Model Diagnostics – Prediction Uncertainty

Aijun Zhang, Ph.D.
Corporate Model Risk, Wells Fargo

Model Validation Class, UNCC Master of Data Science, Fall 2023

Disclaimer: This material represents the views of the presenter and does not necessarily reflect those of Wells Fargo.

Model Diagnostics – Prediction Uncertainty

- **Machine Learning Model Uncertainty**
  - Sources of Uncertainty
  - Split Conformal Prediction
- Conformal Prediction for Regression Models
  - Naïve SCP Method
  - Conformal Residual Fitting
  - Conformalized Residual Quantile Regression
- Probability Calibration for Binary Classifiers
- Unreliable Region Detection
  - Feature Identification
  - Segmented Bandwidth/Uncertainty

PiML

An integrated Python toolbox for interpretable machine learning

```
pip install PiML
```

PiML Package: https://github.com/SelfExplainML/PiML-Toolbox

PiML User Guide: https://selfexplainml.github.io/PiML-Toolbox

Google Colab Notebooks:

- CaliforniaHousing Case (Regression)
- SimuCredit Case (Binary Classification)

Medium PiML Tutorials:

- 10/09/2023: Model Diagnostics – Error and Resilience
- 10/21/2023: Model Diagnostics – Overfitting and Robustness
- 10/31/2023: Model Diagnostics – Prediction Uncertainty (Todo)
- 11/xx/2023: Model Diagnostics – Bias and Fairness (Todo)

Machine Learning Model Uncertainty

- Quantifying uncertainty in machine learning prediction is critical for real-world decision making.
  - Additional layer of model transparency, for increasing reliability and level and confidence;
  - Particularly important in high-risk applications where uncertainty leads to serious consequences.
- **SR11-7 Model Risk Management, regarding Outcome Analysis**
  - Establishing expected ranges for those actual outcomes in relation to the intended objectives and assessing the reasons for observed variation
  - Back-testing involves the comparison of actual outcomes with model forecasts not used in model development. The comparison is generally done using expected ranges or statistical confidence intervals around model forecasts.
- **NIST's AI Risk Management Framework (Initial Draft, 03/17/2022)**
  - 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.

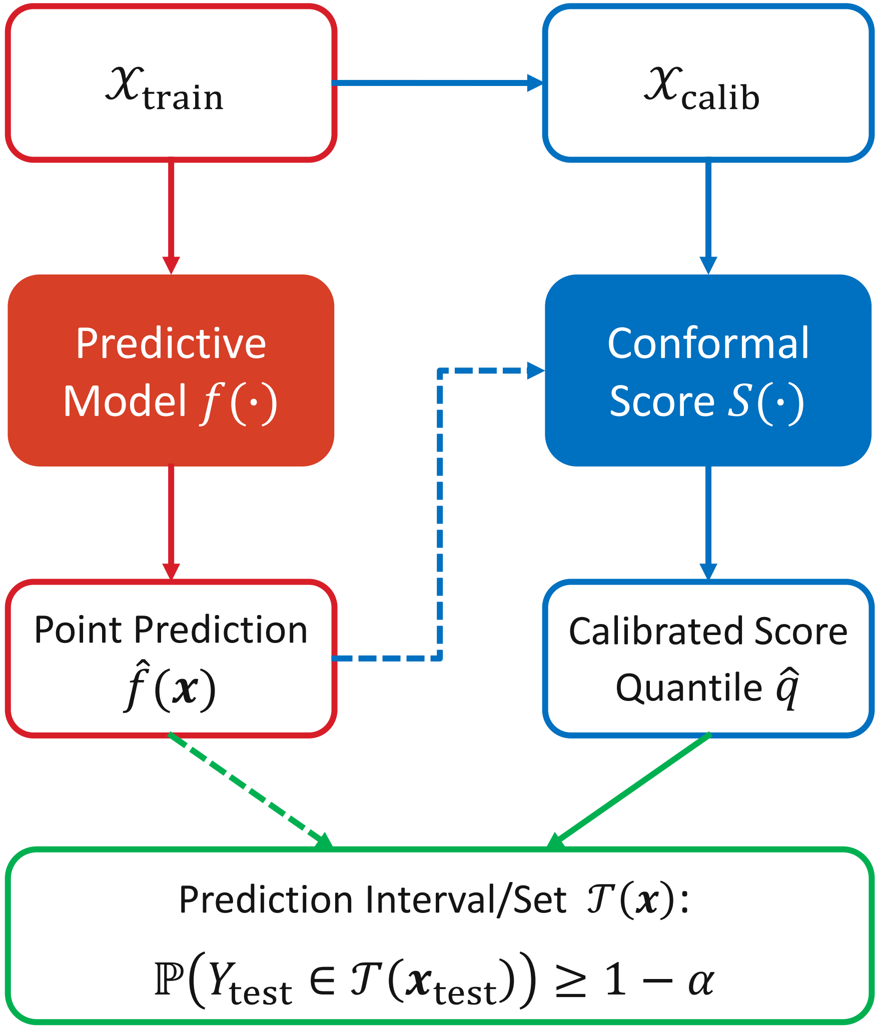
Sources of Uncertainty

- **Data uncertainty**
  - Data with noise: low or high, constant or heterogenous, outliers, distribution shift, etc.
  - Data sparsity in regions, with limited representation or learning capacity
  - Randomness in data partitioning (e.g., train-test-split)
- **Model uncertainty**
  - Hyperparameter tuning and feature selection leads to uncertain model prediction;
  - Stochastic optimization: random state, initialization, early stopping, non-guaranteed optimum.

- **This tutorial:** prediction uncertainty of a pre-trained model, where the uncertainty may come from data noise and sparsity and lack of model fit.
- **Future tutorial:** model retraining uncertainty due to a) random data splitting, b) hyperparameter tuning, and c) stochastic optimization.

Split Conformal Prediction

- **Conformal prediction** is a distribution-free uncertainty quantification (UQ) framework in machine learning:
  - Pioneered by Vladimir Vovk since 1990s; see *Vovk, Gammerman and Shafer (2005; 2022) or* alrw.net
  - A gentle introduction by Angelopoulos and Bates (2023) in *Foundations and Trends in Machine Learning or* arXiv
- **Split conformal prediction** (as illustrated):
  - Simple and easy to implement
  - Model-agnostic, applicable to arbitrary ML models
  - Prediction interval/set can be effectively generated for regression and multi-class problems, but less informative for binary classification;
  - Guaranteed coverage of true response in the unconditional or marginal sense, but impossible in the conditional sense.



PiML Tools for Quantifying Prediction Uncertainty

- PiML toolbox provides a diagnostic suite including the reliability test:
  - **exp.model_diagnose [reliability]:** a novel approach of split conformal prediction for regression models, a conventional approach of probability calibration for binary classification models, both including segmented bandwidth/uncertainty analysis;
  - **exp.model_compare[reliability]:** prediction uncertainty benchmarking analysis.
- PiML reliability test also supports unreliable region detection:
  - Slicing technique based on the quantified bandwidth/uncertainty
  - Distribution shift analysis between unreliable and reliable samples
  - Surrogate modeling for feature identification w.r.t. prediction uncertainty
- Larger bandwidth $\rightarrow$ Wider prediction interval $\rightarrow$ Less reliable prediction

Model Diagnostics – Prediction Uncertainty

- Machine Learning Model Uncertainty
  - Sources of Uncertainty
  - Split Conformal Prediction
- **Conformal Prediction for Regression Models**
  - Naïve SCP Method
  - Conformal Residual Fitting
  - Conformalized Residual Quantile Regression
- Probability Calibration for Binary Classifiers
- Unreliable Region Detection
  - Feature Identification
  - Segmented Bandwidth/Uncertainty

Math Behind Conformal Prediction

- Suppose $W_1, \dots, W_n, W_{n+1}$ are exchangeable (i.e., permutation invariant, weaker than i.i.d.)

a) The rank of W_{n+1} is uniformly distributed over $1, 2, \dots, n+1$.

b) For $\alpha \in [0,1]$, th probability that W_{n+1} is among the $\lceil (n+1)(1-\alpha) \rceil$ smallest of $W_1, \dots, W_{n+1}$ is given by

$$\mathbb{P}(\mathrm{rank}(W_{n+1}) \leq \lceil (n+1)(1-\alpha) \rceil) = \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1}$$

c) Let $\hat{q} = \mathrm{Quantile}\left(\{W_1, \dots, W_n\}; \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1}\right)$

d) We can derive that $\mathbb{P}(W_{n+1} \leq \hat{q}) = \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1} \in \left[1-\alpha, 1-\alpha+\frac{1}{n+1}\right)$.

Split Conformal Prediction: Procedure

- Given a pre-trained regression model $\hat{f}(\boldsymbol{x})$, a hold-out calibration data $\mathcal{X}_{\text{calib}}$, a test sample $\boldsymbol{x}_{\text{test}}$, and an error rate α (say 0.1). Define a conformal score measuring the prediction uncertainty, $S(\boldsymbol{x}, y, \hat{f}) \in \mathbb{R}$, which is assumed exchangeable among calibration and testing samples.
  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 $\hat{q} = \text{Quantile}\left(\{S_1, \dots, S_n\}; \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1}\right)$;
  3) Construct the prediction set for the test sample $\boldsymbol{x}_{\text{test}}$ by $\mathcal{T}(\boldsymbol{x}_{\text{test}}) = \left\{y: S\left(\boldsymbol{x}_{\text{test}}, y, \hat{f}(\boldsymbol{x}_{\text{test}})\right) \le \hat{q}\right\}$.
- Under the exchangeability condition of conformal scores, we have the coverage guarantee

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

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

Split Conformal Prediction: Scores

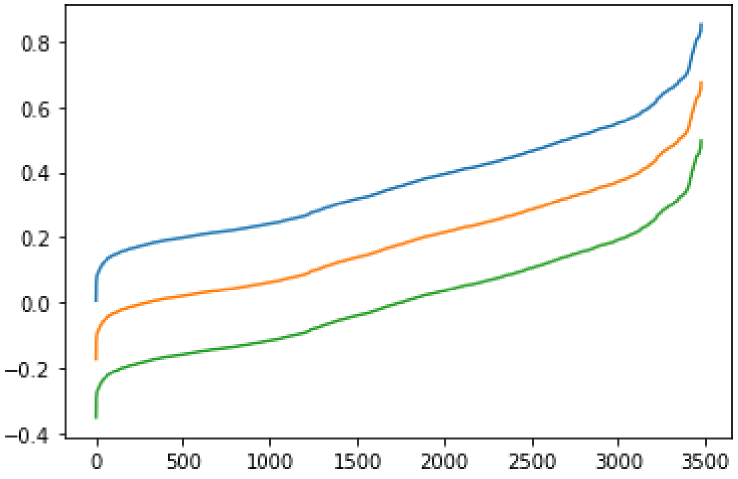
- **Naïve SCP** based on $S(\boldsymbol{x}, y, \hat{f}) = |y - \hat{f}(\boldsymbol{x})|$, generates prediction intervals with constant bandwidth $\mathcal{T}(\boldsymbol{x}_{\text{test}}) = \{y\colon |y - \hat{f}(\boldsymbol{x}_{\text{test}})| \le \hat{q}\}$

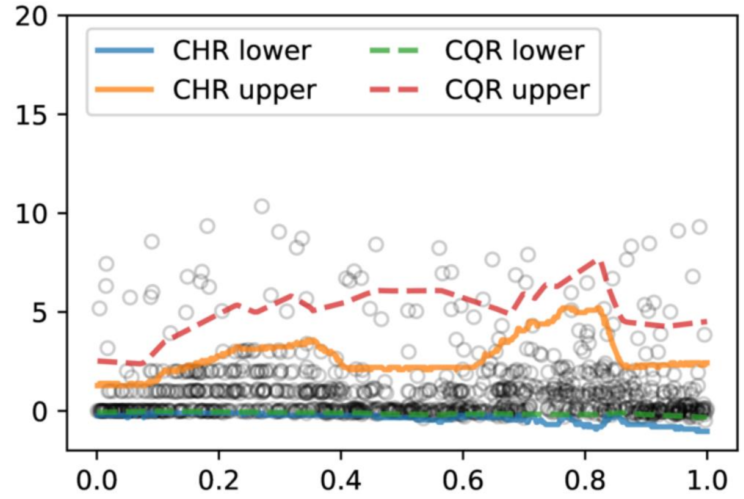
- **Conformal residual fitting (CRF)** with locally adaptive conformal score $S(\boldsymbol{x}, y, \hat{f}) = \frac{|y-\hat{f}(\boldsymbol{x})|}{\sigma(\boldsymbol{x})}$, where $\sigma(\boldsymbol{x})$ is trained with an auxiliary model on hold-out sample $\{(\boldsymbol{x}, |y - \hat{f}(\boldsymbol{x})|), \boldsymbol{x} \in \mathcal{X}_{\text{res}}\}$.

- Both naïve SCP and CRF can be implemented by MAPIE.

- **CQR** (conformalized quantile regression, by Romano, et al. 2019) and **CHR** (conditional histogram regression, by Sesia and Romano, 2021) are not directly suitable for pre-trained model diagnostics.

- PiML-reliability test: **residual-based CQR** for pre-trained models.





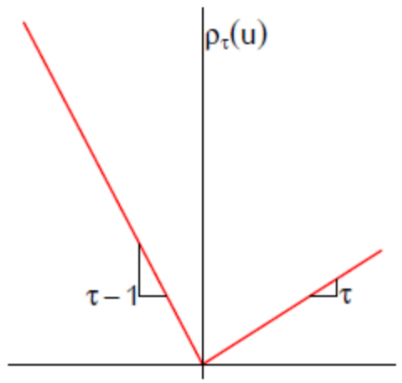
CRQR (Conformalized Residual Quantile Regression)

1. Fit a GBM with quantile loss on $\{\boldsymbol{x}_i, y_i - \hat{f}(\boldsymbol{x}_i), i \in \mathcal{X}_{\text{res}}\}$ (holdout sample) to predict the residual quantiles $[\hat{g}_{\alpha/2}(\boldsymbol{x}), \hat{g}_{1-\alpha/2}(\boldsymbol{x})]$;
2. Define score $S(\boldsymbol{x}, y, \hat{f}) = \max\{\hat{g}_{\alpha/2}(\boldsymbol{x}) - y + \hat{f}(\boldsymbol{x}),\ y - \hat{f}(\boldsymbol{x}) - \hat{g}_{1-\alpha/2}(\boldsymbol{x})\}$
3. Calculate $\hat{q} = \text{Quantile}\left(\{S_1, \dots, S_n\}; \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1}\right)$, using $S(\boldsymbol{x}, y, \hat{f})$ on $\mathcal{X}_{\text{calib}}$
4. Construct the prediction interval for the test sample $\boldsymbol{x}_{\text{test}}$ by

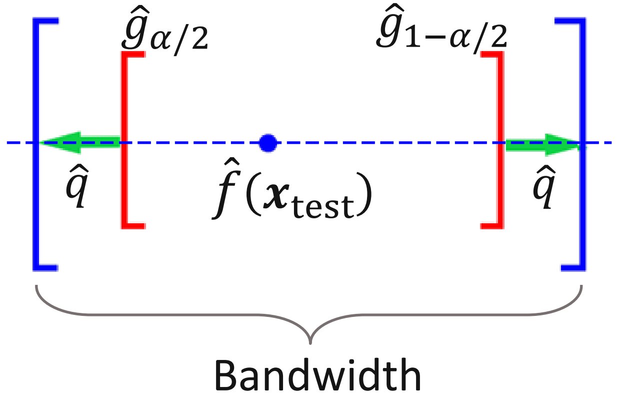
$$\mathcal{T}(\boldsymbol{x}_{\text{test}}) = \left[\hat{f}(\boldsymbol{x}_{\text{test}}) + \hat{g}_{\alpha/2}(\boldsymbol{x}_{\text{test}}) - \hat{q},\ \ \hat{f}(\boldsymbol{x}_{\text{test}}) + \hat{g}_{1-\alpha/2}(\boldsymbol{x}_{\text{test}}) + \hat{q}\right]$$

Interpretation of $\mathcal{T}(\boldsymbol{x}_{\text{test}})$:

the final prediction interval is composed of three terms, namely the original prediction, the fitted residual quantiles, and the calibrated adjustment.



Quantile loss



PiML Demo: Regression Case

- Consider the CaliforniaHousing case with existing data and model pipelines (Google Colab Notebook)

```
from xgboost import XGBRegressor
XGB = XGBRegressor(max_depth=5, n_estimators=500)
exp.model_train(model=XGB, name='XGB5')
```

```
exp.model_diagnose(model="XGB5", show="reliability_table", alpha=0.1)
```

	Empirical Coverage	Average Bandwidth
0	0.890975	0.27162

```
from sklearn.neural_network import MLPRegressor
DNN = MLPRegressor(hidden_layer_sizes=[40]*4,
                   activation="relu", random_state=0)
exp.model_train(model=DNN, name='ReLUDNN')
```

```
exp.model_diagnose(model="ReLUDNN", show="reliability_table", alpha=0.1)
```

	Empirical Coverage	Average Bandwidth
0	0.896426	0.307637

Model Diagnostics – Prediction Uncertainty

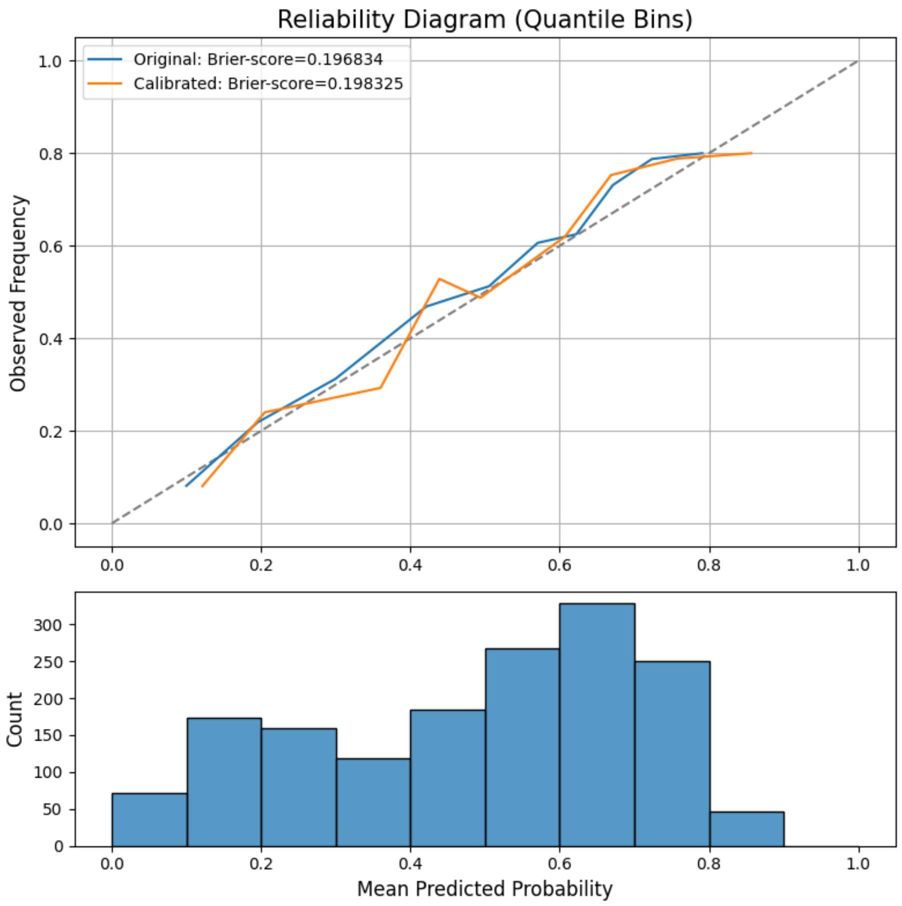
- Machine Learning Model Uncertainty
  - Sources of Uncertainty
  - Split Conformal Prediction
- Conformal Prediction for Regression Models
  - Naïve SCP Method
  - Conformal Residual Fitting
  - Conformalized Residual Quantile Regression
- **Probability Calibration for Binary Classifiers**
- Unreliable Region Detection
  - Feature Identification
  - Segmented Bandwidth/Uncertainty

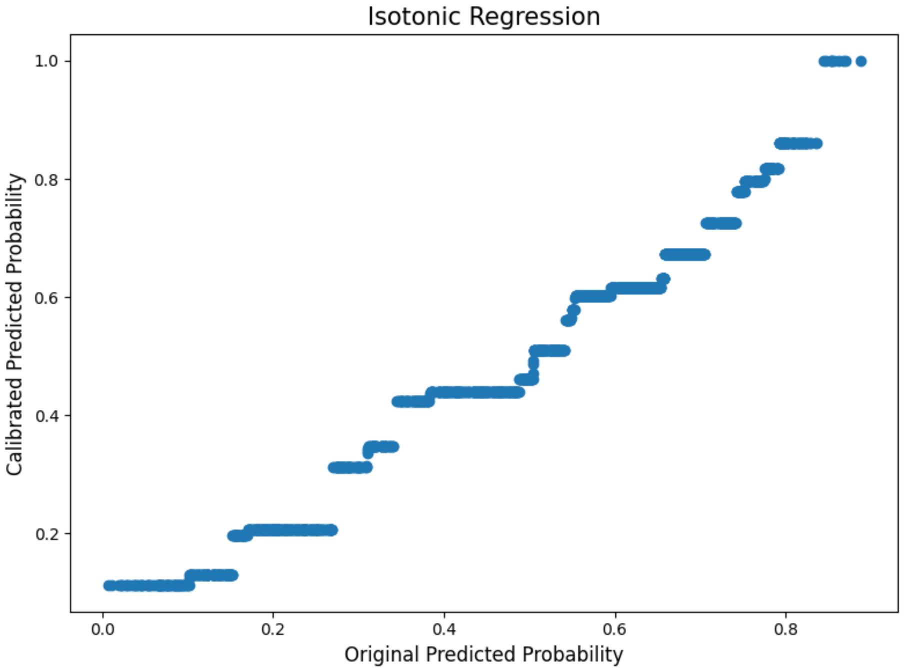
Probability Calibration for Binary Classifiers

- The simple and easy conformal prediction does not work as effectively for the binary classification case.
- We take a conventional approach of using **predict_proba** $\hat{p} = \mathbb{P}(Y = 1|\boldsymbol{x})$ and measure the uncertainty by the quantity $\sqrt{\hat{p}(1-\hat{p})}$ for each point prediction.
- **Caveat:** there is no statistical guarantee of correct coverage of the true class.
- However, probability calibration is needed for raw predict_proba by some ML models, so the predicted probabilities align with the observed class frequencies, as shown by the reliability diagram or measured through the Brier score.
- There are lots of tutorials online, so we don't repeat here.
- In PiML, we adopt the isotonic regression to calibrate the predicted probabilities as a monotonic step function; while Platt scaling is a parametric sigmoid curve.

PiML Demo: Binary Classification Case

- Consider the SimuCredit case with existing data and model pipeline (Google Colab Notebook)





Model Diagnostics – Prediction Uncertainty

- Machine Learning Model Uncertainty
  - Sources of Uncertainty
  - Split Conformal Prediction
- Conformal Prediction for Regression Models
  - Naïve SCP Method
  - Conformal Residual Fitting
  - Conformalized Residual Quantile Regression
- Probability Calibration for Binary Classifiers
- **Unreliable Region Detection**
  - Feature Identification
  - Segmented Bandwidth/Uncertainty

Unreliable Region Detection

- PiML reliability test supports unreliable region detection, by utilizing the slicing technique on the test sample-wise bandwidth/uncertainty quantification.

- **A Practical User Guide:**

  1) Identify the features sensitive to prediction uncertainty

     – Distribution shift analysis between unreliable and reliable samples (thresholding), or

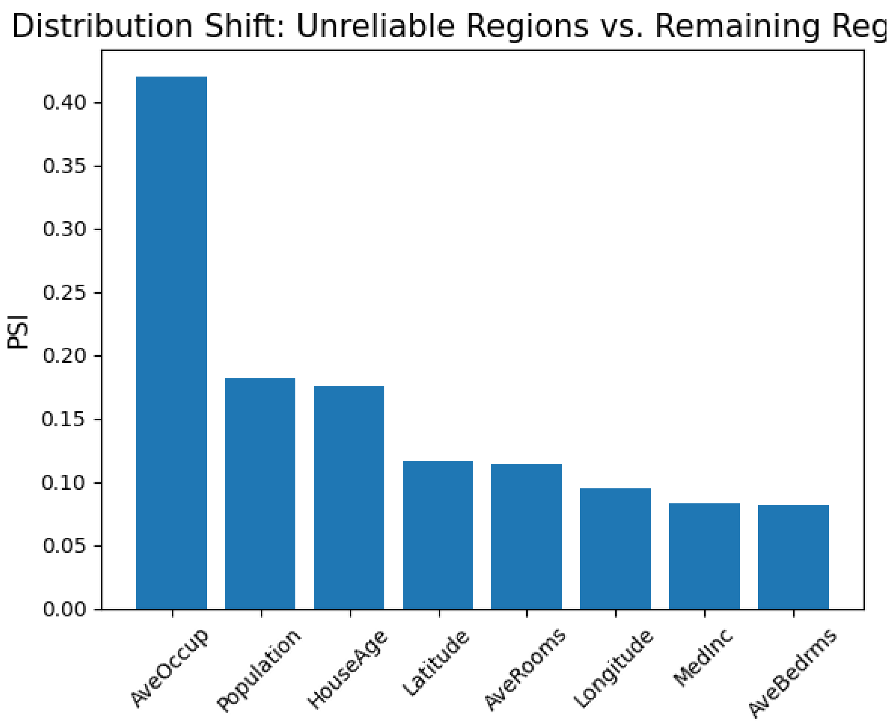
     – Feature importance of a surrogate model fitted on $\{(\boldsymbol{x}_i, Bandwidth(\boldsymbol{x}_i)), i \in \mathcal{X}_{\text{test}}\}$

  2) Perform segmented bandwidth analysis (i.e., slicing) according to identified features.

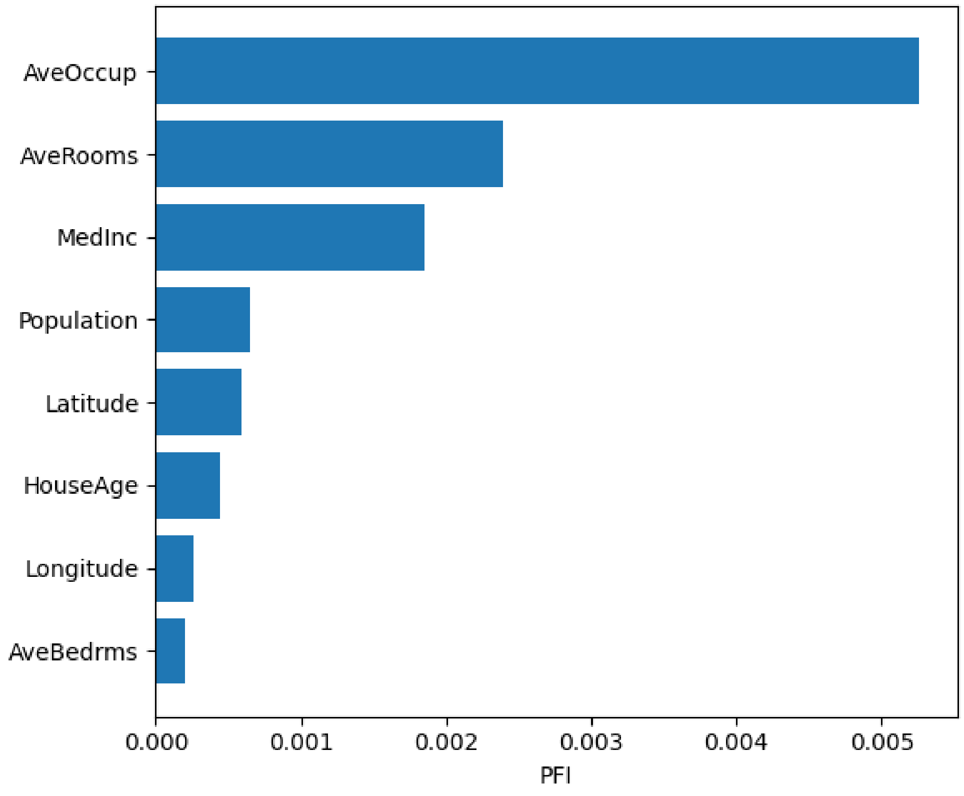
  3) Verify the diagnostic result jointly with weak spot and other tests.

Feature Identification w.r.t. Prediction Uncertainty

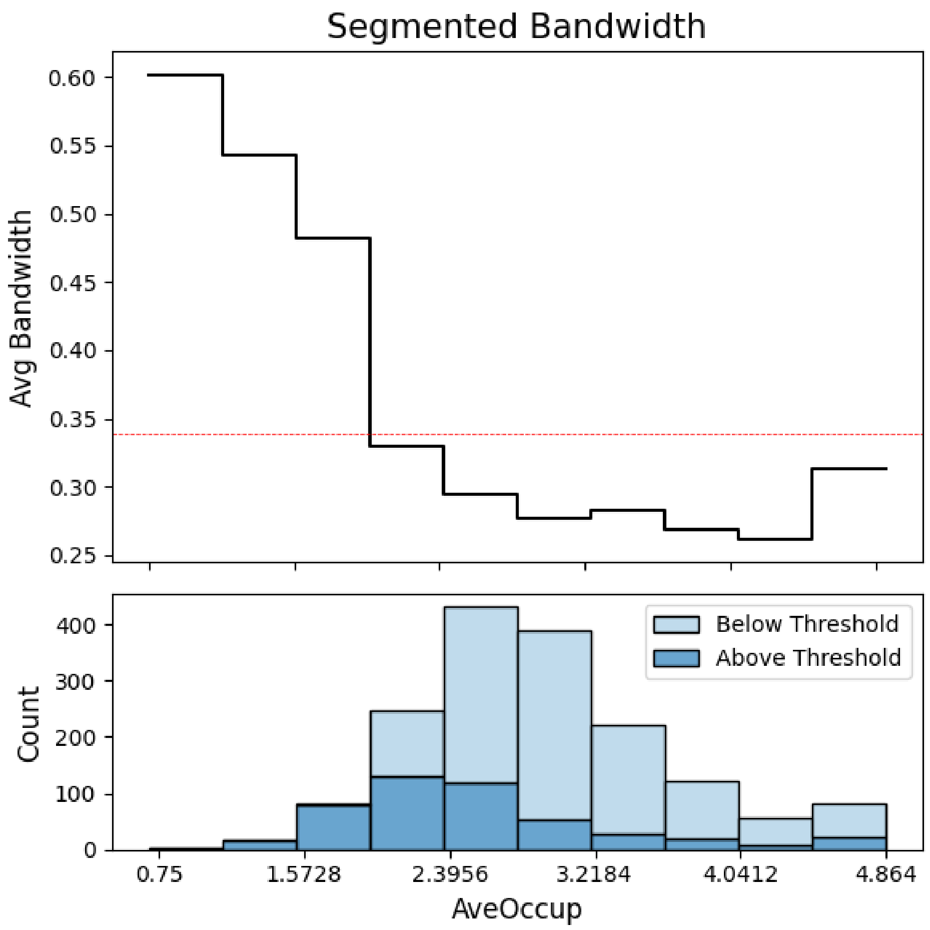
```
# Measuring distribution shift between unreliable and reliable samples
exp.model_diagnose(model="ReLUDNN",
                   show="reliability_distance", alpha=0.1,
                   threshold=1.1, distance_metric="PSI",
                   figsize=(6,5))
```

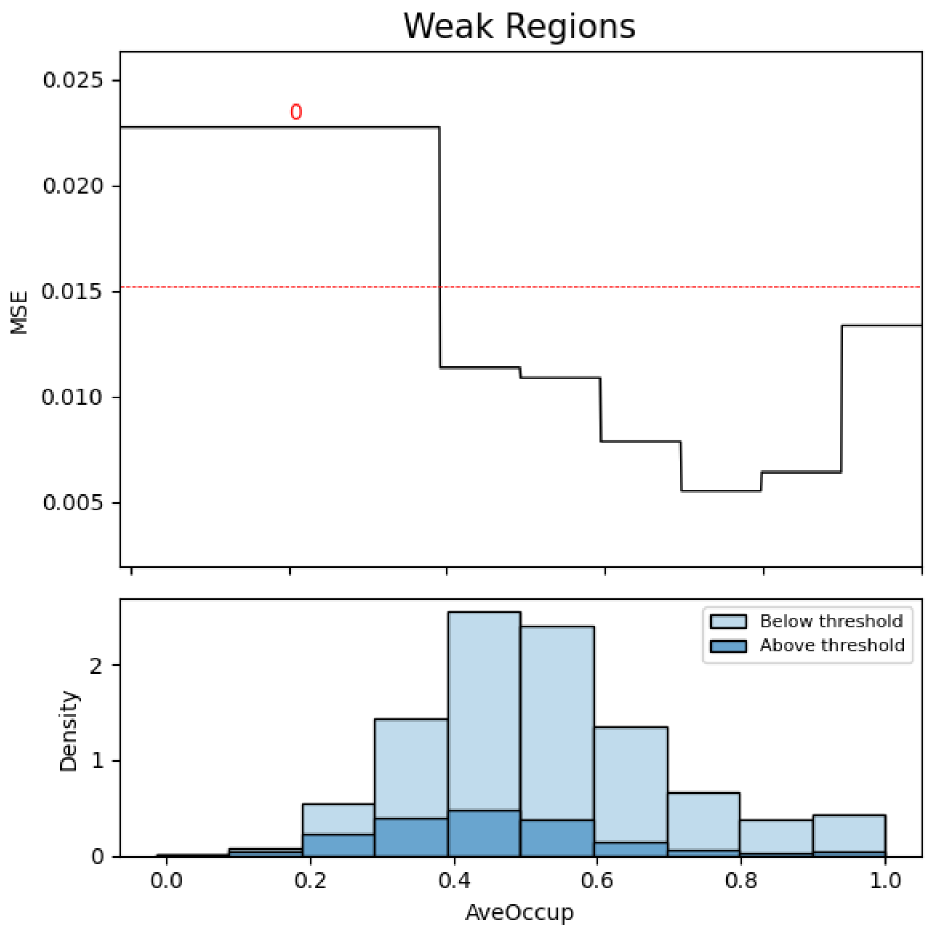



```
PredUQ_SurrogatePFI(model_name="ReLUDNN", alpha=0.1,
                    max_depth=1, max_trees=200)
```



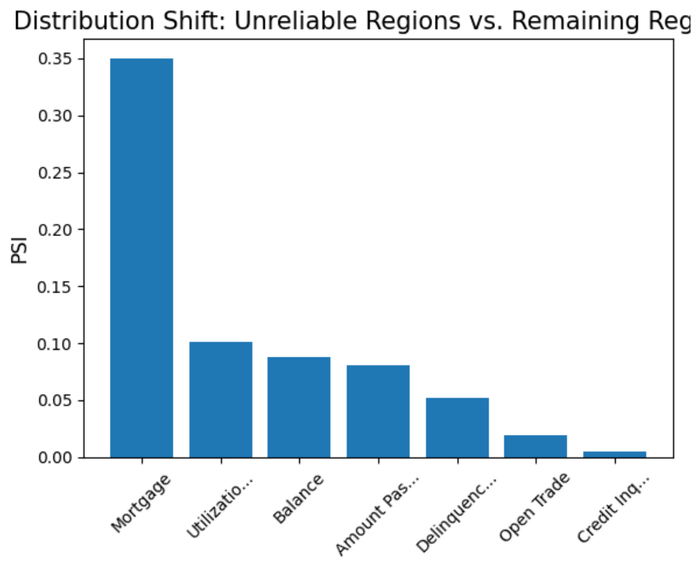
PiML Demo: CaliforniaHousing Regression Case

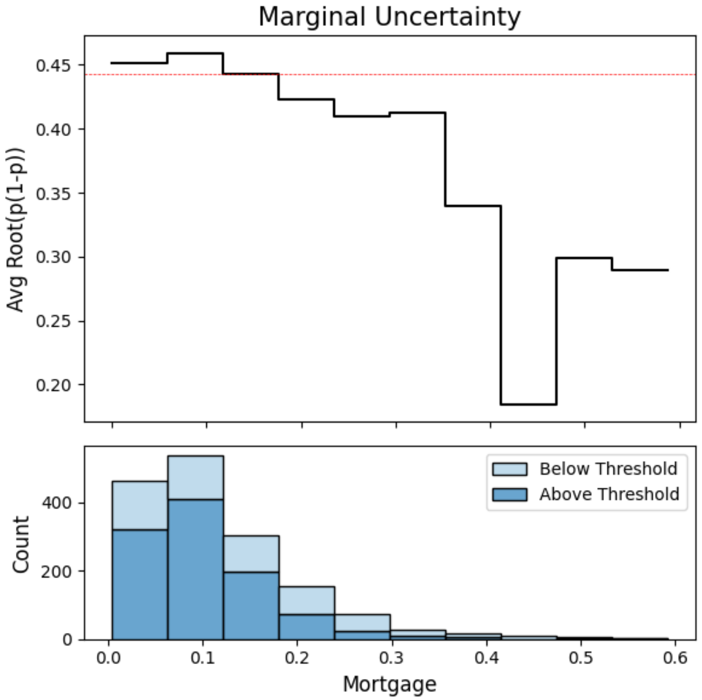


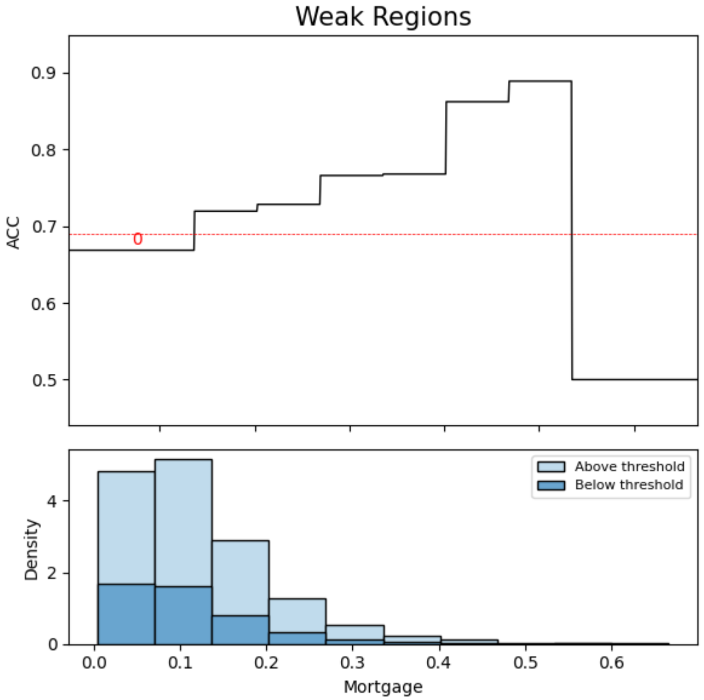


PiML Demo: SimuCredit Binary Classification Case

```
# Measuring distribution shift between unreliable and reliable samples
exp.model_diagnose(model="ReLUDNN",
                   show="reliability_distance",
                   threshold=1.1, distance_metric="PSI",
                   figsize=(6,5))
```







Thank you

Aijun Zhang, Ph.D.

Email: Aijun.Zhang@wellsfargo.com

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

An integrated Python toolbox for interpretable machine learning

```
pip install PiML
```

PiML Package: https://github.com/SelfExplainML/PiML-Toolbox

PiML User Guide: https://selfexplainml.github.io/PiML-Toolbox

Google Colab Notebooks:

- CaliforniaHousing Case (Regression)
- SimuCredit Case (Binary Classification)

Medium PiML Tutorials:

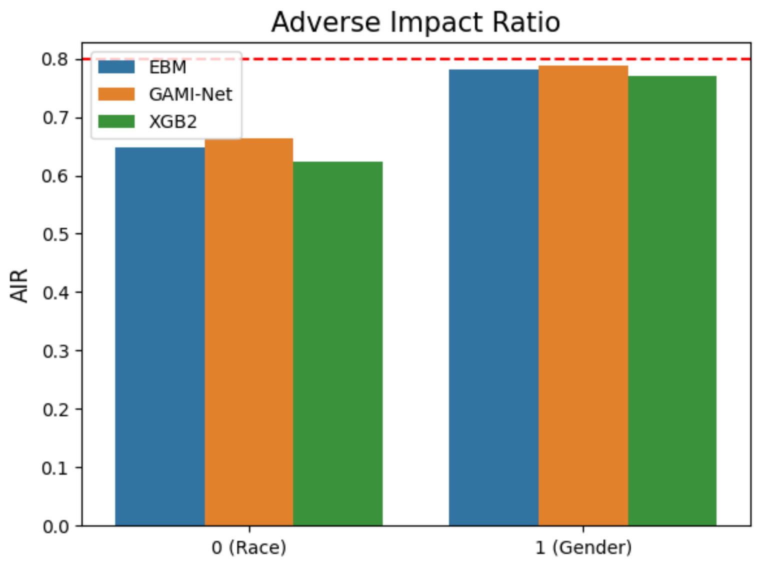
- 10/09/2023: Model Diagnostics – Error and Resilience
- 10/21/2023: Model Diagnostics – Overfitting and Robustness
- 10/31/2023: Model Diagnostics – Prediction Uncertainty (Todo)
- 11/xx/2023: Model Diagnostics – Bias and Fairness (Todo)

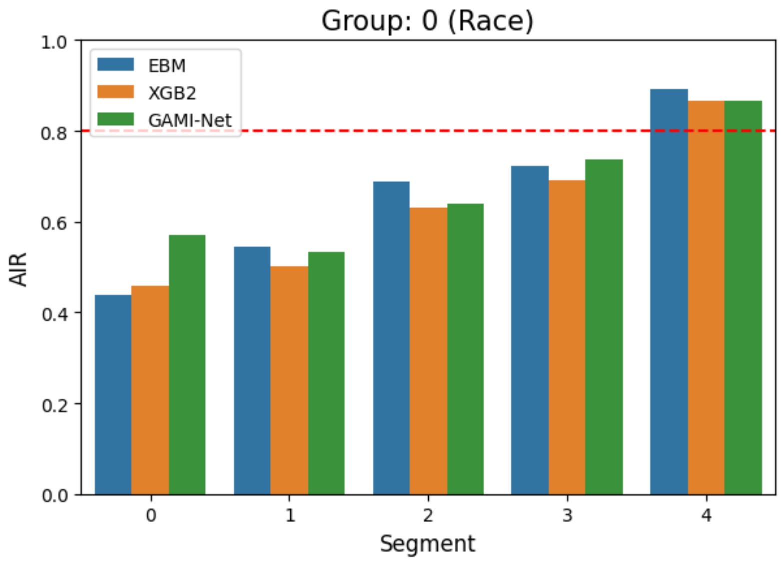
Bias and Fairness

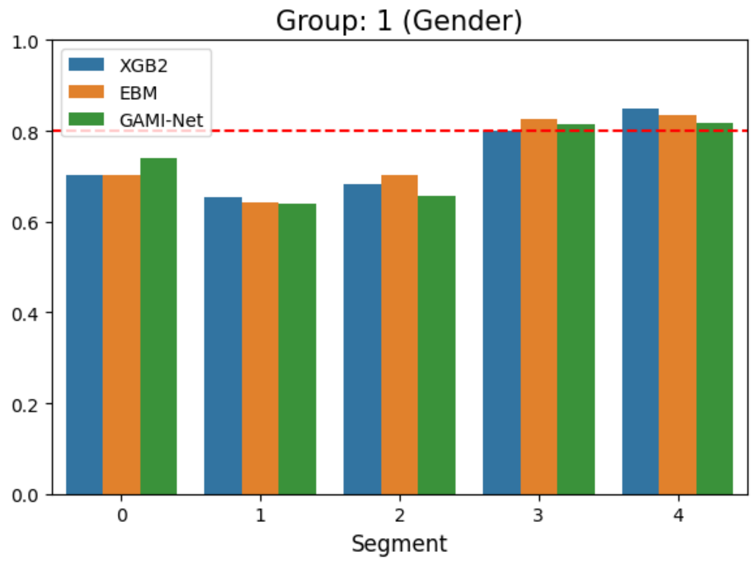
- For each demographic feature (Race, Gender), consider AIR between protected group vs reference group.

$$AIR = \frac{(TP_p + FN_p)/n_r}{(TP_r + FN_r)/n_p}$$

- AIR below 0.8 is a sign of bias and unfairness.
- PiML provides segmented metrics conditional on a modeling variable (e.g., Balance below). It also provides methods to debias through feature binning and decision thresholding.

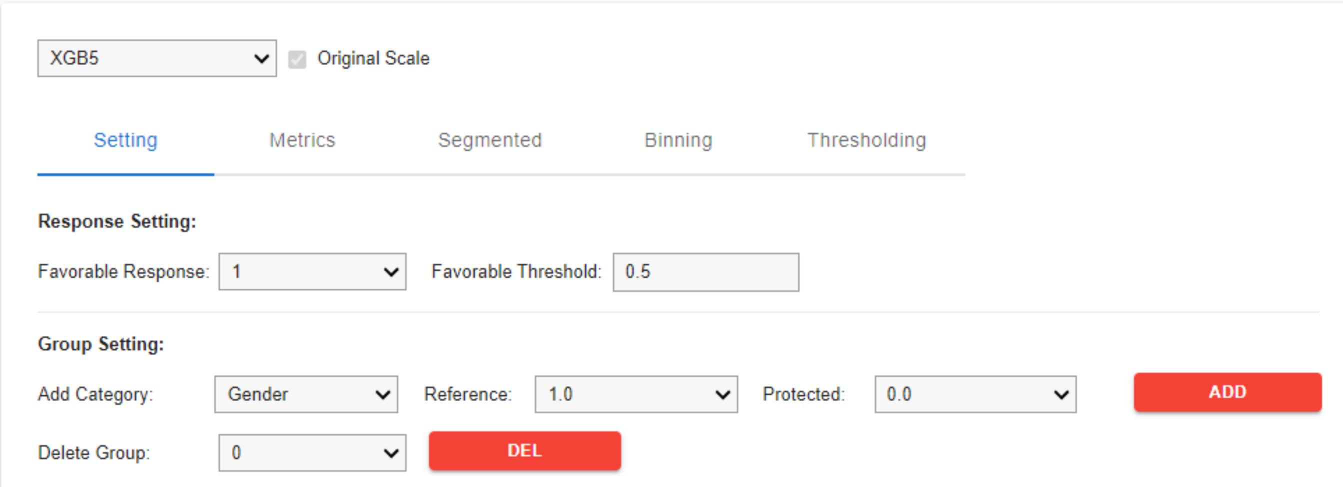




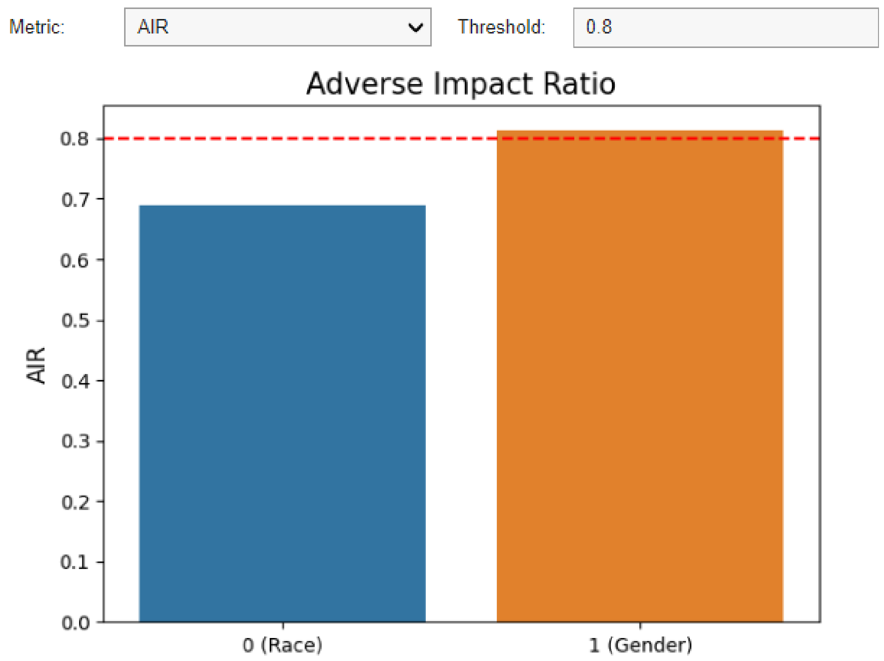


PiML Demo: Bias and Fairness

```
exp.model_fairness()
```



Group Index	Group Category	Reference Group	Protected Group
0	Race	1.0	0.0
1	Gender	1.0	0.0



Group Index	Group Category	Reference Group	Protected Group	AIR
0	Race	1.0	0.0	0.689346
1	Gender	1.0	0.0	0.813465