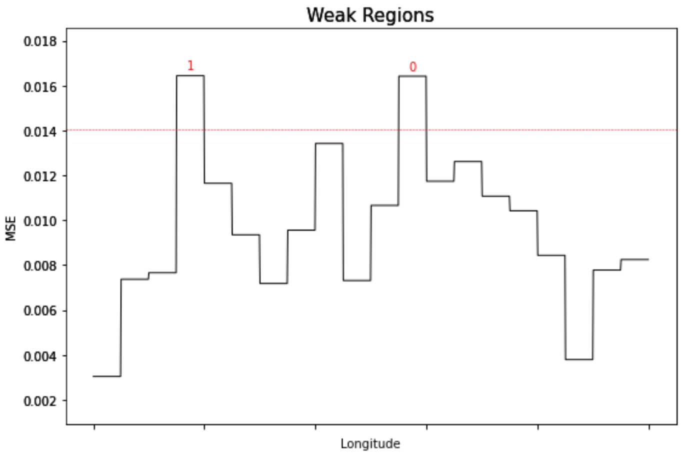
- ML model performance is often measured by **accuracy,** as examined via standard ML metrics (e.g. MSE, MAE, R2, ACC, AUC, F1-score, Precision and Recall).

- 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.

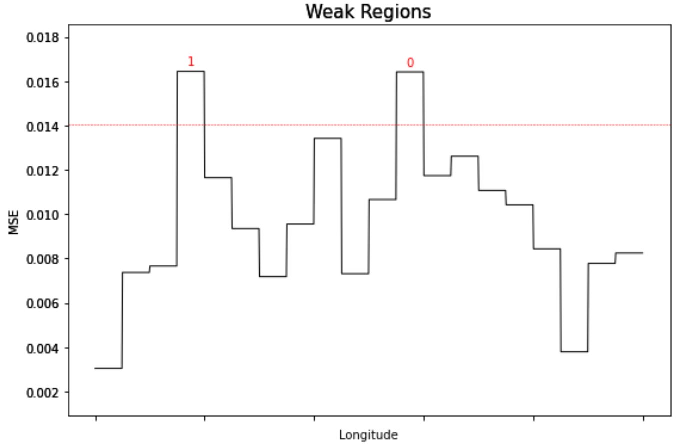
- **PiML toolbox** employs several slicing techniques for WeakSpot.

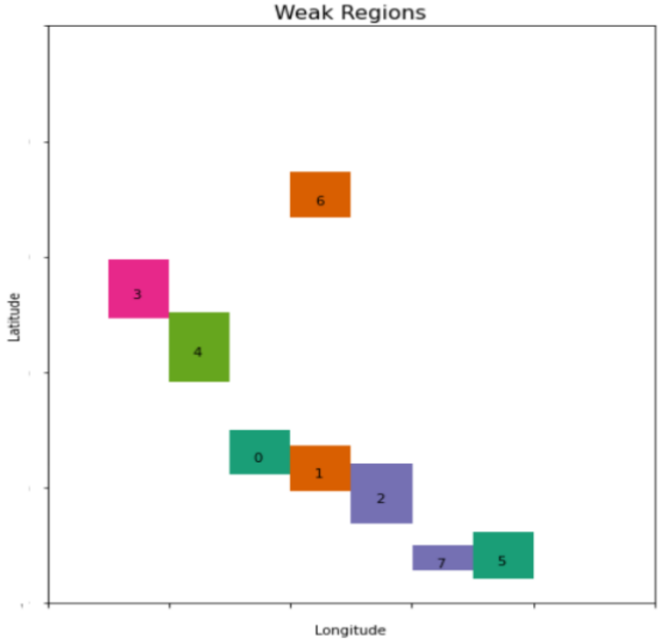




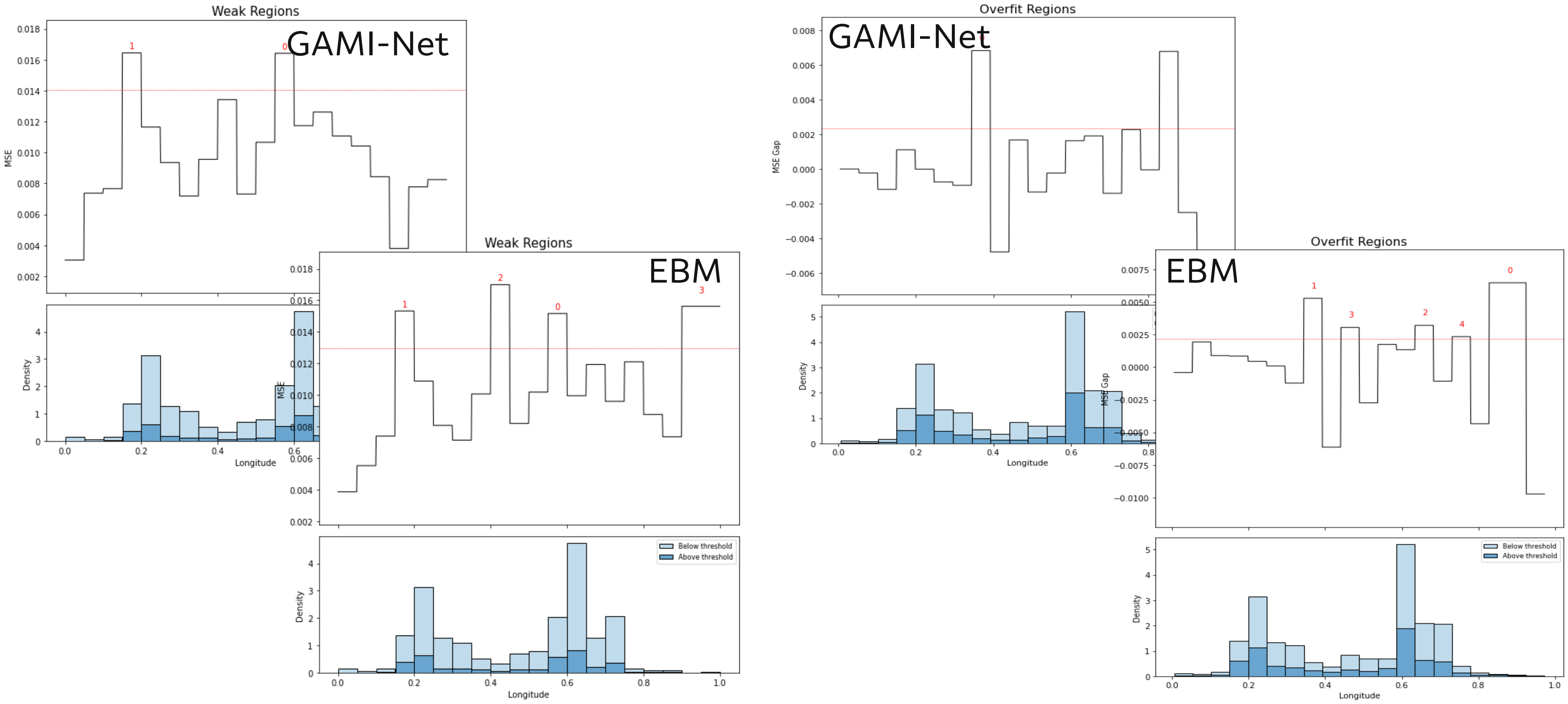
Error Analysis by Slicing Techniques

1. **Specify an appropriate metric** based on individual prediction residuals: e.g., MSE for regression, ACC for classification, train-test performance gap (for checking overfit), uncertainty bandwidth, …
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.





PiML Example: WeakSpot and Overfit



Example: WeakSpot and Overfit analysis for CaliforniaHousing data fit by GAMI-Net and EBM

Uncertainty Quantification

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

  Wider prediction interval $\rightarrow$ Less reliable prediction

- Quantification of prediction uncertainty 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

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

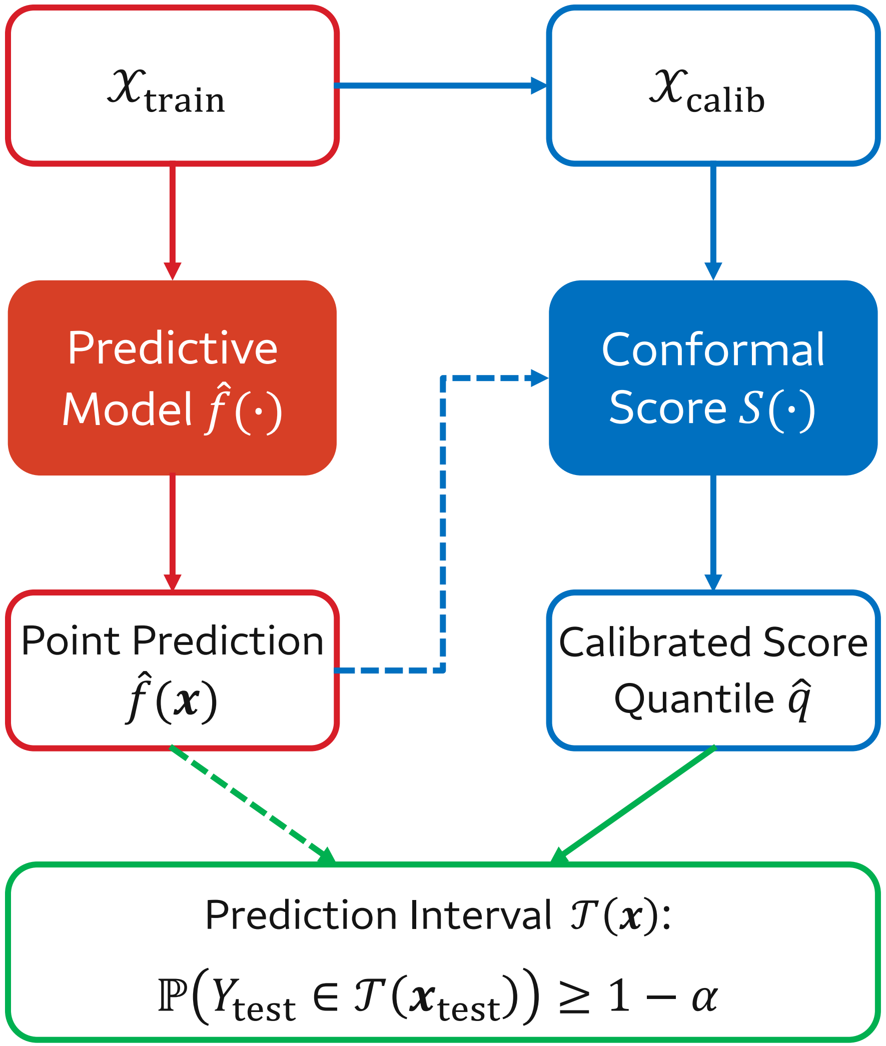
3. Construct the prediction set for the test sample $\boldsymbol{x}_{\text{test}}$ by

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

Under the exchangeability condition of conformal scores, we have that

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

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



Conformalized Residual Quantile Regression

Directly evaluate prediction uncertainty 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 $\left[\hat{g}_{\alpha/2}(\boldsymbol{x}),\ \hat{g}_{1-\alpha/2}(\boldsymbol{x})\right]$;

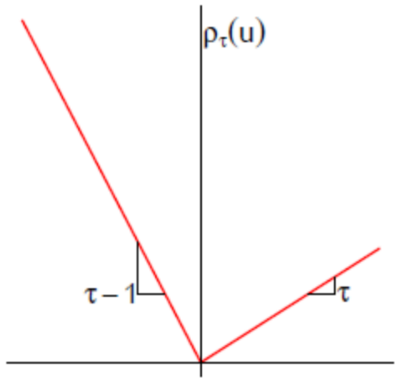
2. Define score $S\left(\boldsymbol{x}, y, \hat{f}\right) = \max\left\{\hat{g}_{\alpha/2}(\boldsymbol{x}) - y + \hat{f}(\boldsymbol{x}),\ y - \hat{f}(\boldsymbol{x}) - \hat{g}_{1-\alpha/2}(\boldsymbol{x})\right\}$

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

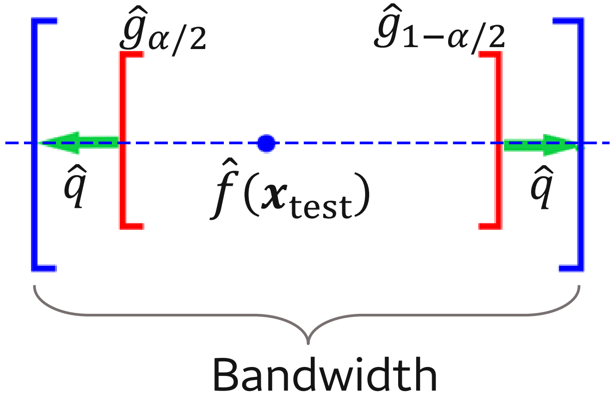
4. Construct the prediction interval for the test sample $\boldsymbol{x}_{\text{test}}$ by

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

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

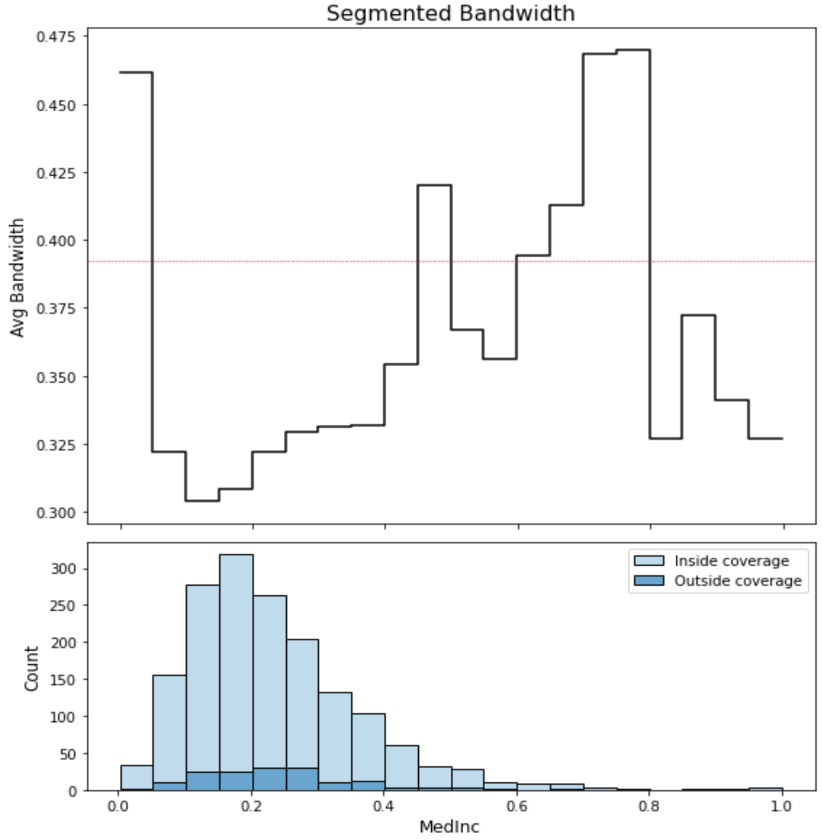


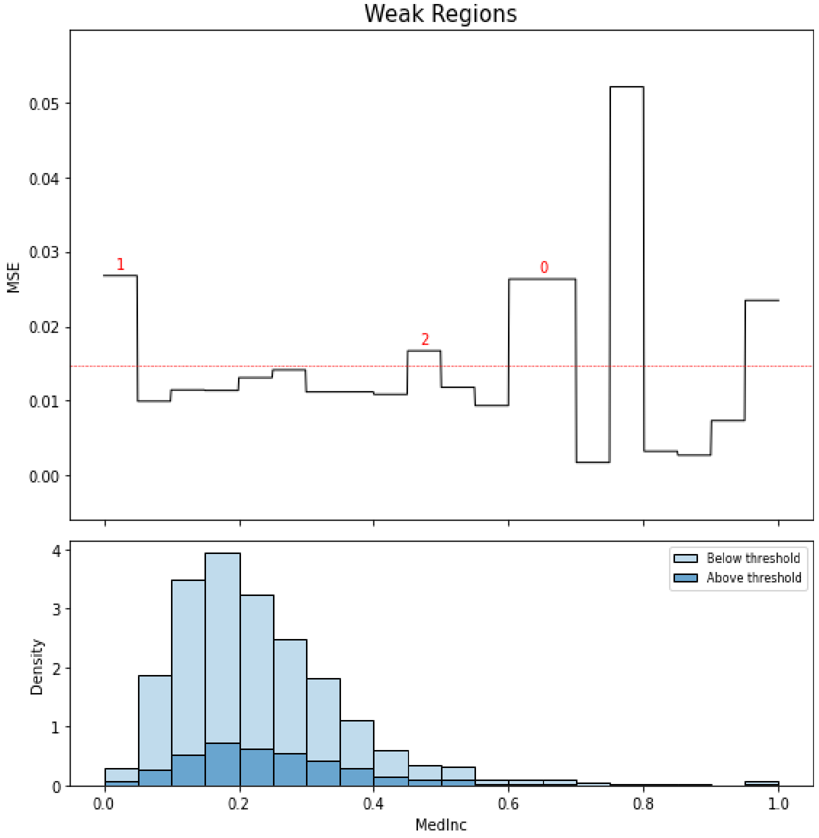
Quantile loss



PiML Example: Uncertainty Quantification

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

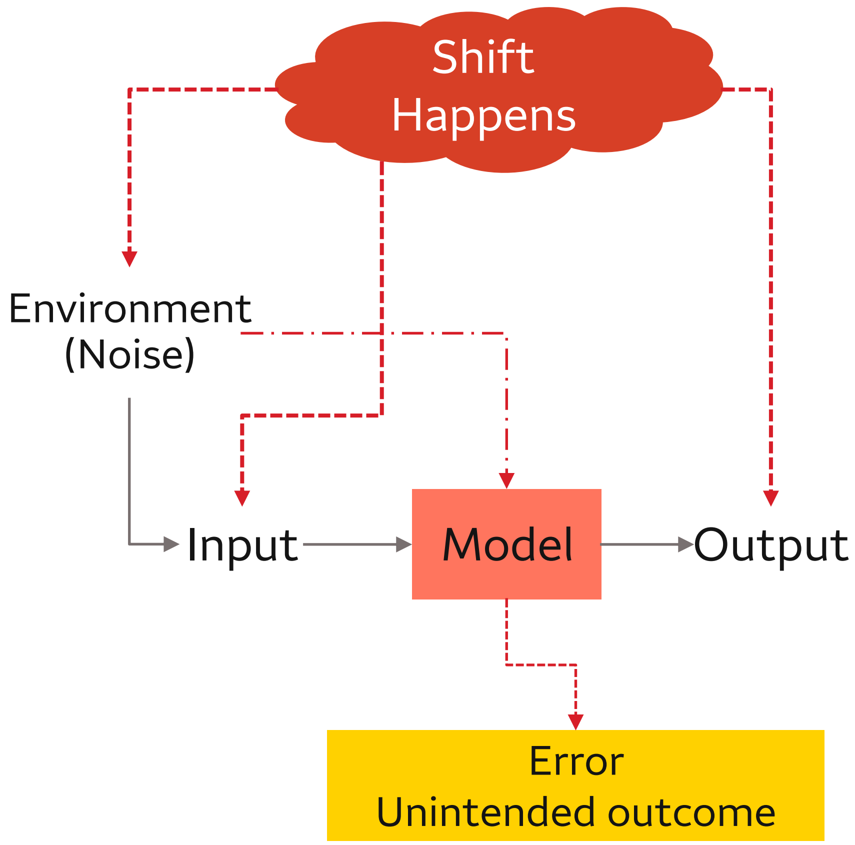




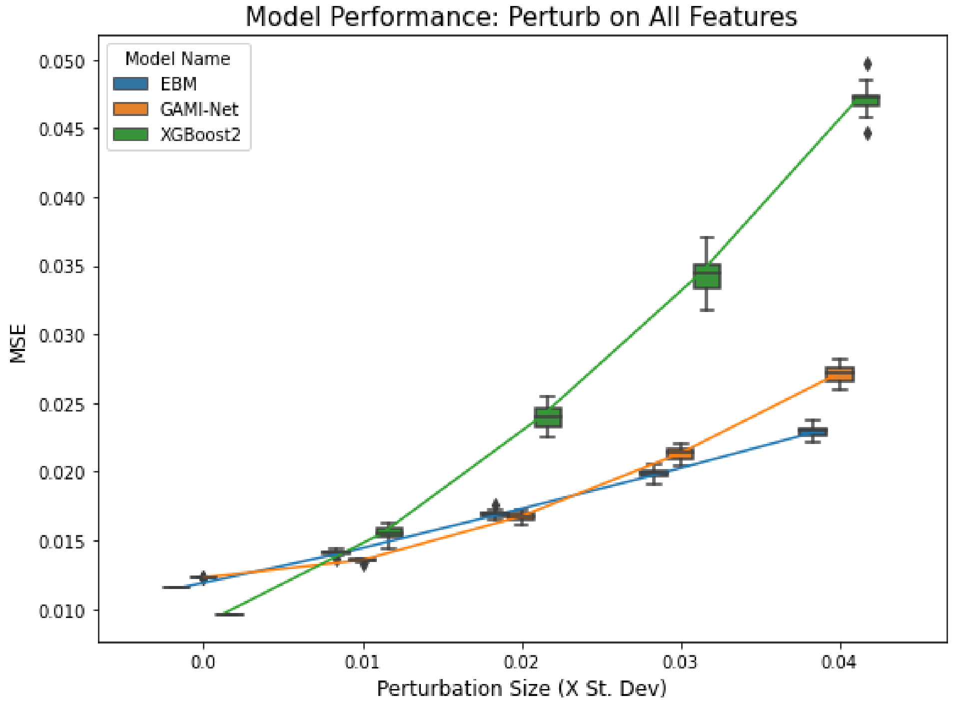
Example: Prediction Uncertainty for CaliforniaHousing data fit by GAMI-Net.

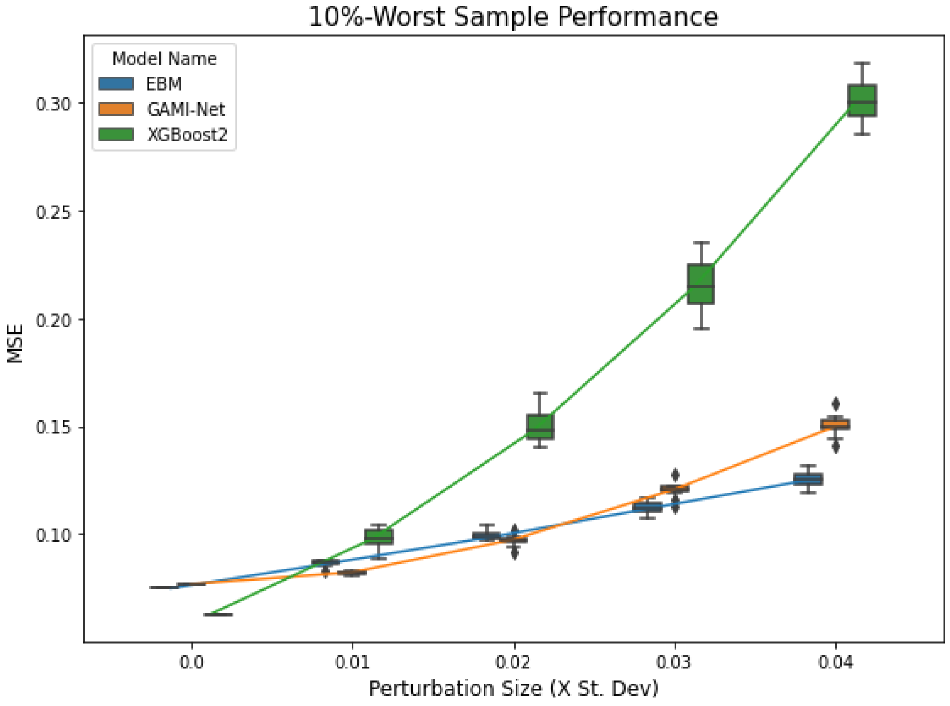
Robustness Test

- 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.
- Related topic: **resilience test** for various other out-of-distribution scenarios.



PiML Example: Robustness Testing





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

Examples using the PiML Toolbox

- Example: CaliforniaHousing data
- PiML Demo Examples based on Google Colab
- https://github.com/SelfExplainML/PiML-Toolbox/tree/main/docs/Workshop/202306-RiskLearning

- See also:
  - PiML User Guide: https://selfexplainml.github.io/PiML-Toolbox/
  - https://selfexplainml.github.io/PiML-Toolbox/_build/html/guides/cases/Example_CaliforniaHousing.html

WELLS FARGO

Thank you

Aijun Zhang, Ph.D.

Email: Aijun.Zhang@wellsfargo.com

LinkedIn: https://www.linkedin.com/in/ajzhang/