Session 58, Does It Really Work? - Validating Predictive Models

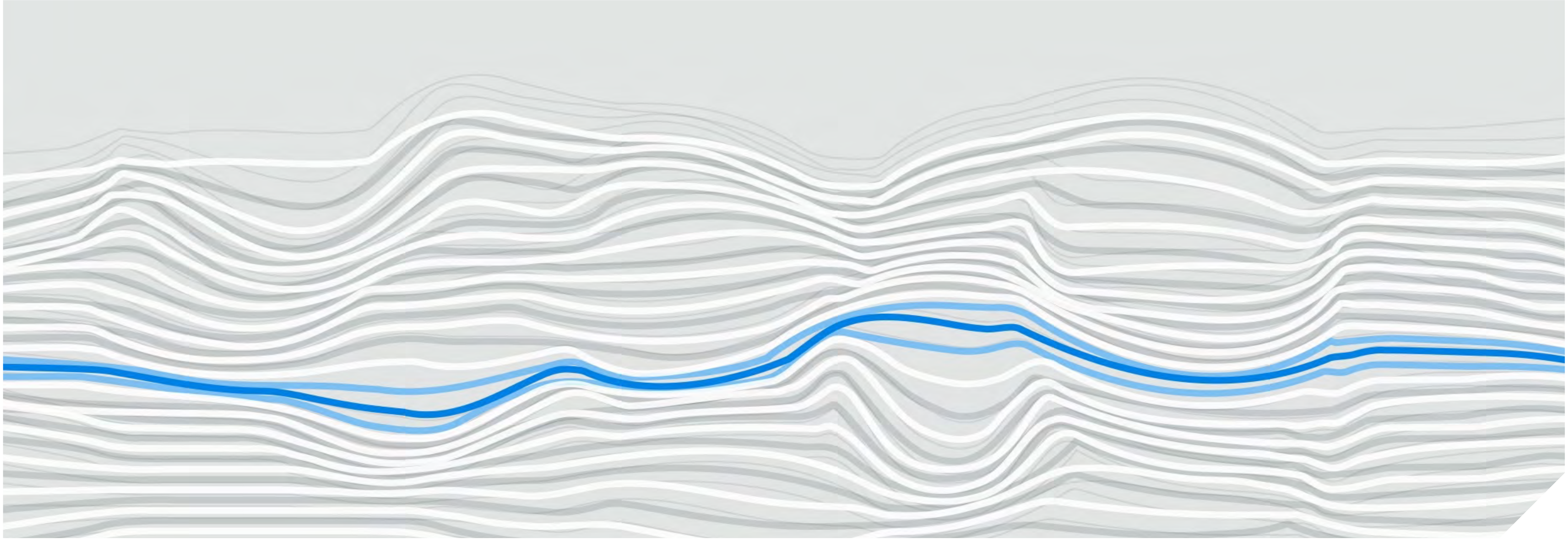
SOA Antitrust Disclaimer
SOA Presentation Disclaimer

Milliman

2019 Health SOA Meeting

Session 058: Does It Really Work? - Validating Predictive Models

Michael Niemerg, FSA, MAAA
Erica Rode, PhD, FSA, MAAA



SOA Antitrust Compliance Guidelines

Active participation in the Society of Actuaries is an important aspect of membership. While the positive contributions of professional societies and associations are well-recognized and encouraged, association activities are vulnerable to close antitrust scrutiny. By their very nature, associations bring together industry competitors and other market participants.

The United States antitrust laws aim to protect consumers by preserving the free economy and prohibiting anti-competitive business practices; they promote competition. There are both state and federal antitrust laws, although state antitrust laws closely follow federal law. The Sherman Act, is the primary U.S. antitrust law pertaining to association activities. The Sherman Act prohibits every contract, combination or conspiracy that places an unreasonable restraint on trade. There are, however, some activities that are illegal under all circumstances, such as price fixing, market allocation and collusive bidding.

There is no safe harbor under the antitrust law for professional association activities. Therefore, association meeting participants should refrain from discussing any activity that could potentially be construed as having an anti-competitive effect. Discussions relating to product or service pricing, market allocations, membership restrictions, product standardization or other conditions on trade could arguably be perceived as a restraint on trade and may expose the SOA and its members to antitrust enforcement procedures.

While participating in all SOA in person meetings, webinars, teleconferences or side discussions, you should avoid discussing competitively sensitive information with competitors and follow these guidelines:

- -**Do not** discuss prices for services or products or anything else that might affect prices
- -**Do not** discuss what you or other entities plan to do in a particular geographic or product markets or with particular customers.
- -**Do not** speak on behalf of the SOA or any of its committees unless specifically authorized to do so.
- -**Do** leave a meeting where any anticompetitive pricing or market allocation discussion occurs.
- -**Do** alert SOA staff and/or legal counsel to any concerning discussions
- -**Do** consult with legal counsel before raising any matter or making a statement that may involve competitively sensitive information.

Adherence to these guidelines involves not only avoidance of antitrust violations, but avoidance of behavior which might be so construed. These guidelines only provide an overview of prohibited activities. SOA legal counsel reviews meeting agenda and materials as deemed appropriate and any discussion that departs from the formal agenda should be scrutinized carefully. Antitrust compliance is everyone’s responsibility; however, please seek legal counsel if you have any questions or concerns.

Limitations

- The views expressed in this presentation are those of the presenters, and not those of Milliman. Nothing in this presentation is intended to represent a professional opinion or be an interpretation of actuarial standards of practice.

Validating Predictive Models

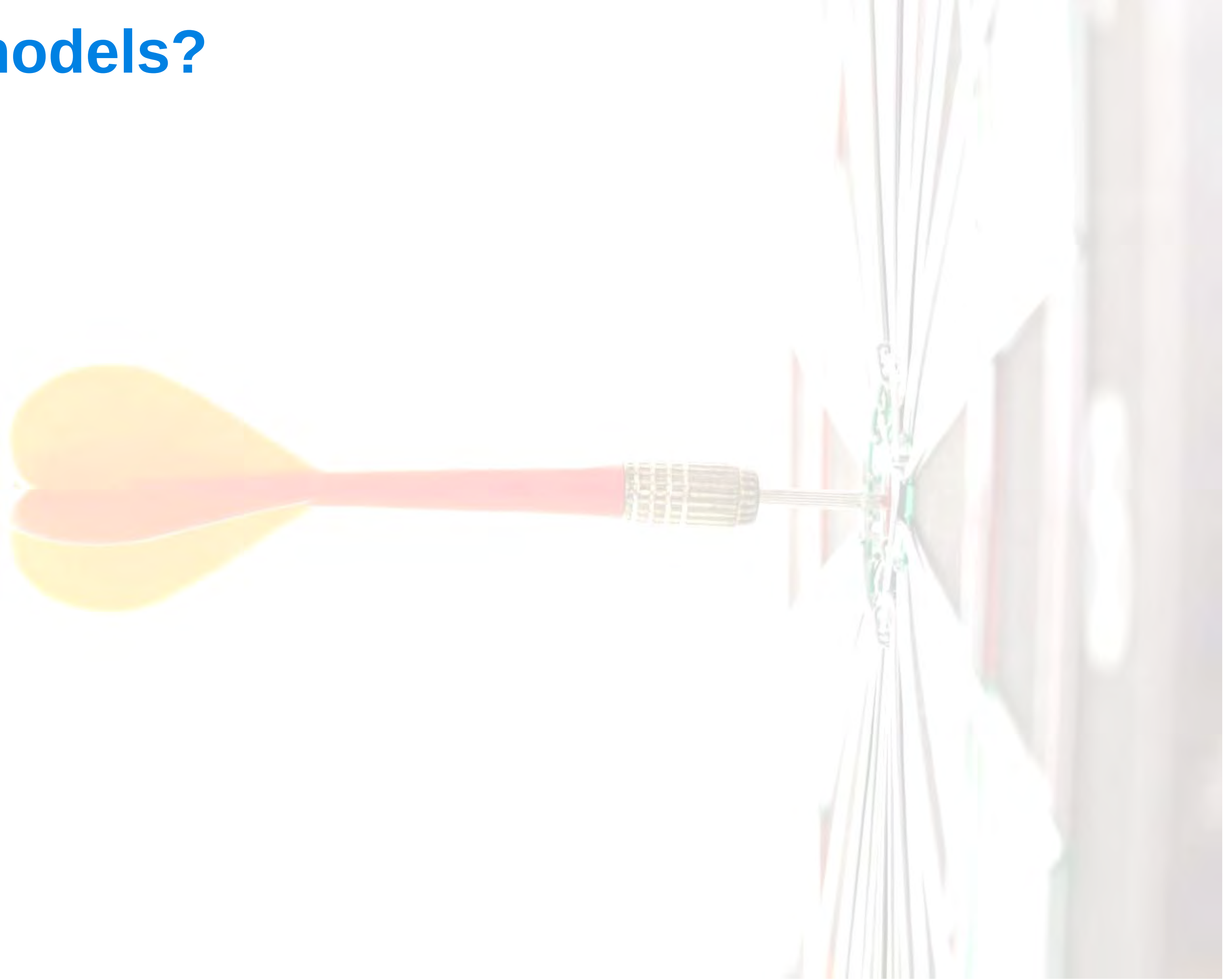
Erica Rode, PhD, FSA, MAAA

JUNE 25, 2019

Milliman

Why do we validate models?

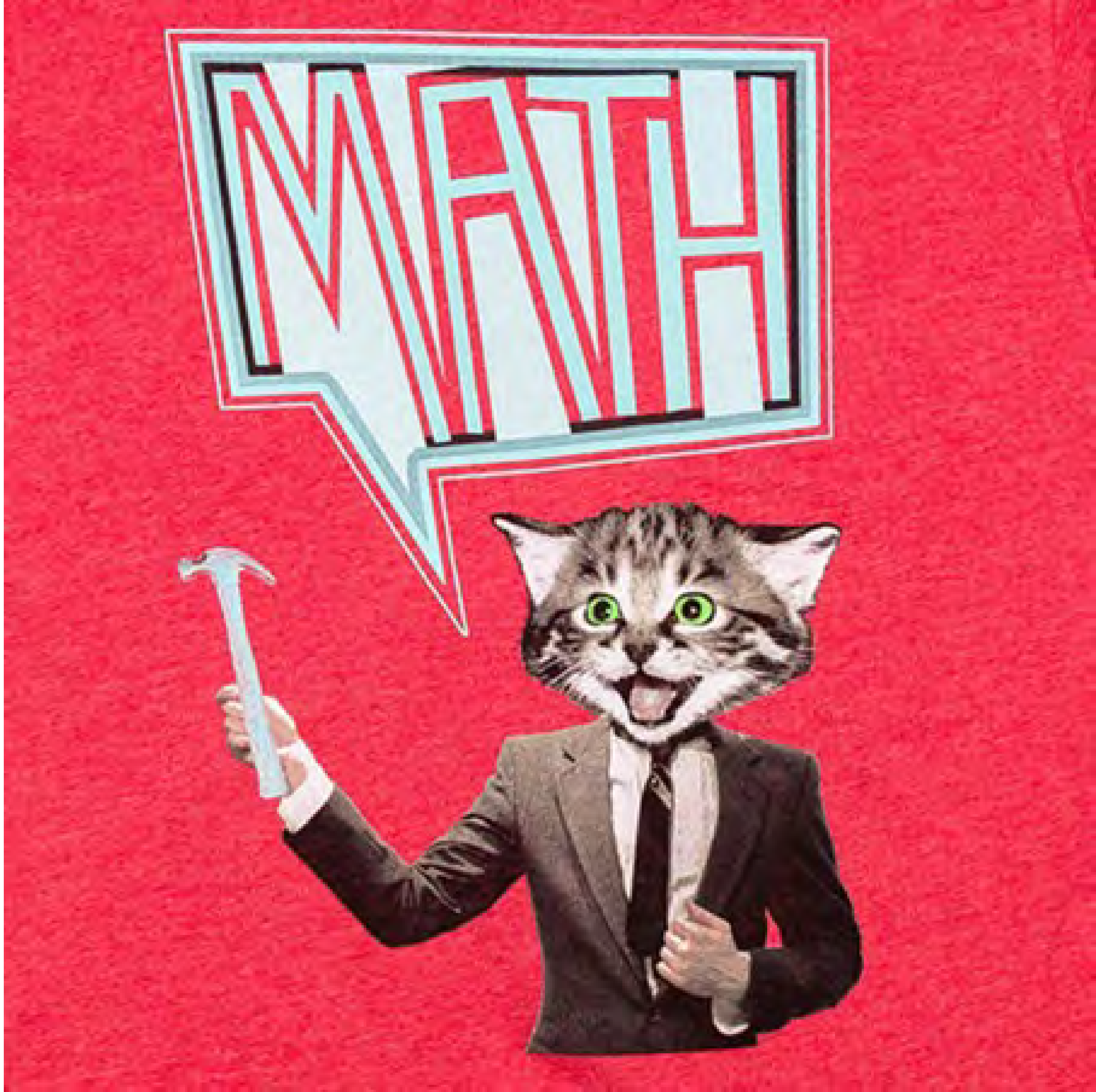
- Identify potential problems
  - Bias
  - Overfitting
  - Face validity problems
- Ensure proper implementation
- Monitoring over time



Let's get this out of the way…

- R^2
- MAPE

How we *don't* validate a model

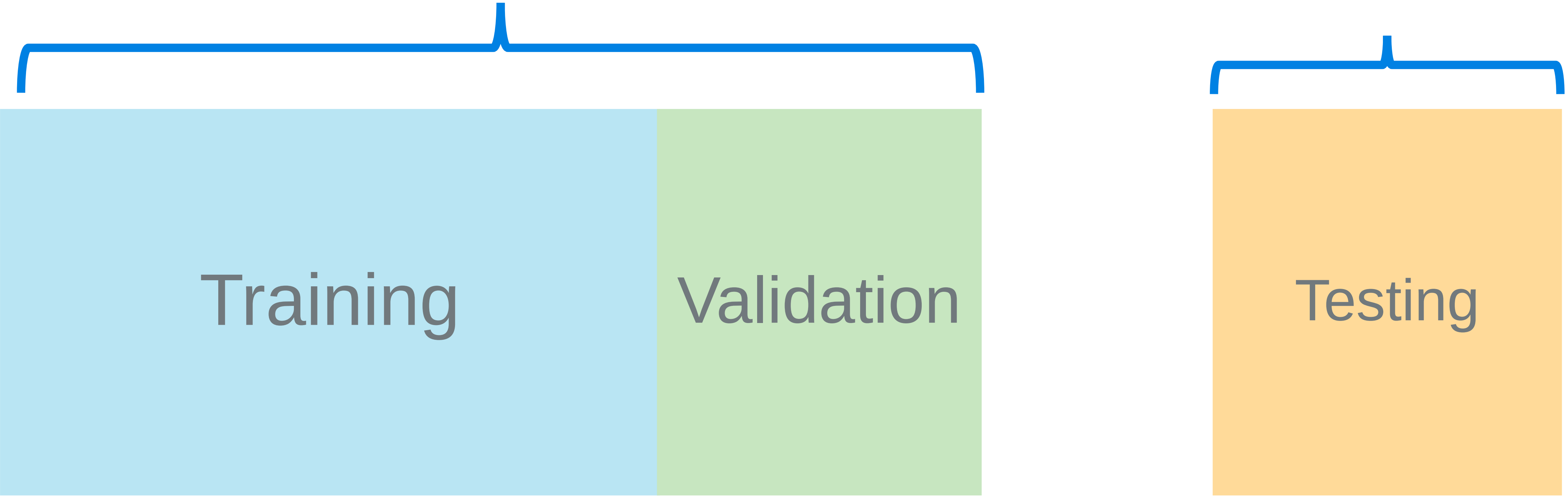


Partitioning the data

Training, validating, and testing

Used in building the model

Sits on the bench waiting for its turn

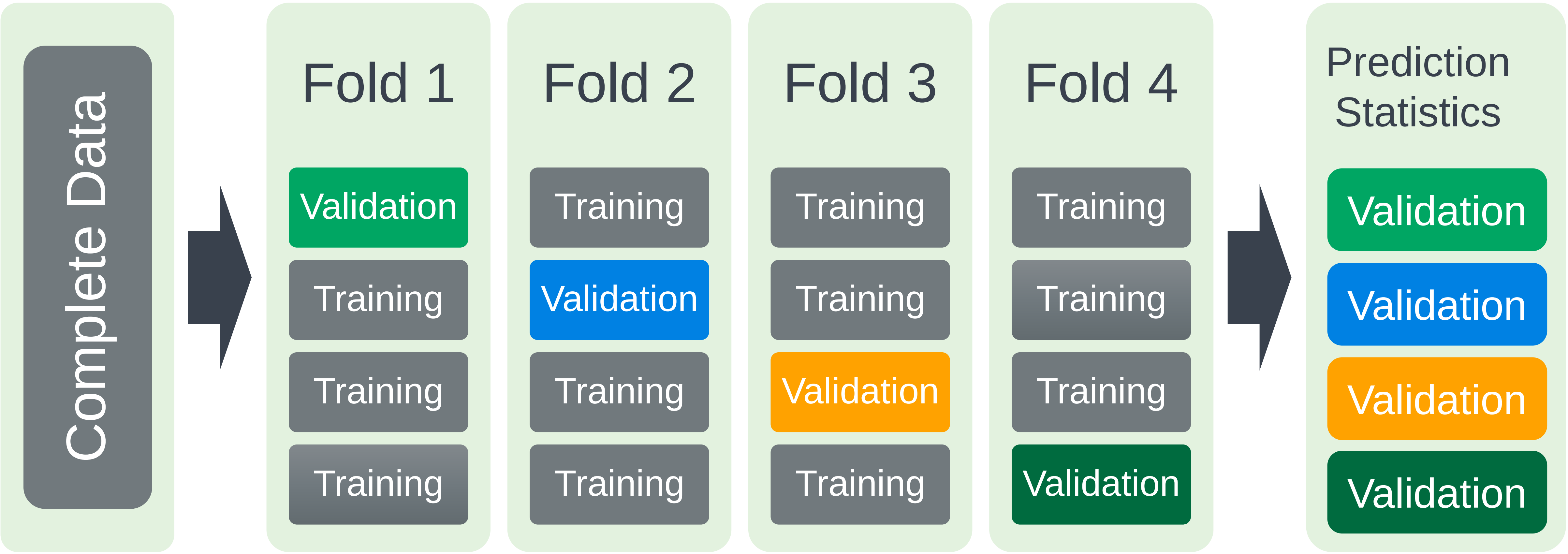


Milliman

The testing data

- Totally unseen during calibration
  - A subset of the training data carved out
  - From a completely different source
  - From a different time period
- Consider biased subsets
  - Age, gender, condition cohort
  - Cost or risk score strata

Cross validation



Testing model performance

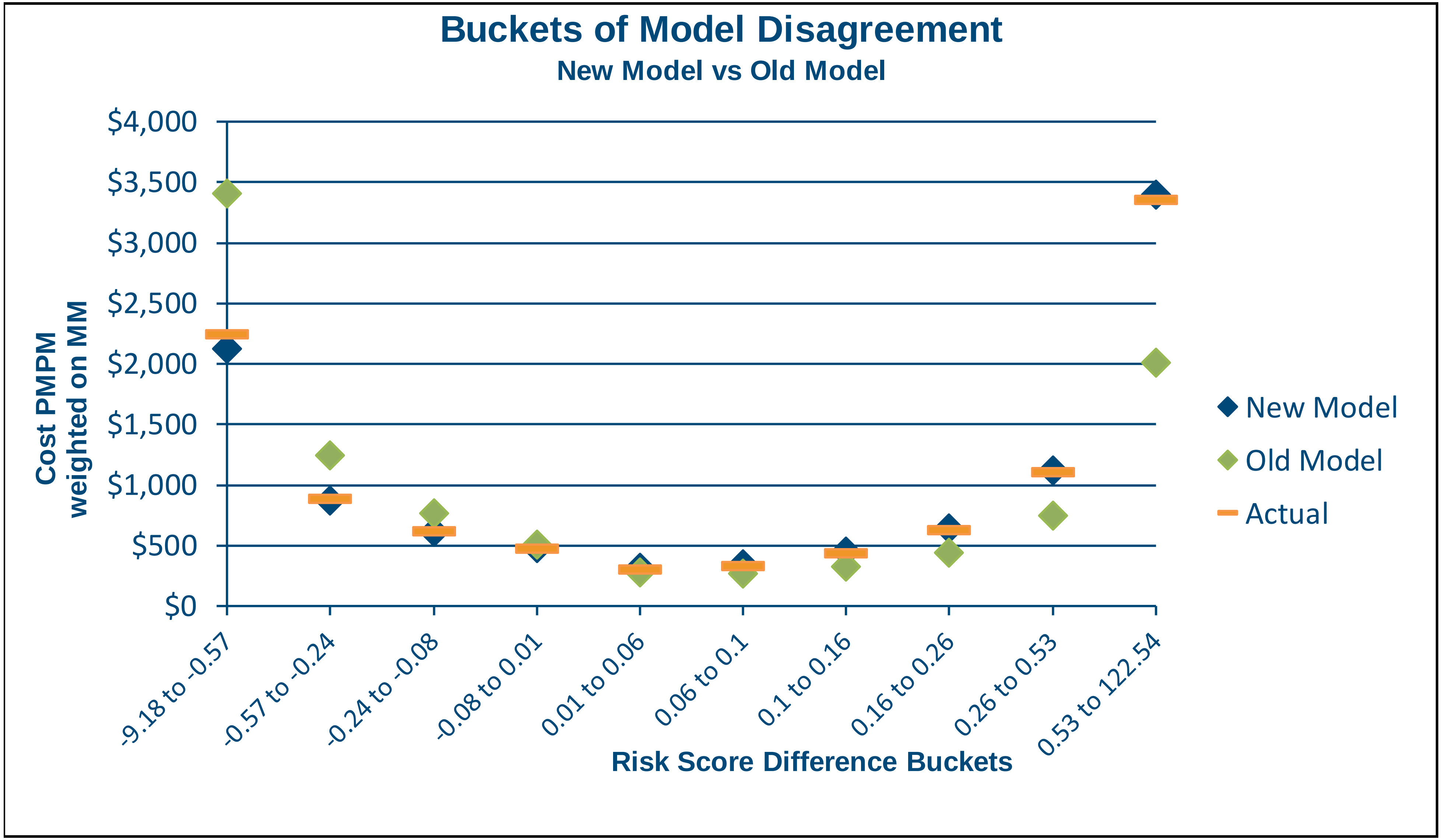
Milliman

What metrics should you use?

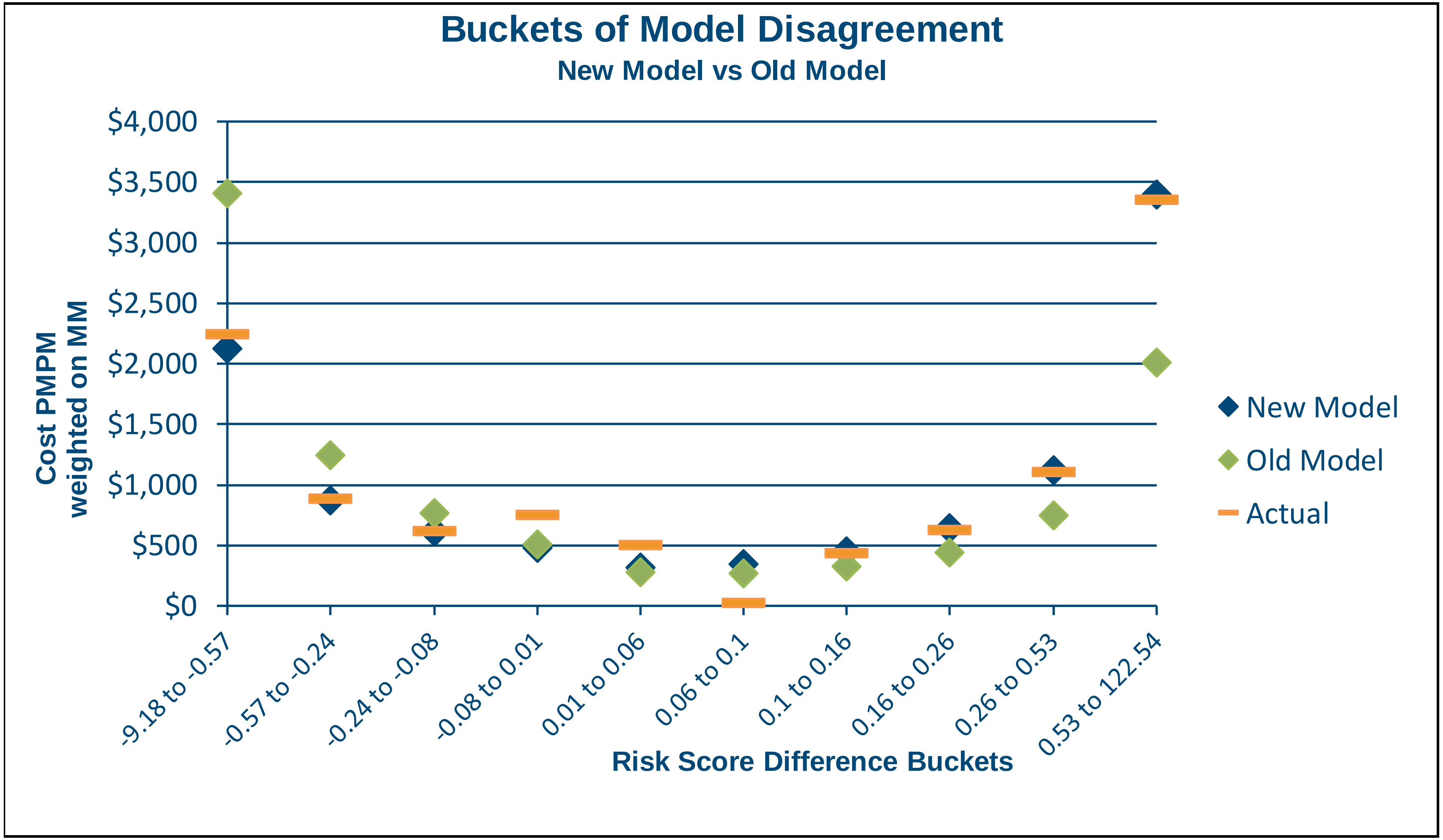
Depends on what you're trying to do…

Milliman

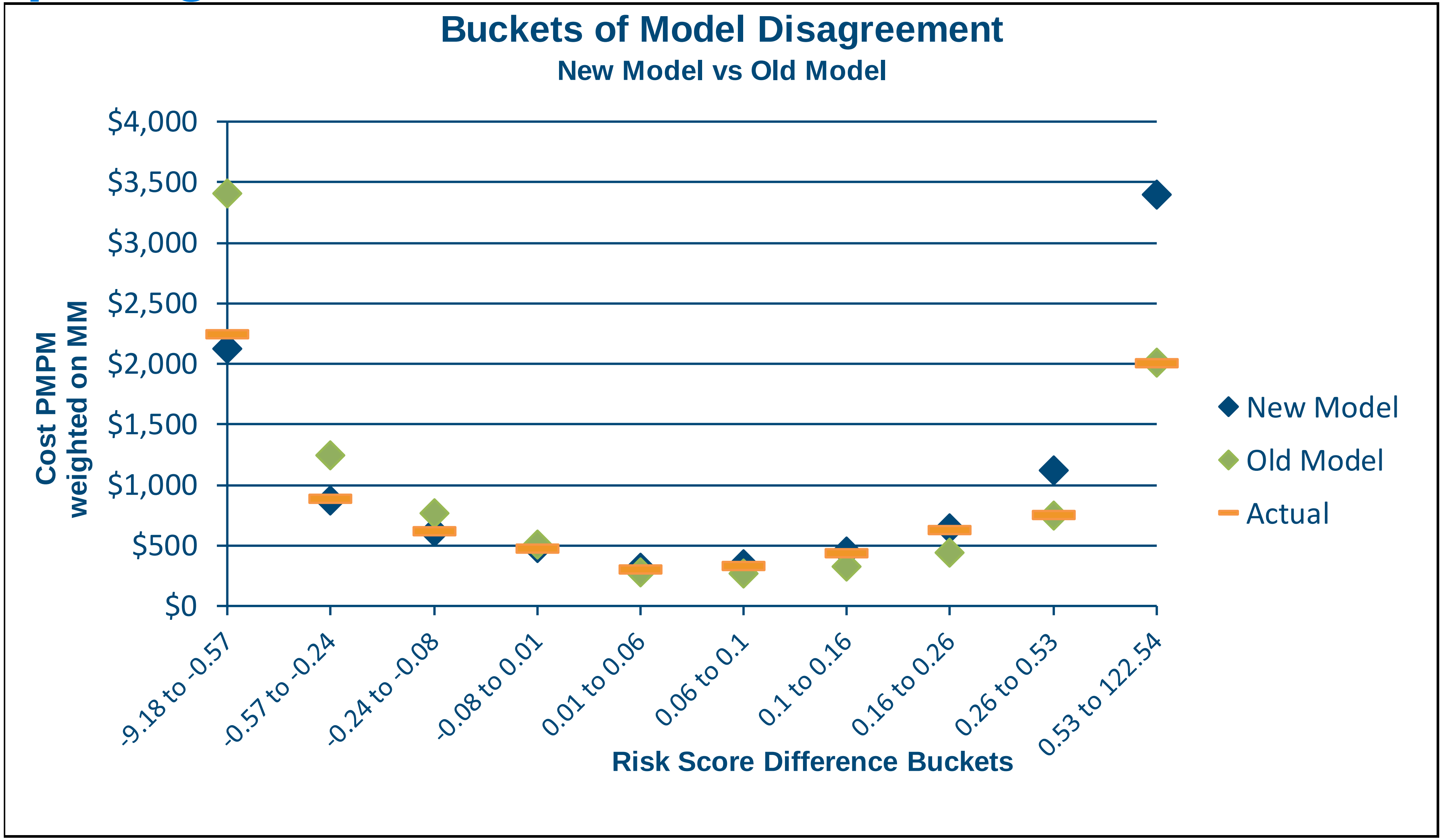
Comparing two models



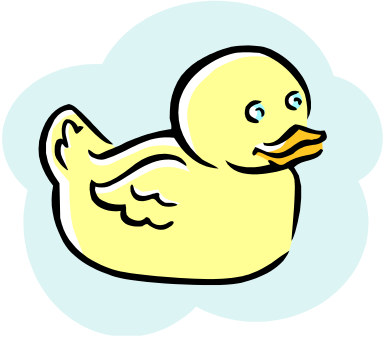
Comparing two models

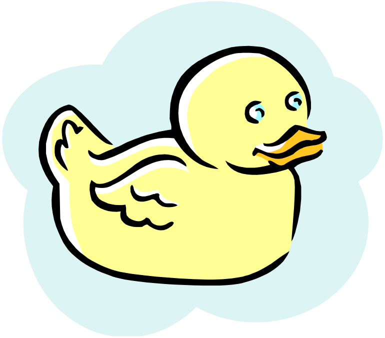


Comparing two models

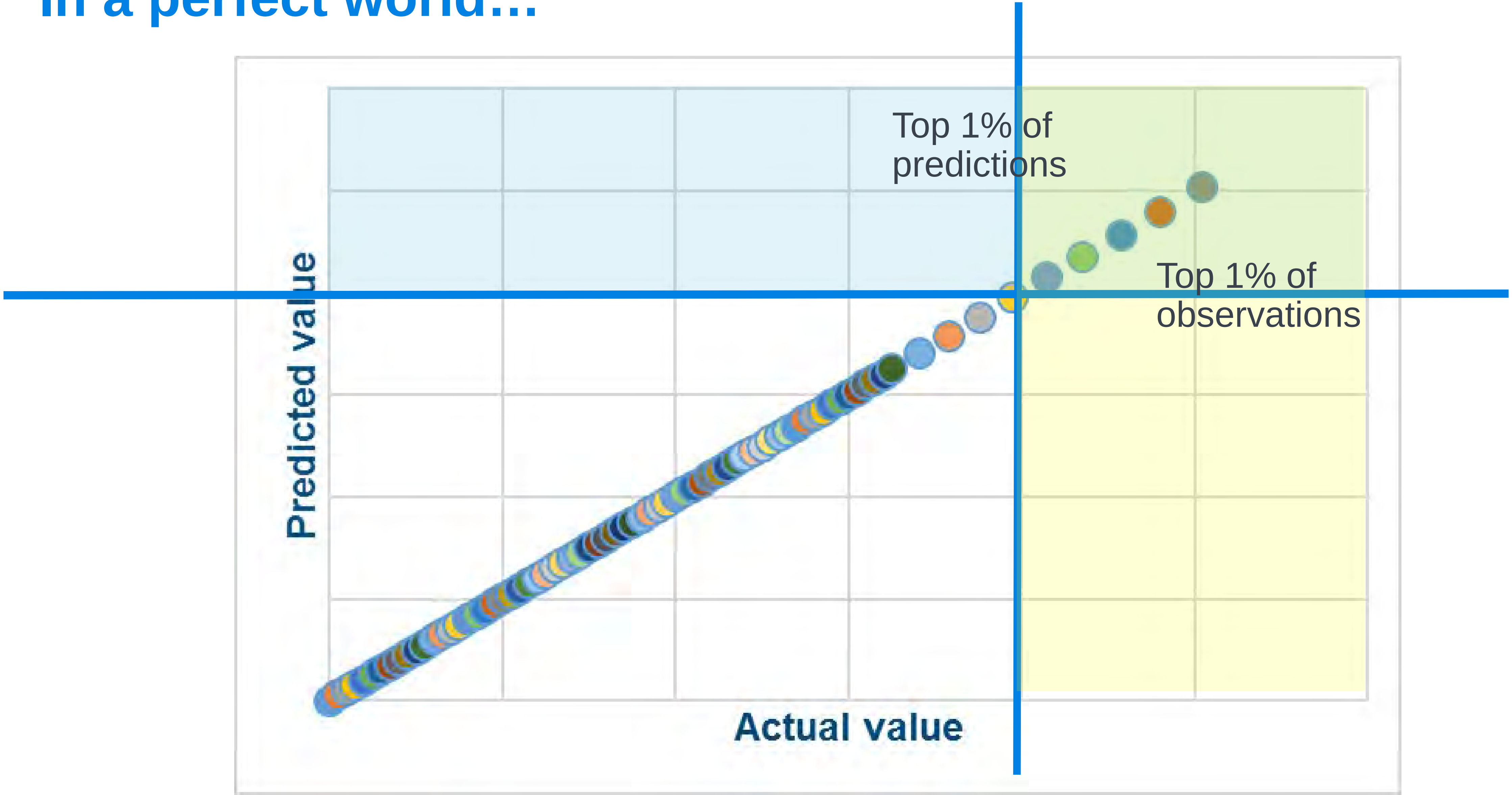


Classification metrics

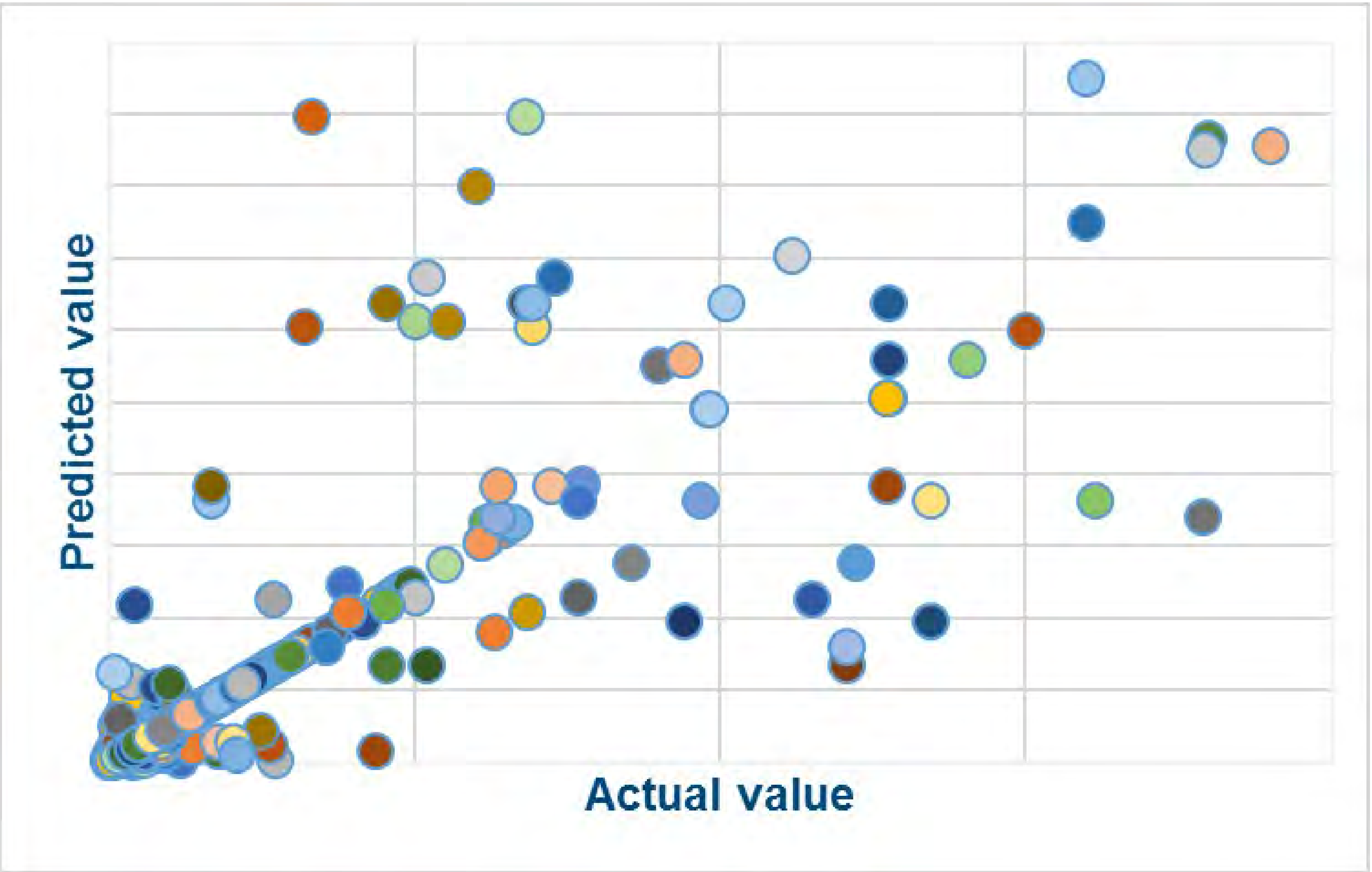




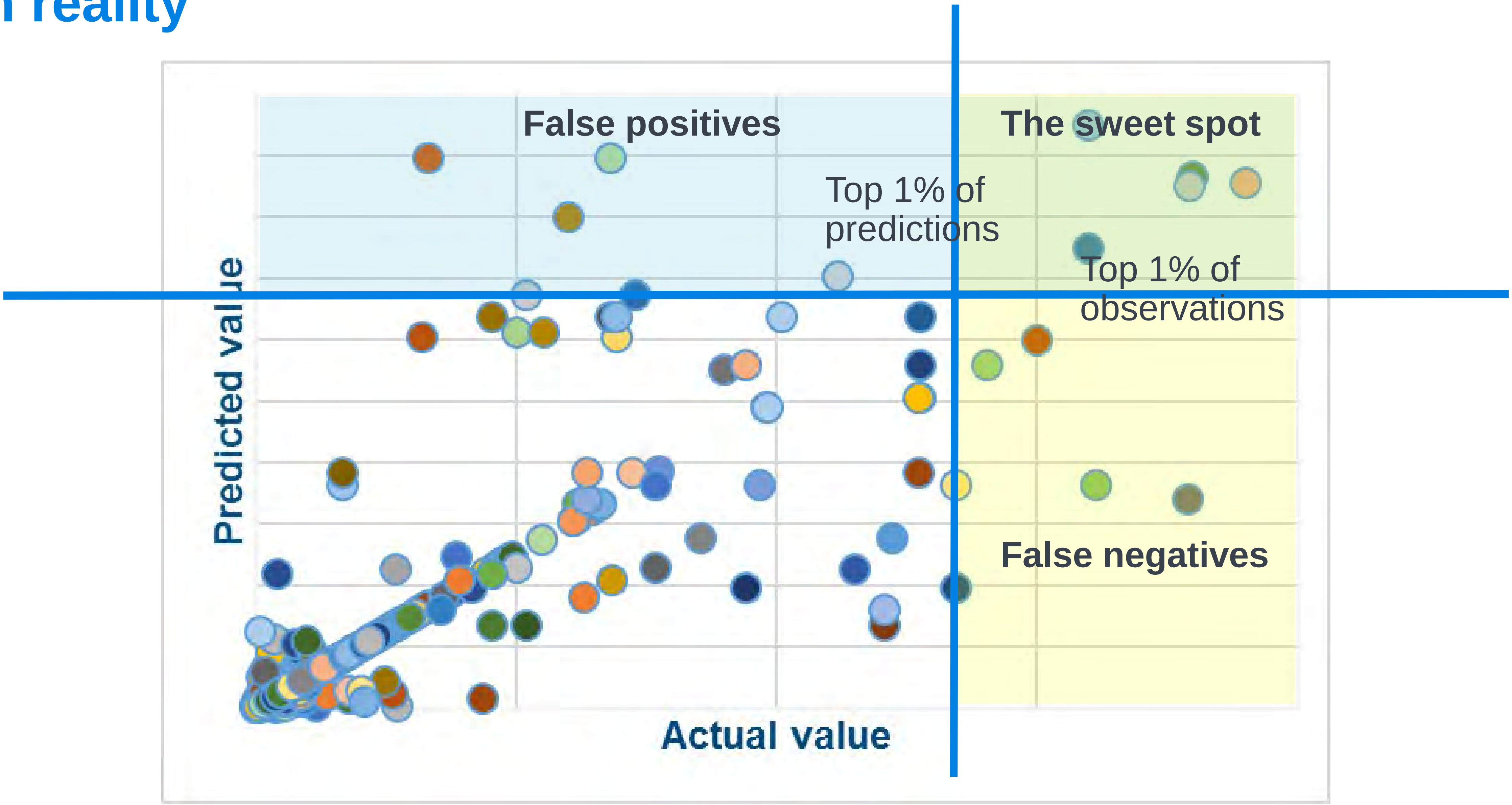
In a perfect world…



In reality



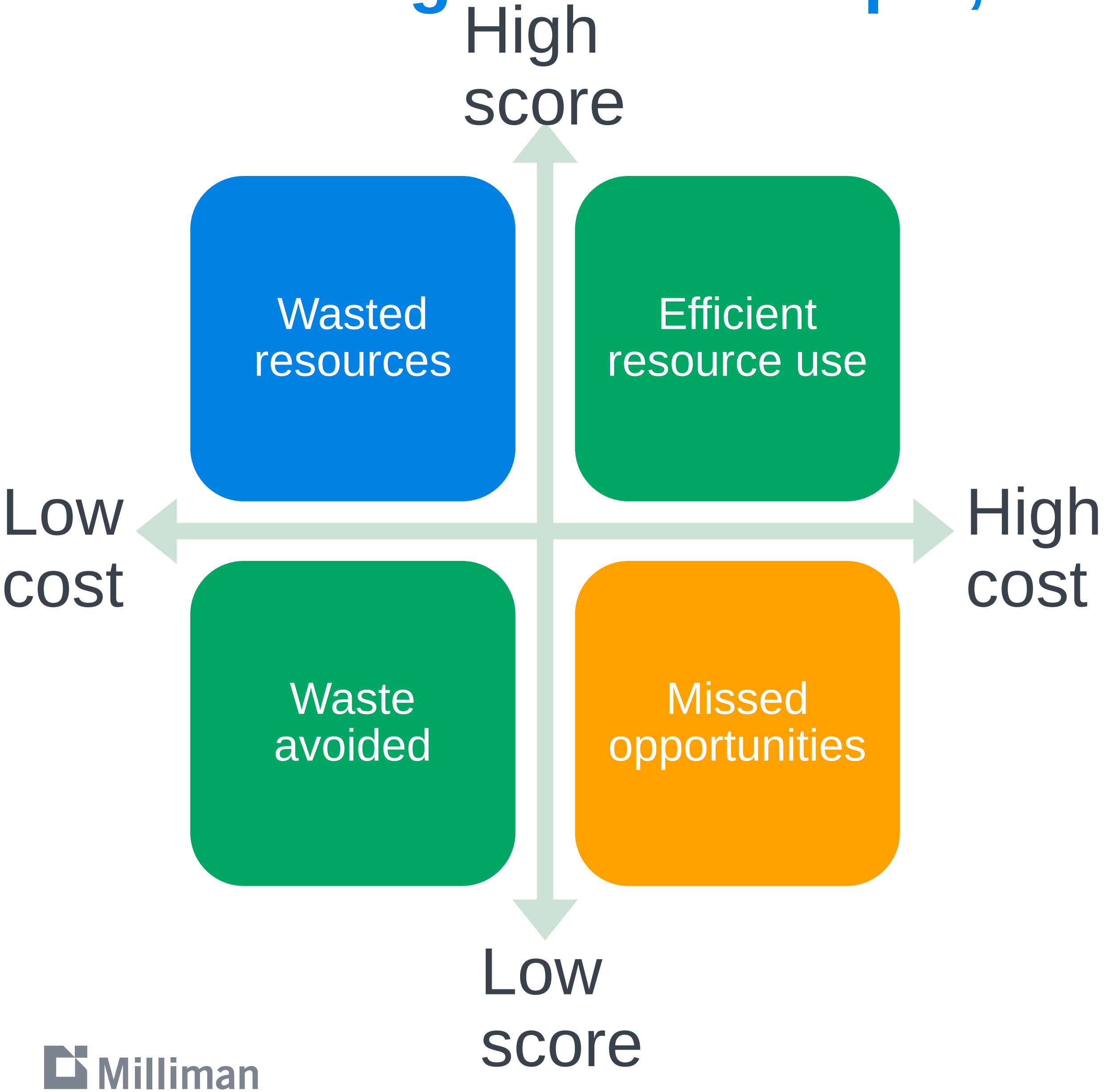
In reality


False positives
The sweet spot
Top 1% of
predictions
Top 1% of
observations
Predicted value
False negatives
Actual value

Example: care management interventions

- Target: individuals in the top 5% of costs
- Method: Use top 5% of risk scores
- Questions:
  - What percentage of high cost members am I identifying?
  - What percentage of members am I targeting that aren’t high cost?
  - Is there a better risk score threshold to use?

Care management example, cont’d



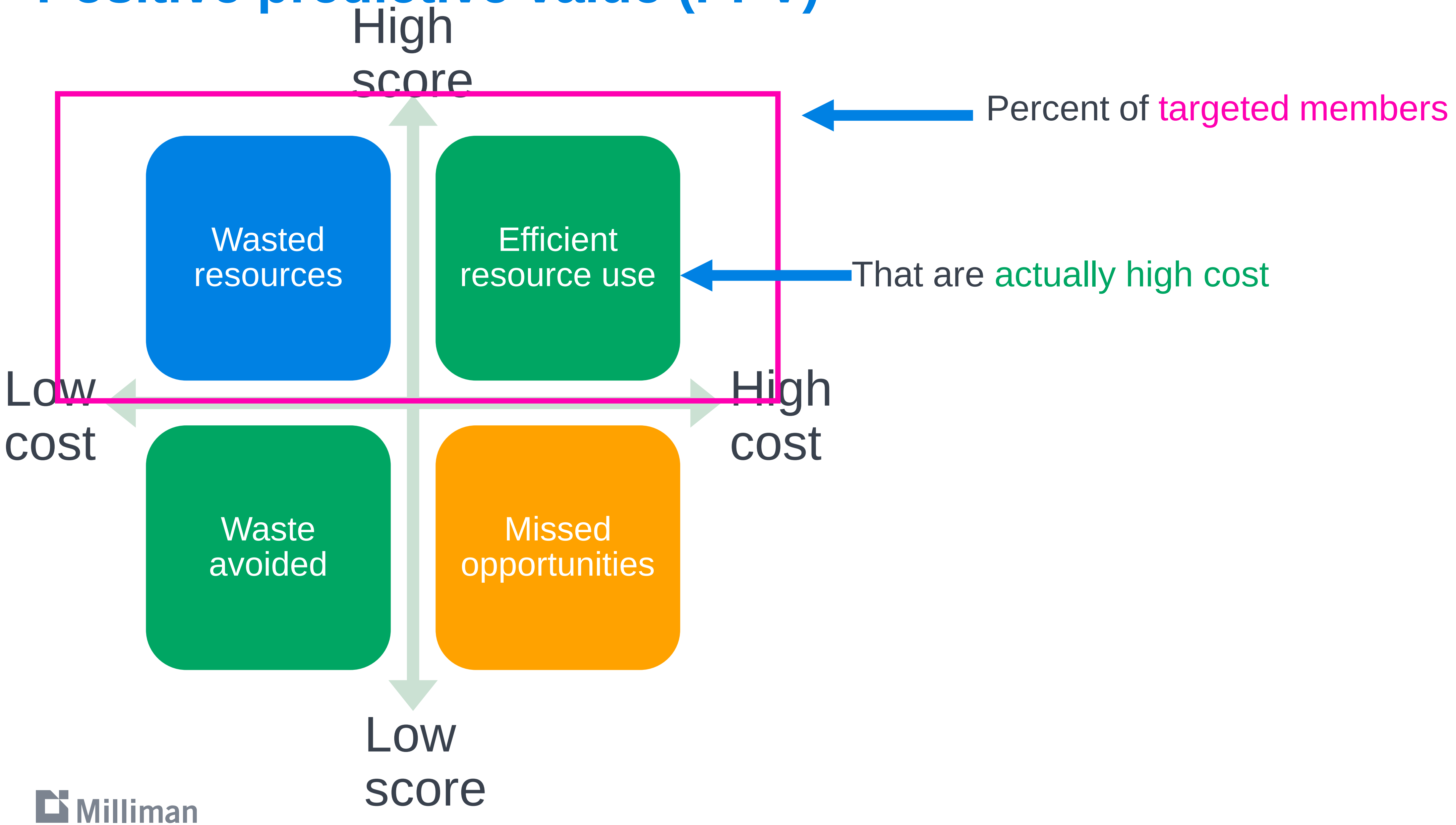
Questions to ask

- What percentage of the time am I in the right quadrants?
  - PPV, NPV, Sensitivity, Specificity

- Where should I draw the line to maximize the time in the right quadrants?
  - ROC, AUC

Milliman

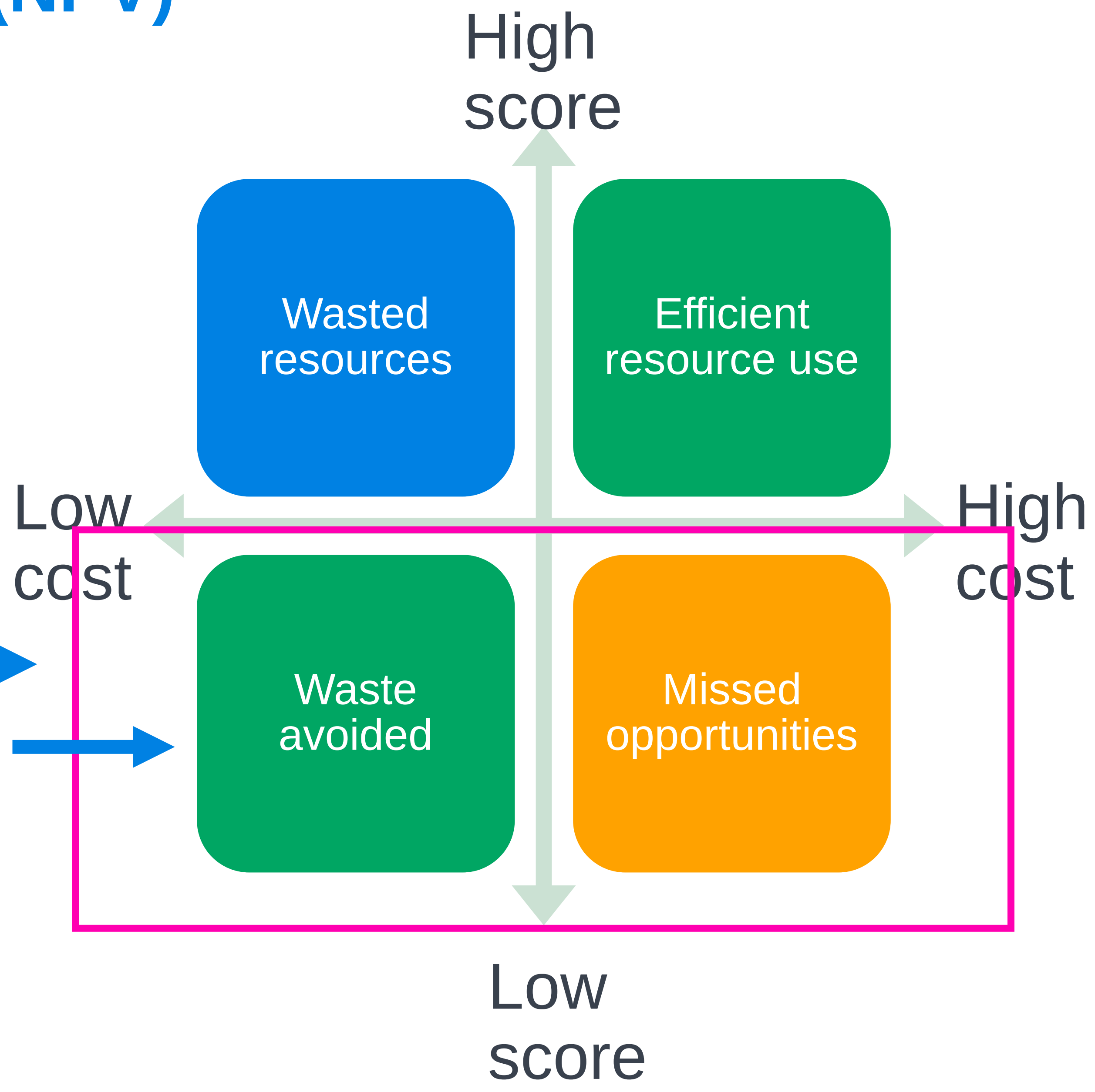
Positive predictive value (PPV)



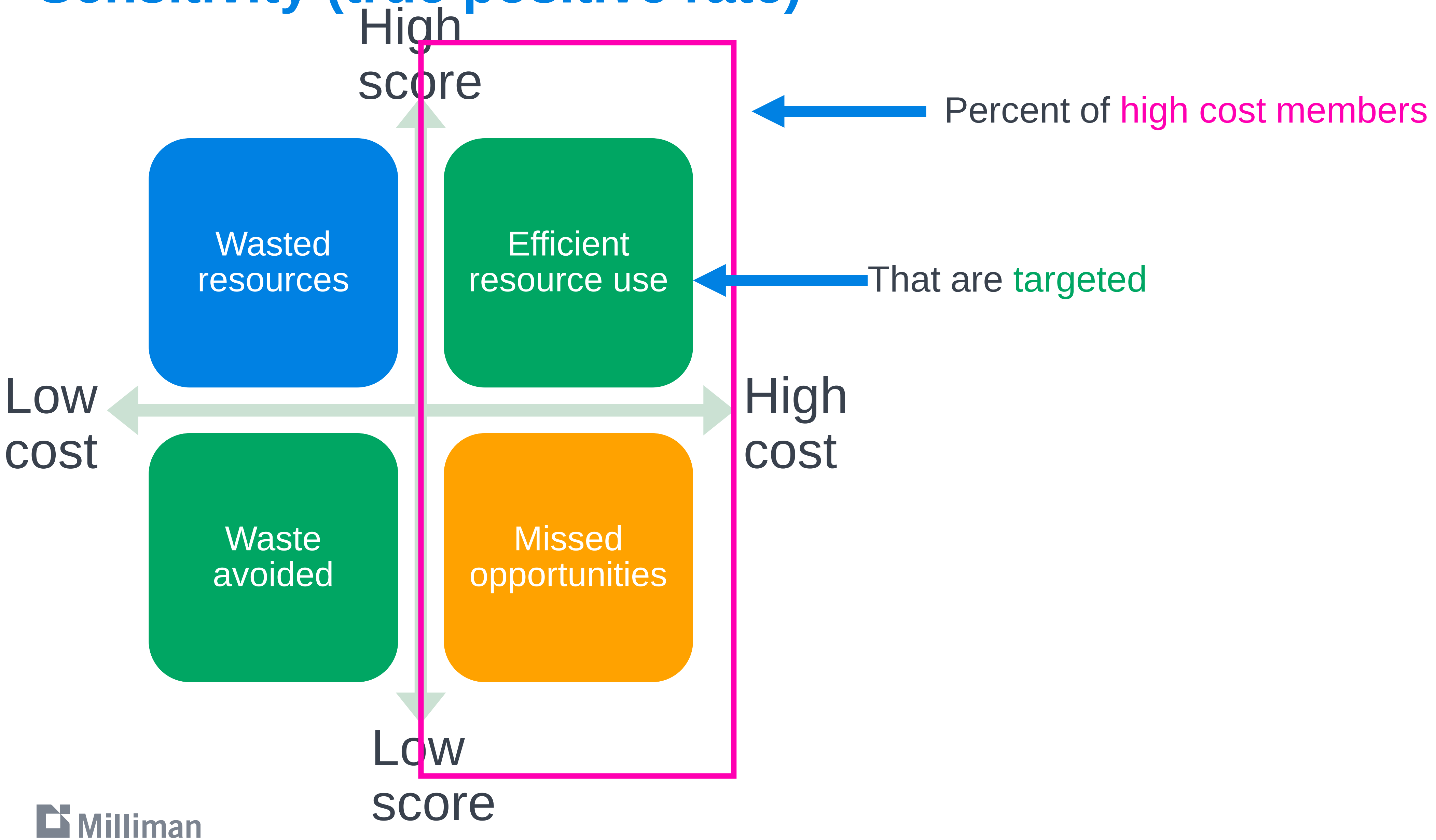
Negative predictive value (NPV)

Percent of untargeted members

That are actually low cost



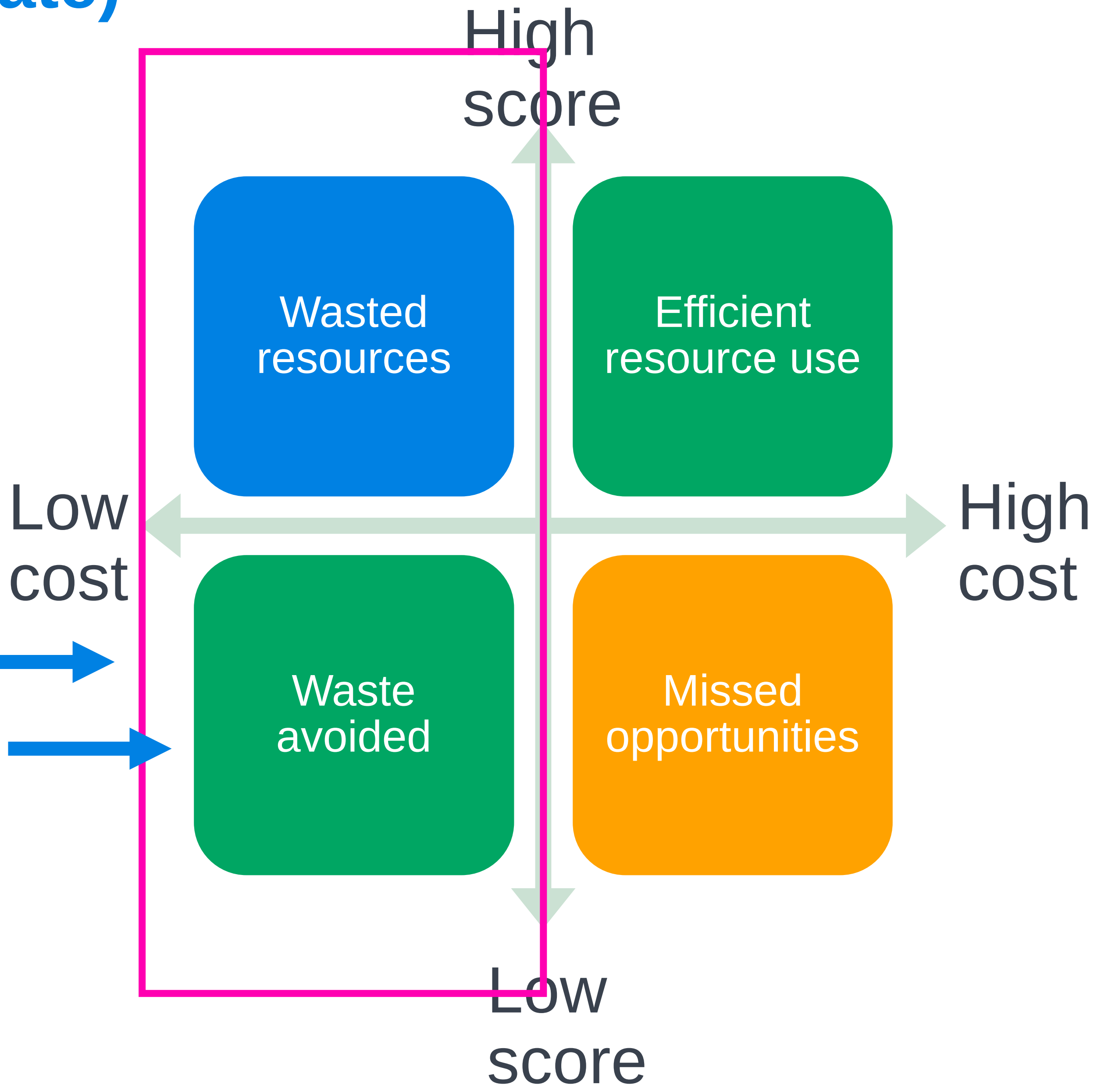
Sensitivity (true positive rate)



Specificity (true negative rate)

Percent of low cost members

That are not targeted

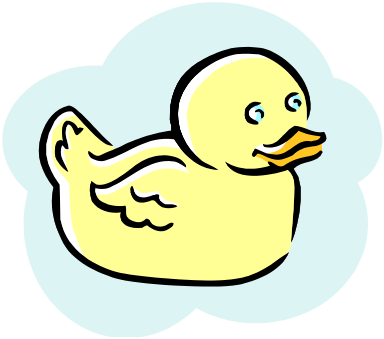


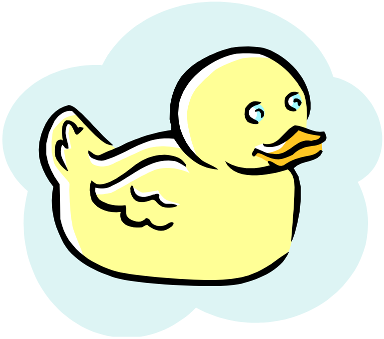
The tradeoff between sensitivity and specificity

Matthews Correlation Coefficient

- Accounts for true/false positives/negatives
- Classes can have very different sizes
- Acts like a correlation coefficient between observed and predicted classifications
  - +1 means perfect classification
  - 0 equivalent to coin flip
  - -1 means perfect disagreement

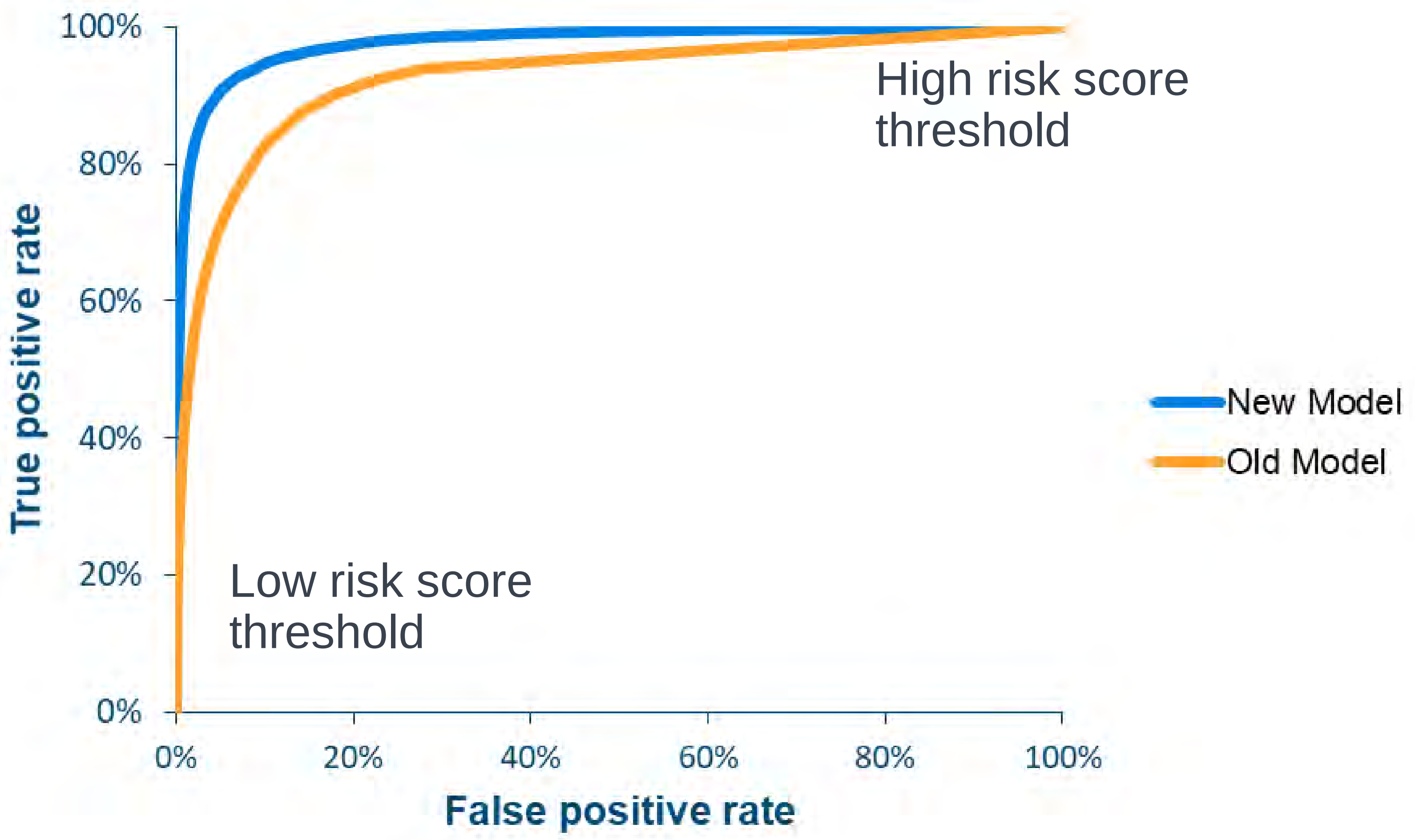
$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$





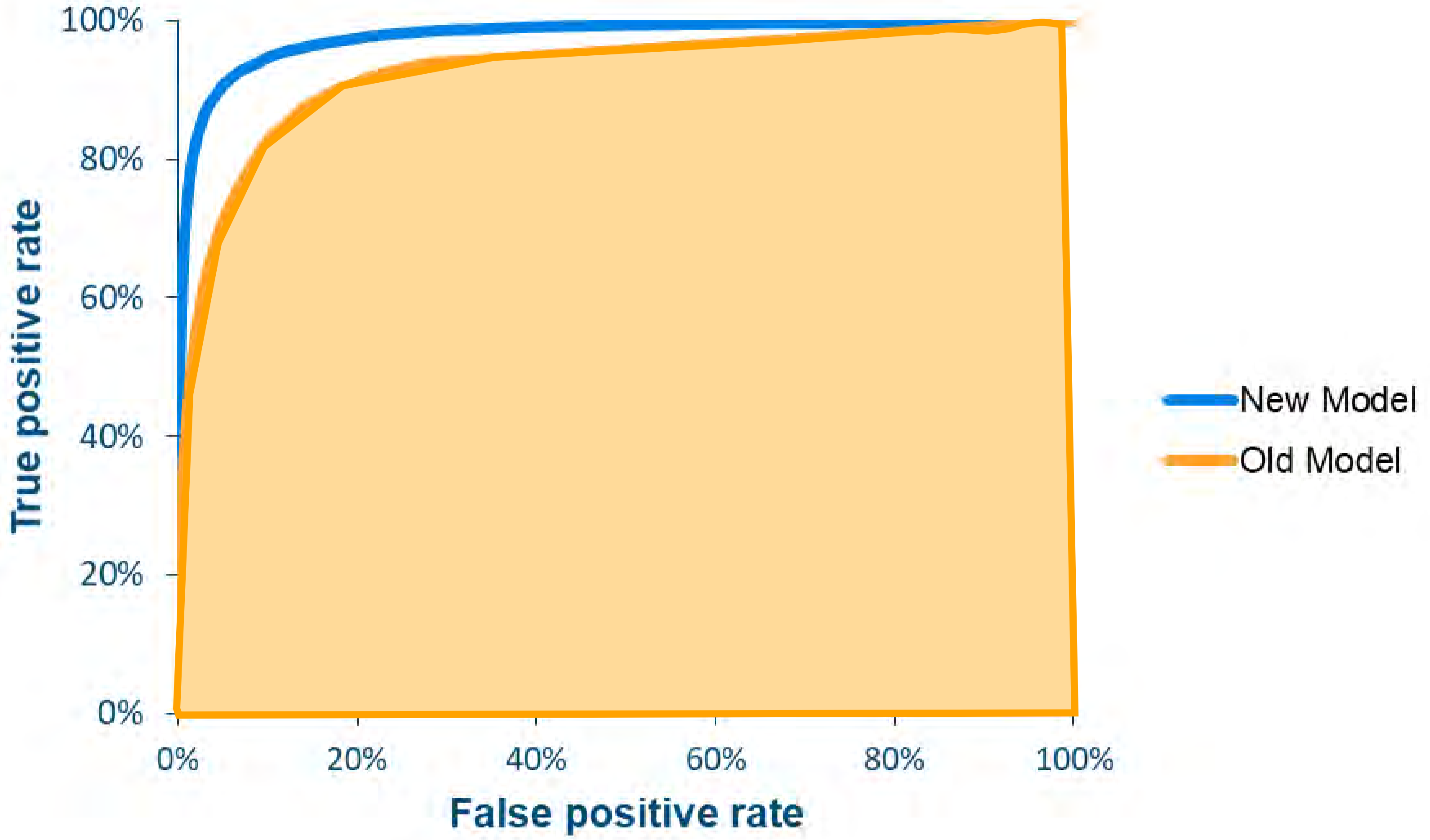
Receiver operating characteristic (ROC) curves

Predicting the top 1% of high cost members



Area under the curve

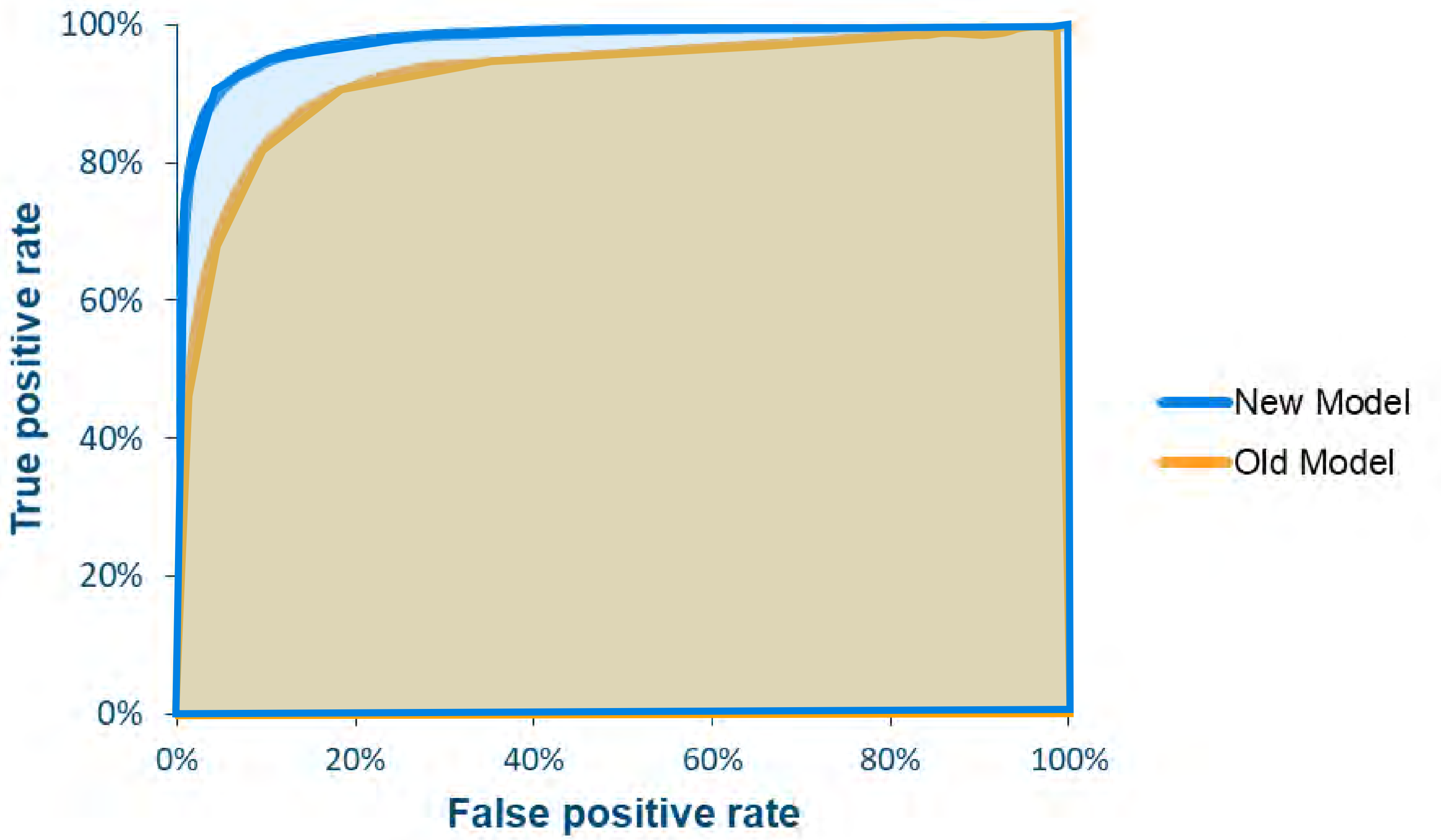
Predicting the top 1% of high cost members



Milliman

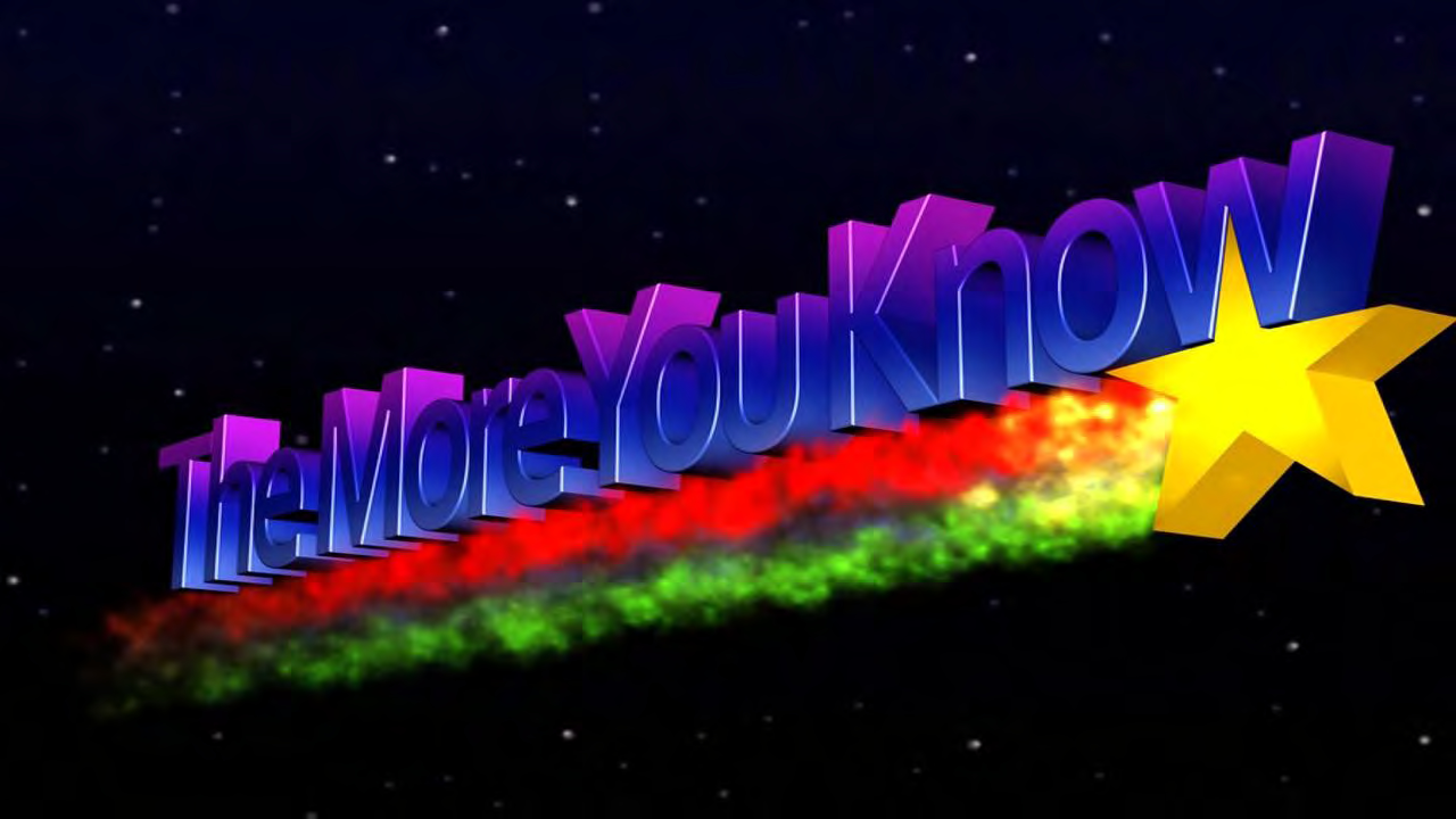
Area under the curve

Predicting the top 1% of high cost members

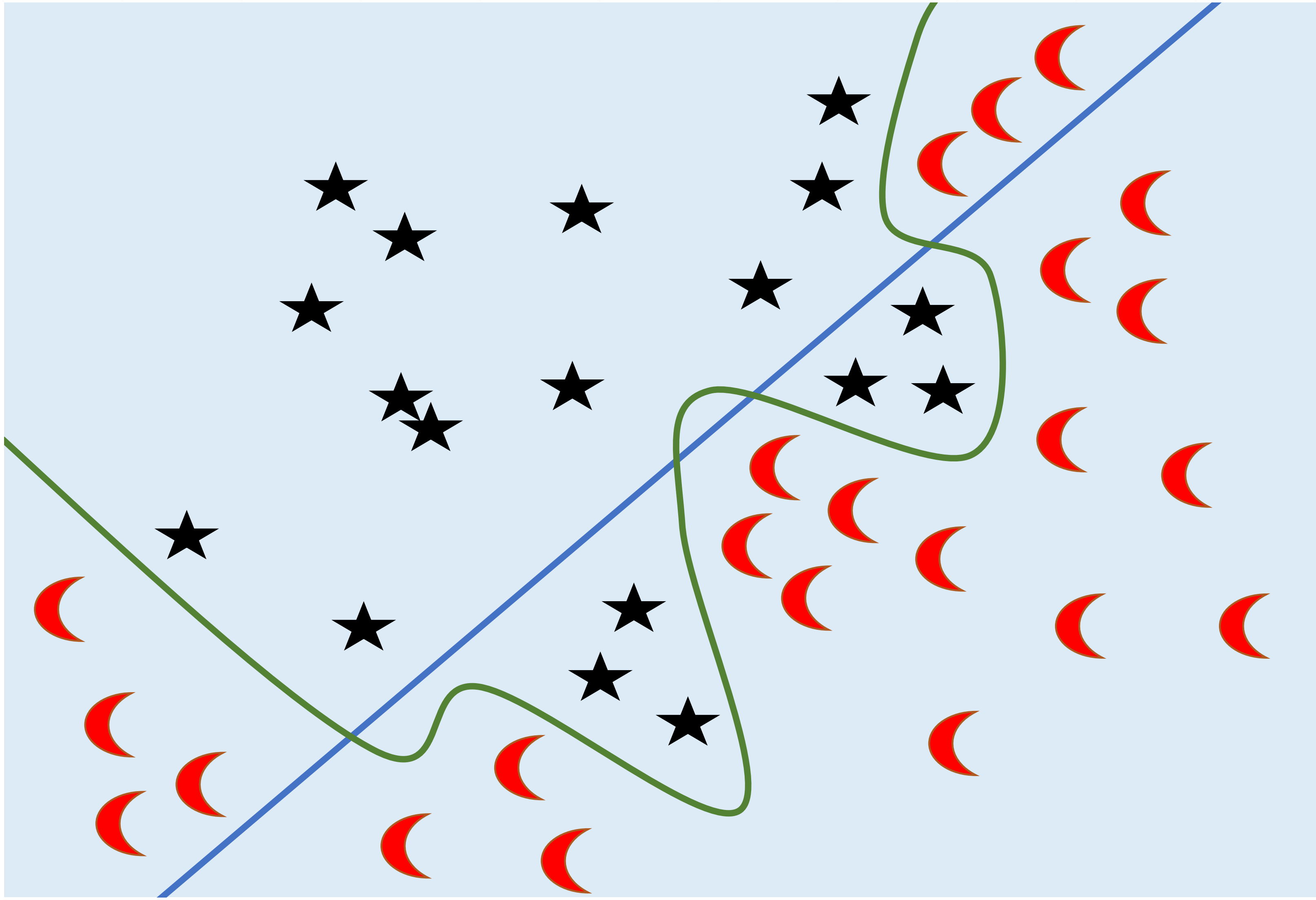


Going Deeper

Milliman


The More You Know

Interpretability vs. Performance



Choices, Choices, Choices

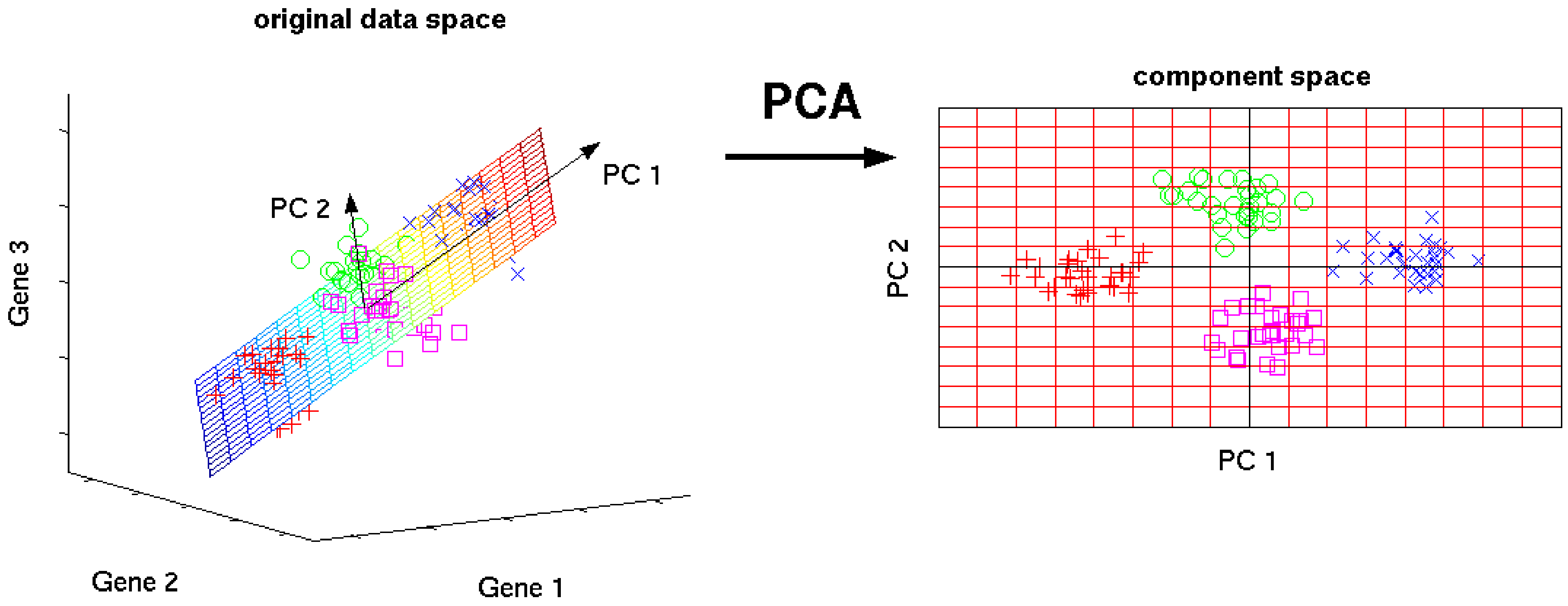
Interpretation Tools

- Dimension Reduction Viz
- Sensitivity Analysis
- Feature Importance
- Partial Dependence Plots
- ICE Plots
- Lorenz Curves and Gini
- Surrogate Models
- Shapley Predictions
- Local interpretable model explanation (LIME)

Gradient Boosting

- XGBFI
- Monotonicity constraints

Visualization



http://www.nlpca.org/pca_principal_component_analysis.html

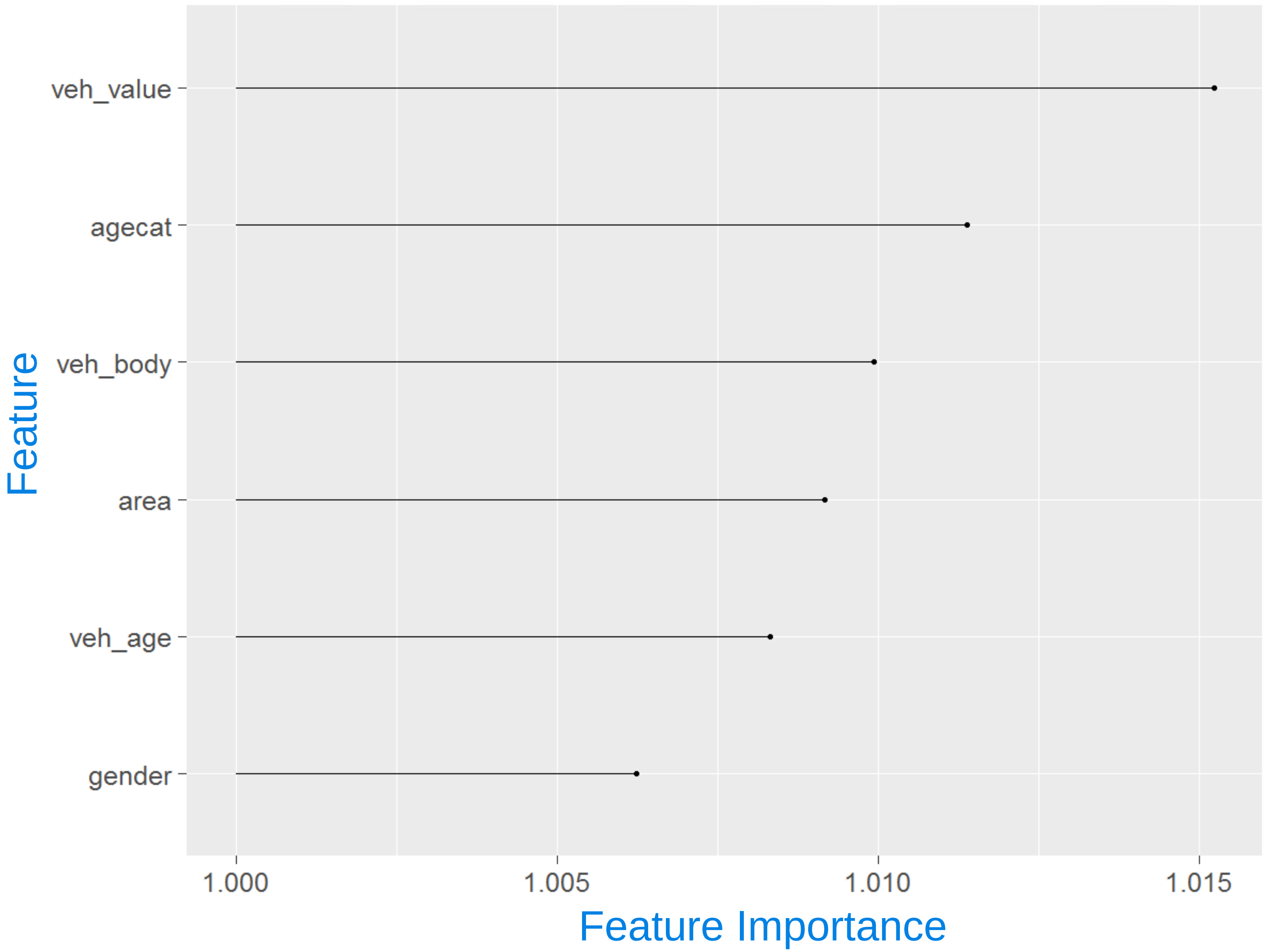
Sensitivity Analysis

- Thoroughly test the model for changes based upon small permutations in features
- Use simulated data representing prototypes for different areas of interest

Feature Importance

- Measures how much a feature contributes to the predictive performance of the model
- Helps us know what is drives predictions at a global level
- Common methods
  - Permute a feature and measure change in model error
  - LOCO – Leave One Covariate Out - Build model with and without feature and compare difference in error

Feature Importance - Visualized



PDP and ICE Plots

Partial Dependence Plot (PDP)

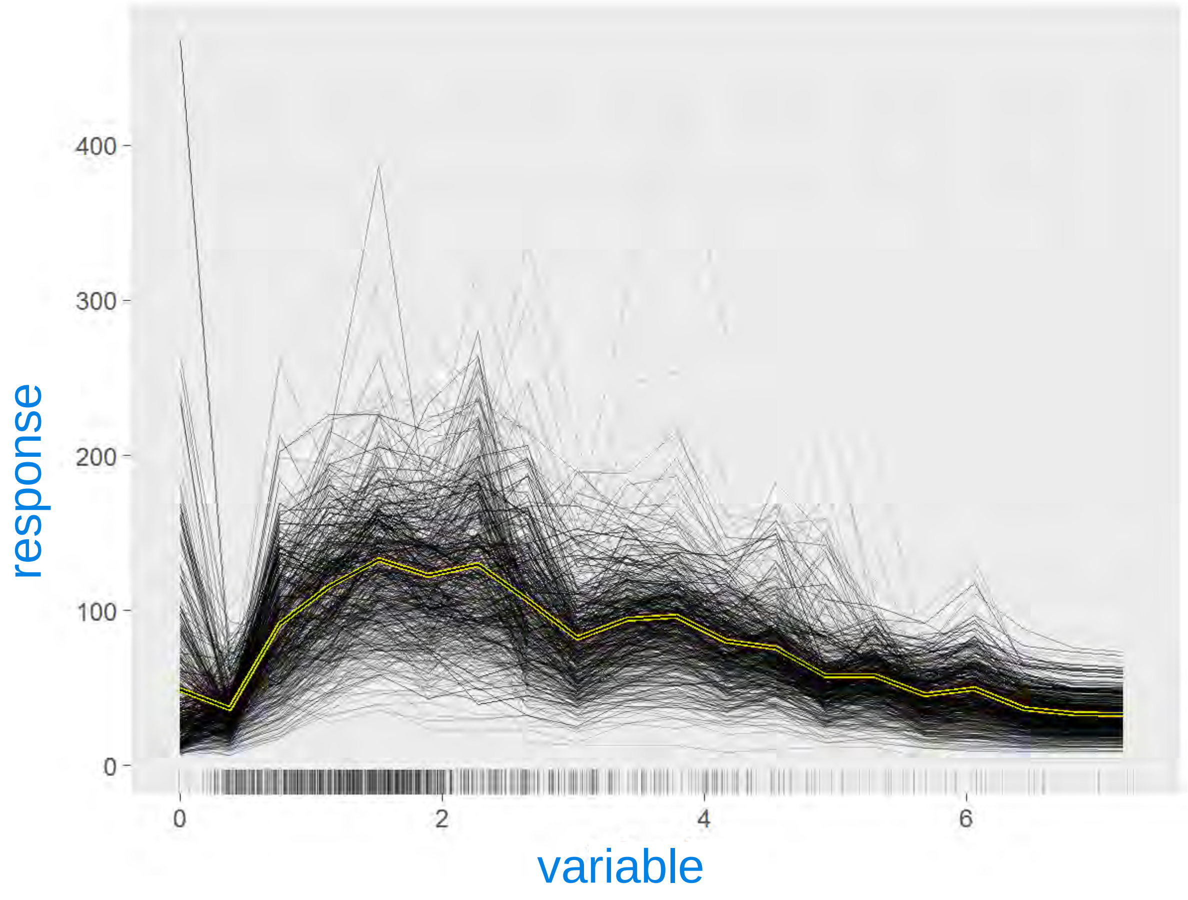
- Displays the marginal impact of a feature on the model – what's happening with "all else equal"
- Shows the relationship between the target and the feature on average
  - Fix the relationship of 1 or 2 predictors at multiple values of interest
  - Average over the other variables
  - Plot response

Individual Conditional Expectation (ICE)

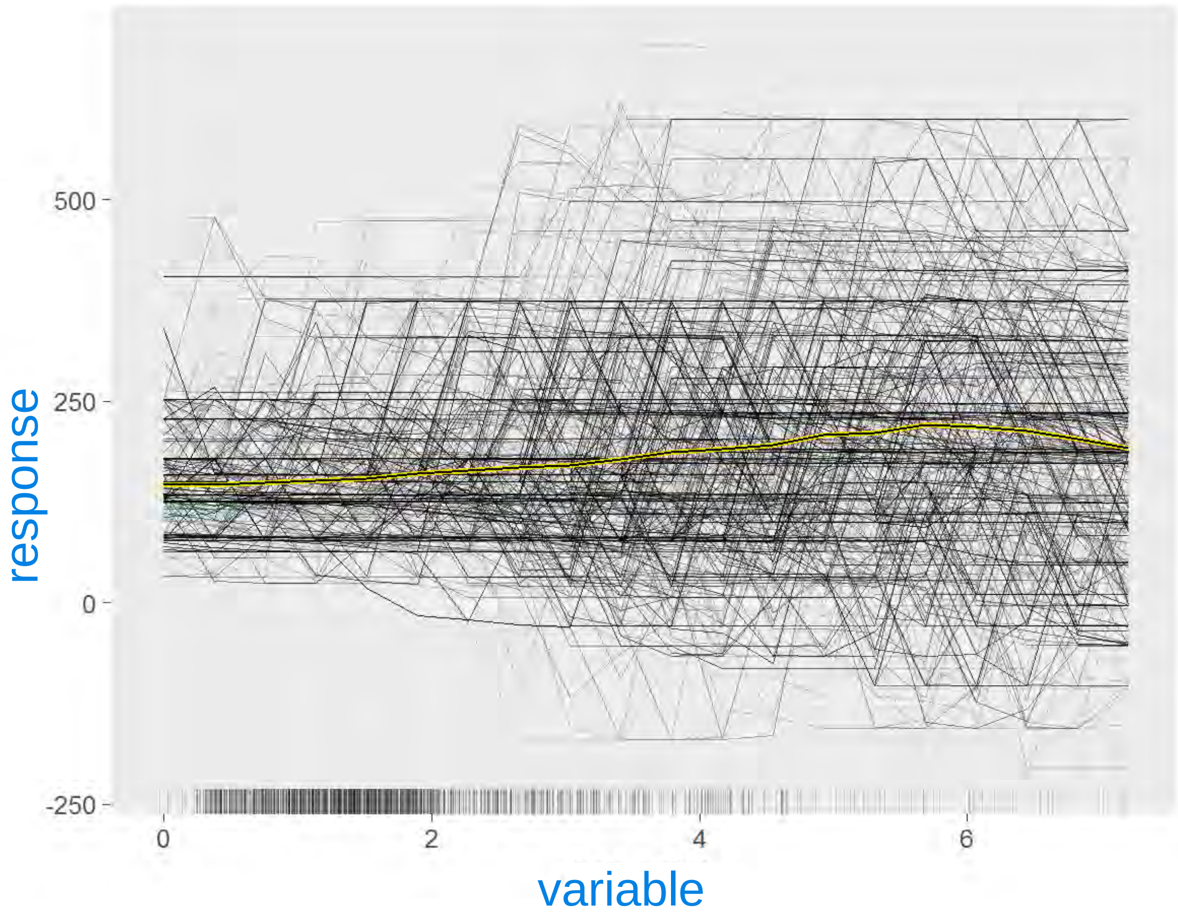
- Shows how a single prediction changes when the value of a single feature is varied
- Run this for multiple predictions and plot results

PDP and ICE Plots - Visualized

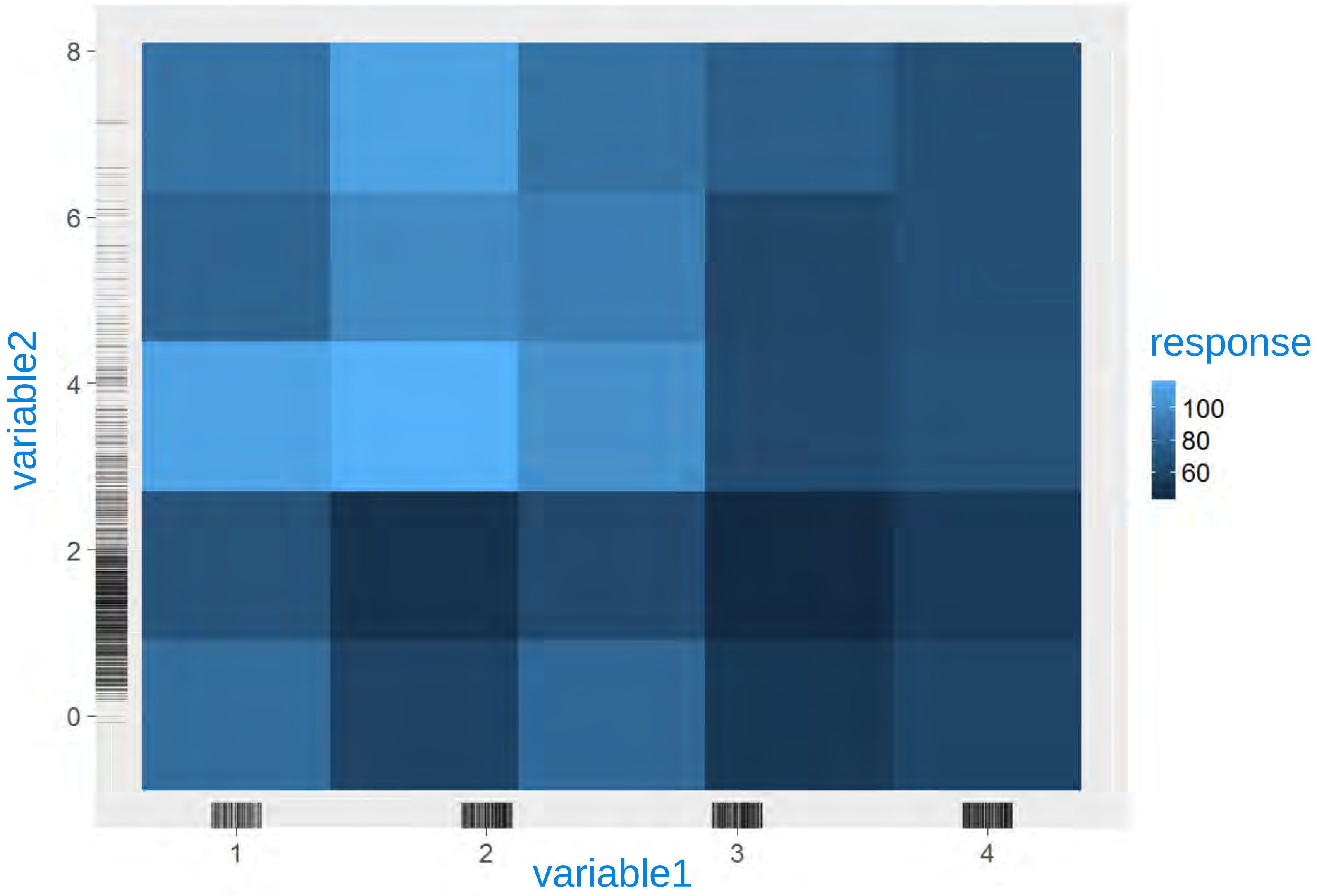
XGBoost



Neural Network



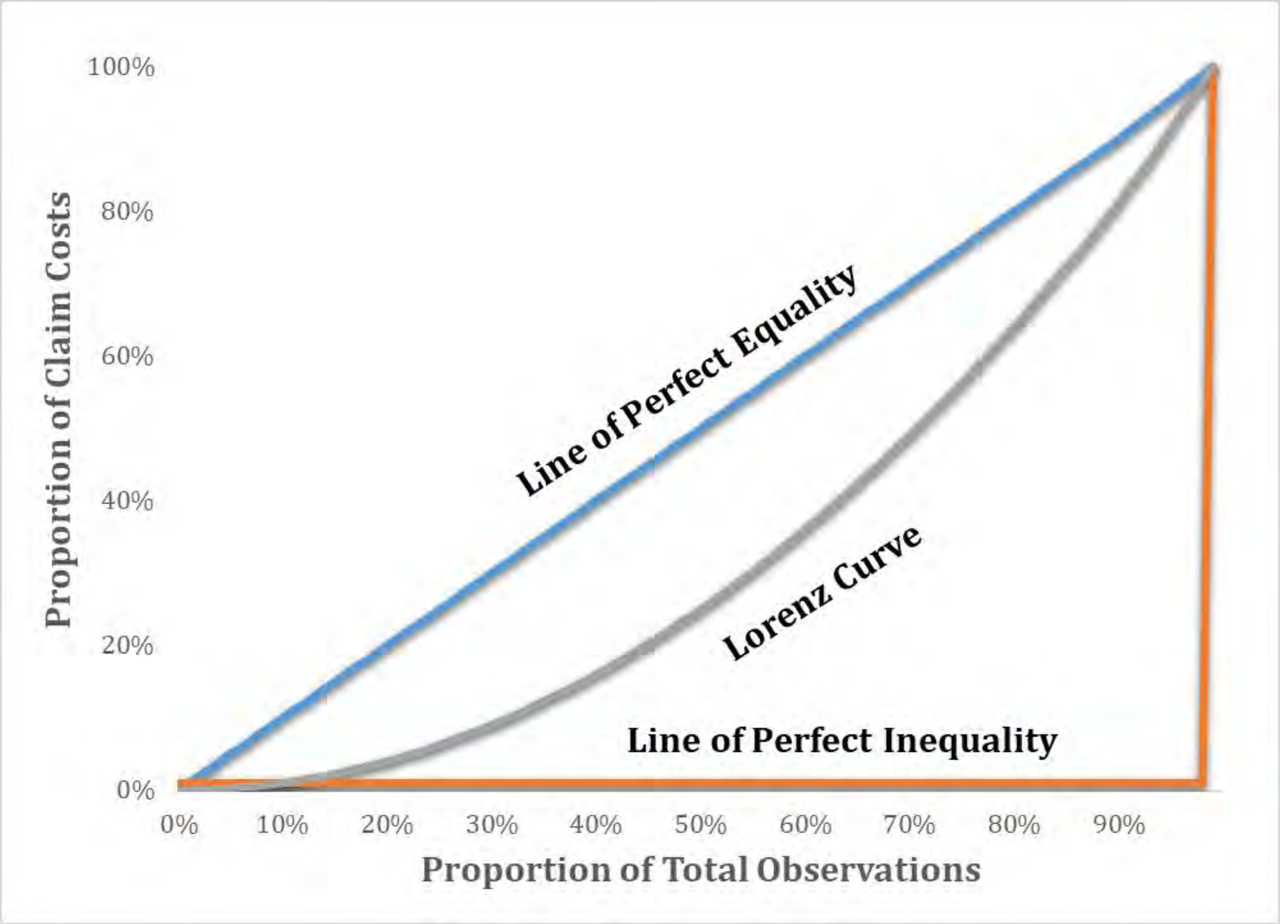
PDP with 2 features - Visualized



Milliman

Ordered Lorenz Curves and Gini Gain

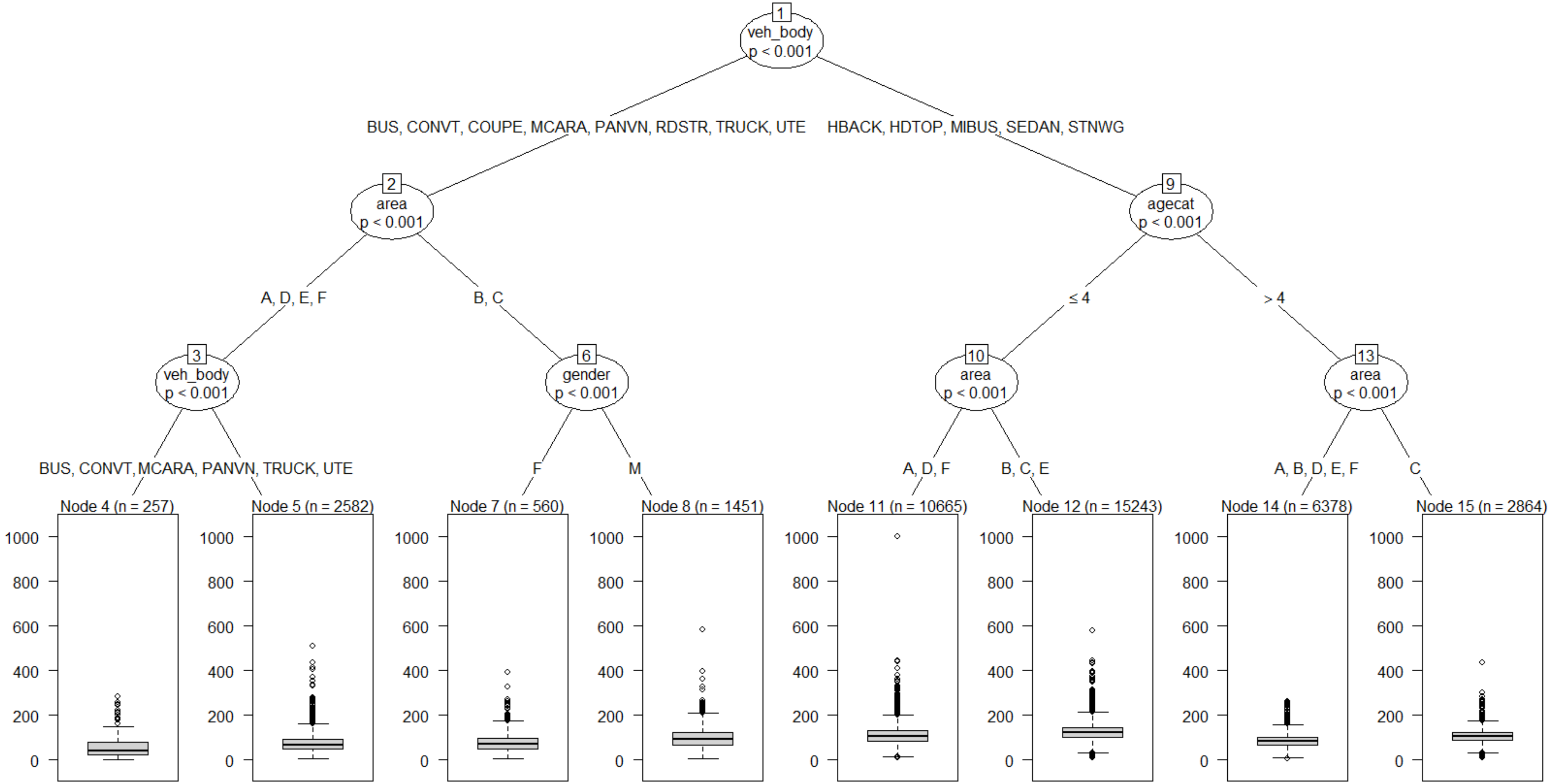
- Ordered Lorenz Curves are useful measures of model stratification
- Gini Gain lets us summarize the lift in a single statistic
  - Equals the area between the Lorenz Curve and the line of perfect Equality



Surrogate Model

- A model trained using another models predictions as its target
  - Decision tree
  - Linear model
- Result is a simpler model that can help interpret the more complex model

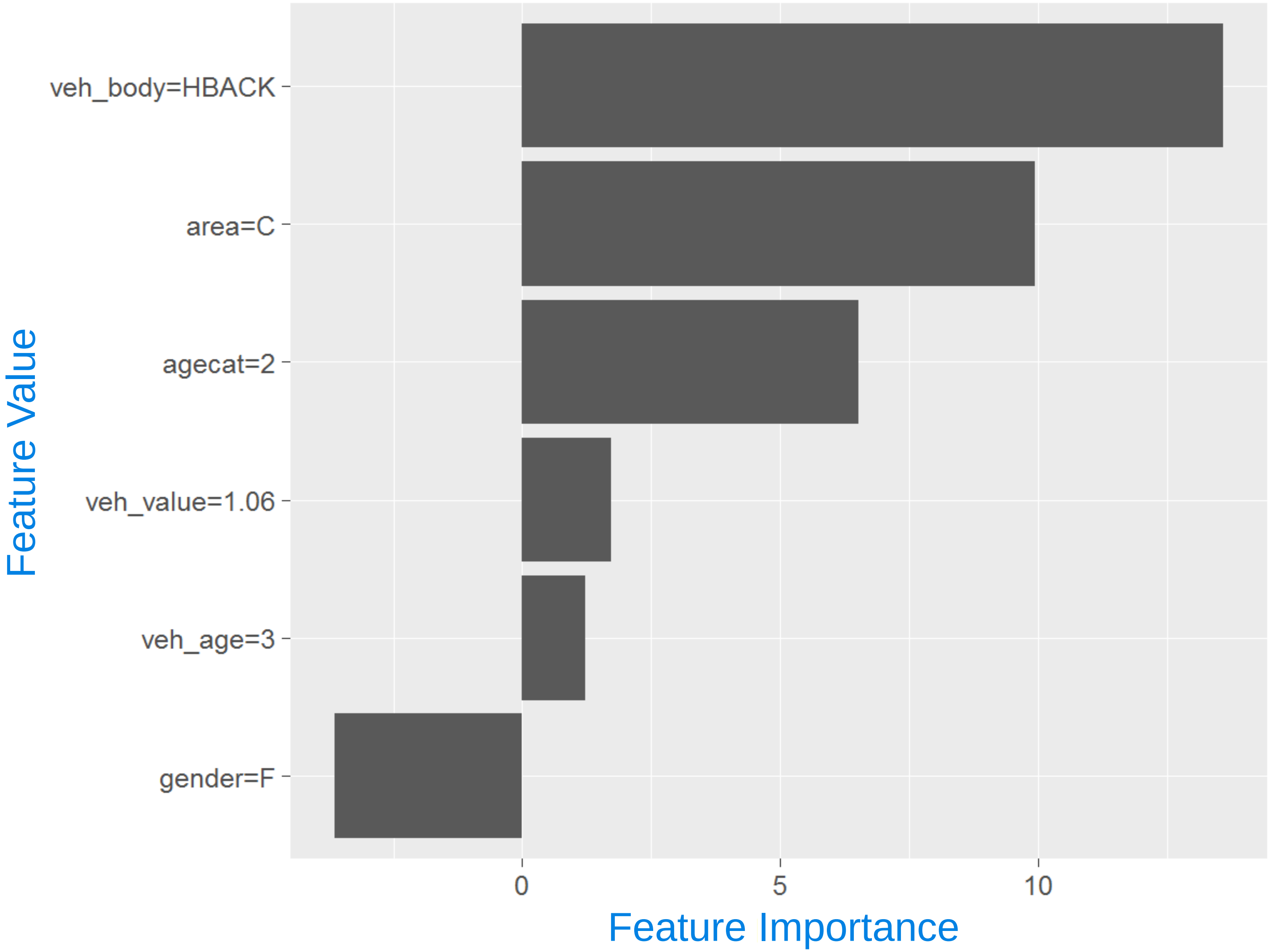
Surrogate Model - Visualized



Shapley Predictions

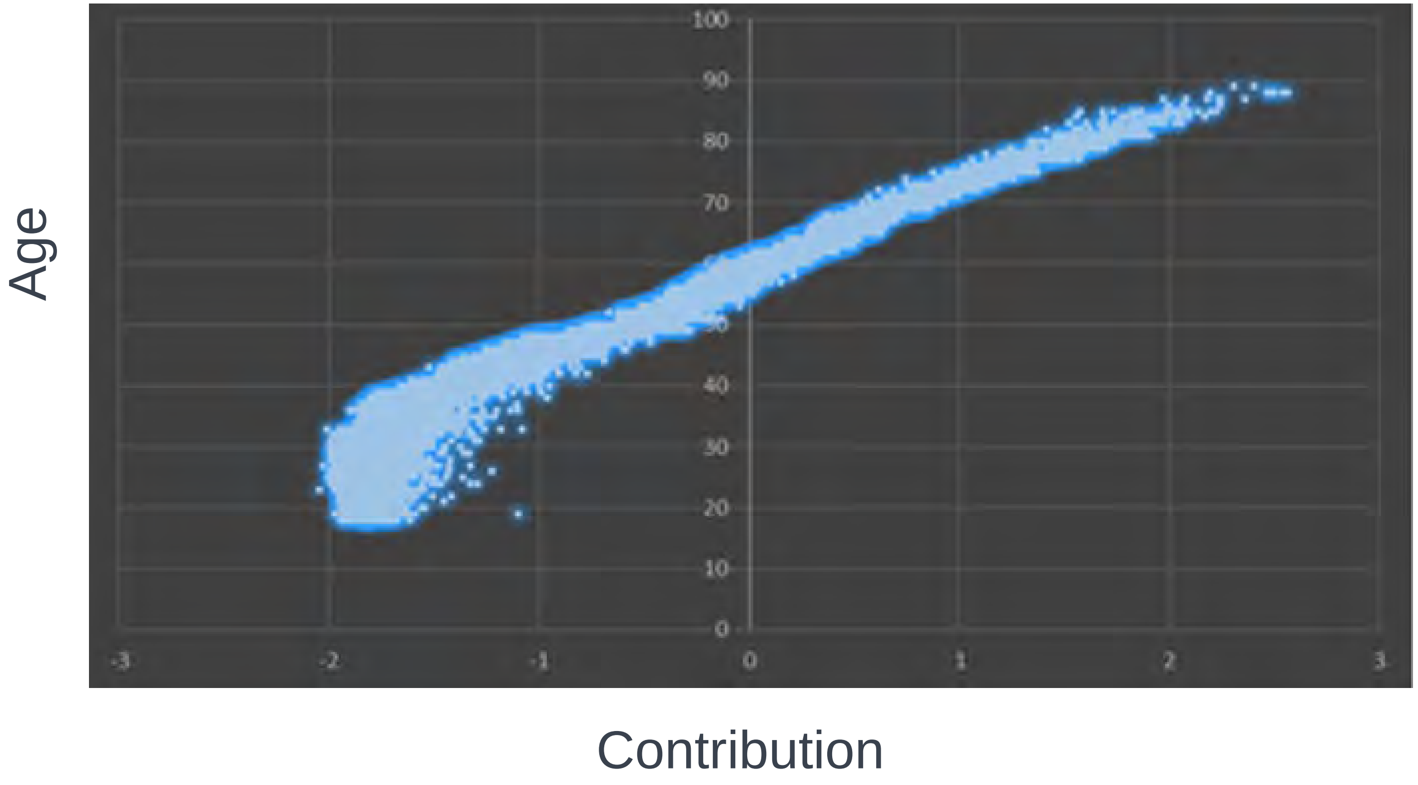
- Provides a measure of local feature contribution for a given prediction
- Basis in game theory
  - Assigns “payout” to players in proportion to marginal contribution
  - “Game” is prediction of an observation

Shapley Visualization



Other Uses of Shapley

- Unsupervised Clustering
- Shapley PDP Plot
- Remove impact of a variable

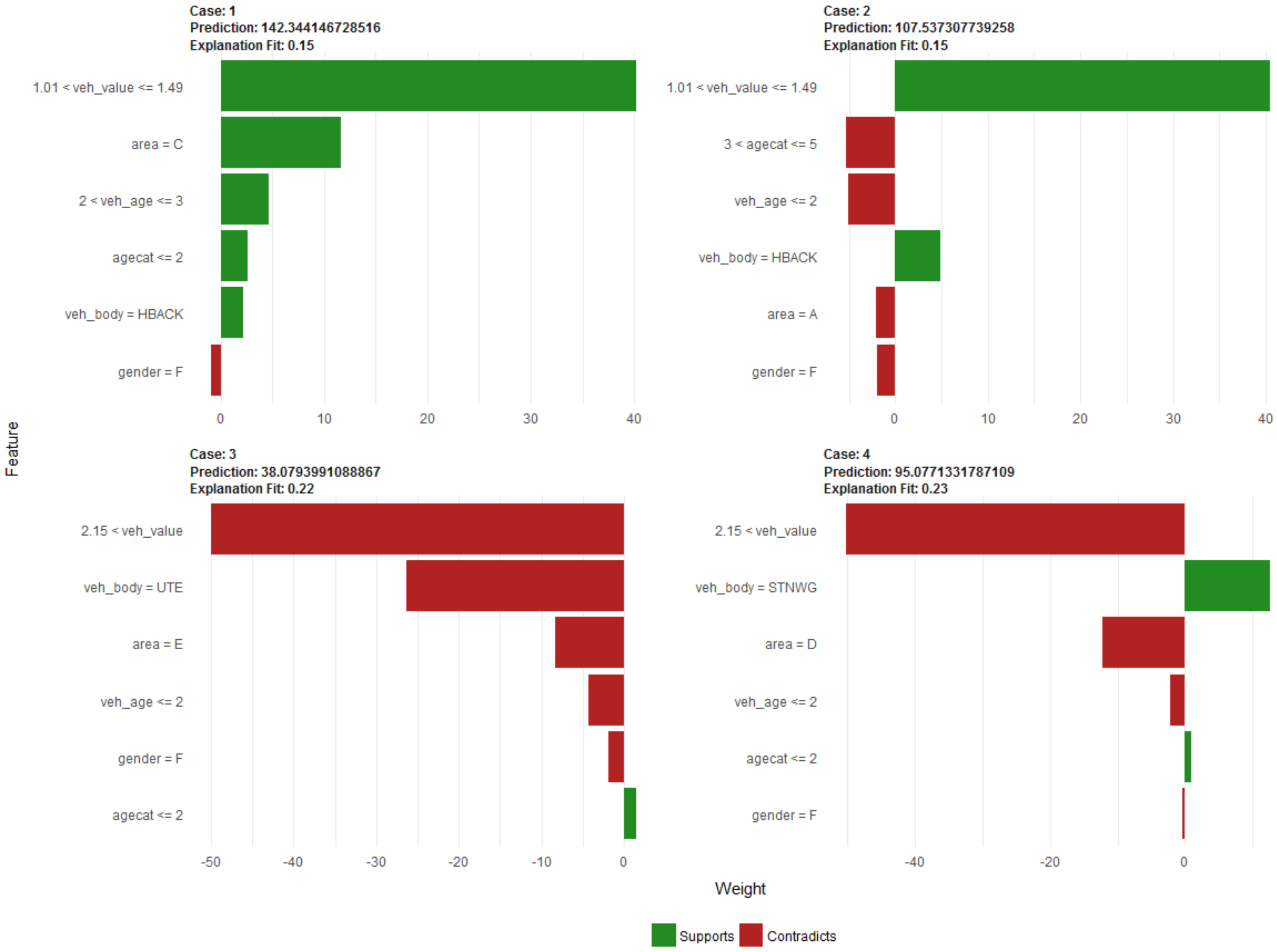


Local Surrogate Models (LIME)

Algorithm

- Choose instances to explain
- Permute instance to create replicated feature data
- Weight permuted instances with the original based on proximity
- Apply “black-box” machine learning model to predict outcomes of permuted data
- Fit a simple model, explaining the complex model outcome with the selected features from the permuted data weighted by its similarity to the original observation
- Explain predictions using this simpler model

LIME - Visualized



XGBFI (XGBoost)

- Computes variable importance and interaction importance ("Gain")
- Shows number of possible splits taken on a feature ("Fscore") and the cut-points chosen
- & more!

Interaction	Gain	FScore
veh_value	4,259,983,149	1,911
area	1,211,945,038	878
veh_body	1,147,646,618	914
veh_age	1,088,228,059	709
agecat	806,955,407	610
gender	707,919,139	514

Interaction	Gain	FScore
veh_value|veh_value	5,970,120,855	1,198
veh_age|veh_value	1,562,875,549	252
agecat|veh_value	1,311,331,233	299
veh_body|veh_value	1,295,426,670	313
area|veh_value	1,100,576,093	327
gender|veh_value	880,025,508	245

XGBFI (XGBoost)

veh_value	
split value	count
0.09	25
0.185	1
0.205	1
0.225	1
0.23	4
0.245	2
0.25	2
0.265	1
0.285	6
0.295	5
0.305	6
0.315	1
0.325	14
0.33	3
0.345	14
0.355	9
0.36	1

area	
split value	count
1.5	226
2.5	198
3.5	178
4.5	161
5.5	115

veh_body	
split value	count
1.5	1
2.5	7
3.5	58
4.5	121
5	1
5.5	47
6	15
6.5	29
7	8
7.5	64
8	2
8.5	18
9	70
9.5	16
10.5	190
11.5	132
12.5	135

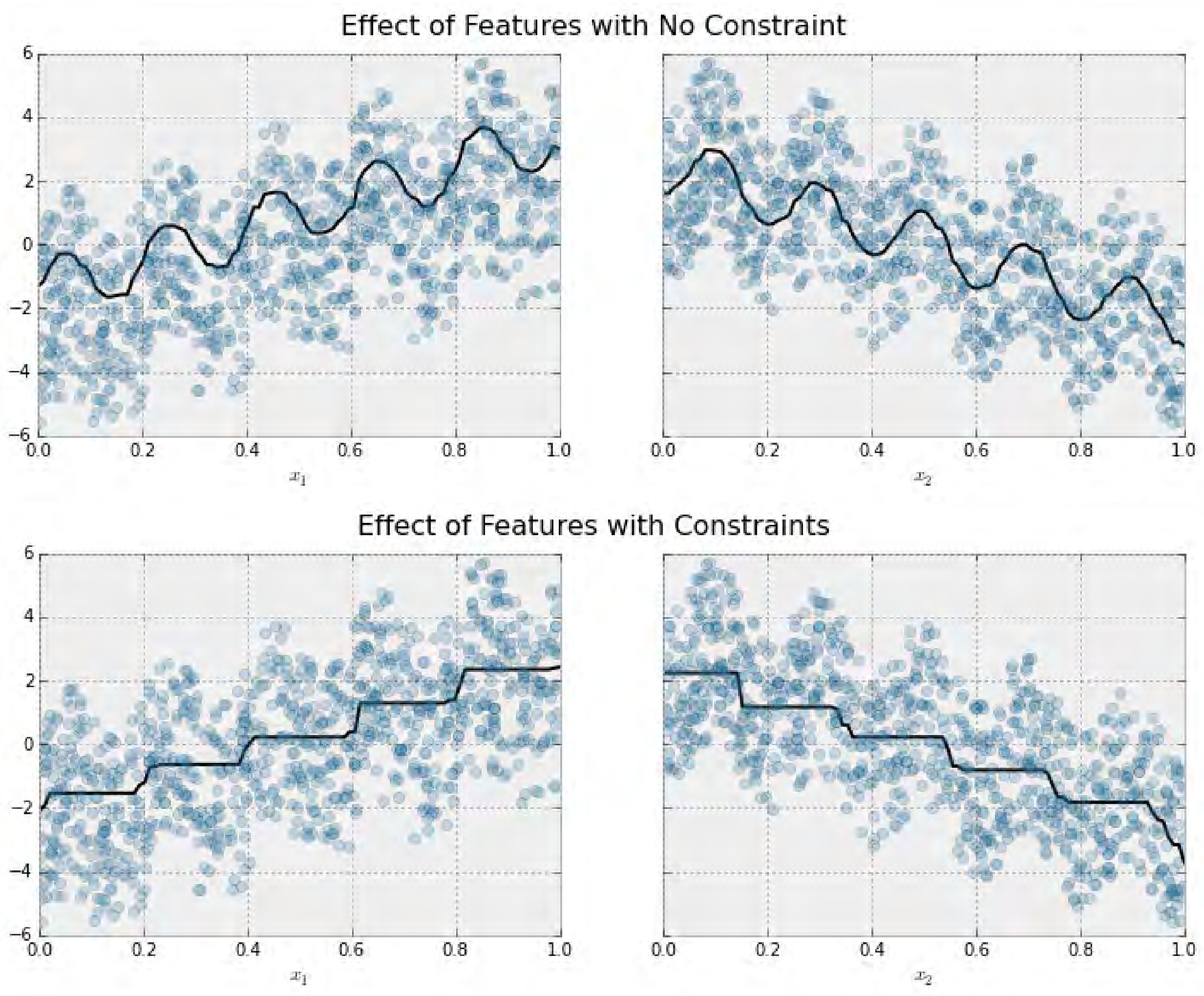
veh_age	
split value	count
1.5	212
2.5	222
3.5	275

agecat	
split value	count
1.5	129
2.5	134
3.5	121
4.5	116
5.5	110

gender	
split value	count
1.5	514

Monotonicity Constraints (XGBoost)

- Enforce a constraint on the model so that the predicted response can only increase / decrease for a given feature



http://xgboost.readthedocs.io/en/latest/tutorials/monotonic.html

Model Calibration

- Idea: The model ranks orders well, but the predictions are biased
- Model Types
  - Platt Scaling
  - Isotonic Regression
  - Polynomial or spline
- Evaluation
  - Brier Score
  - Logloss

Observation	Predicted Relativity	Actual Relativity
1	20%	100%
2	50%	120%
3	120%	140%
4	200%	160%
5	500%	250%

Bias / Fairness

Which model is unfair?

Assume: Two classes, A and B, have equal exposure with a probability of success of 10% and 20% respectively.

Model 1: The models predicts 10% success probability for class A and 20% for class B.

Model 2: The model predicts 15% probability of success regardless of class.

Model 3: The protected class is included in the model, but not statistically important (however, it may be correlated with attributes that are important).

Model 4: The model completely ignores whether someone is in class A or class B in making predictions. It turns out, however, that it gives different average predictions for the two classes.

Possible Definitions of Model Fairness

- **Fairness through Unawareness**: Ignore the protected class, hope that's OK
- **Individual Fairness**: Similar predictions for similar people
- **Demographic Fairness**: Same probability of favorable prediction across classes
$$\Pr\left(\hat{Y} = 1 \middle| \text{Class} = 1, Y = 1\right) = \Pr\left(\hat{Y} = 1 \middle| \text{Class} = 2, Y = 1\right)$$
- **Equality of Opportunity**: Same probability of favorable prediction across classes, conditional on having the positive attribute
$$\Pr\left(\hat{Y} = 1 \middle| \text{Class} = 1, Y = 1\right) = \Pr\left(\hat{Y} = 1 \middle| \text{Class} = 2, Y = 1\right)$$

The Impossible Trifecta

- Calibration across groups
- Calibration for the positive class
- Calibration for the negative class

If there is a true underlying differences across groups, all three criteria cannot be satisfied simultaneously!

Methods to Remove Model Bias

- Case Deletion
- Sampling
- Reweighting
- Shapley Values
- Generative Adversarial Networks

Note: All of these methods require that you have access to the protected attribute

Conclusion

References

- Interpretable Machine Learning: A Guide to Making Black Box Models Explainable https://christophm.github.io/interpretable-ml-book/
- XGBoost: http://xgboost.readthedocs.io/en/latest/
- Z. C. Lipton. The mythos of model interpretability. arXiv preprint arXiv:1606.03490, 2016.
- F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608, 2017.
- F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination.
- https://towardsdatascience.com/preventing-machine-learning-bias-d01adfe9f1fa
- Consistent Individualized Feature Attribution for Tree Ensembles https://arxiv.org/abs/1802.03888
- M. Hardt, E. Price, and N. Srebro. Equality of Opportunity in Supervised Learning https://ttic.uchicago.edu/~nati/Publications/HardtPriceSrebro2016.pdf

Software

- iml (R)
- LIME (R / Python)
- SKATER (Python)
- XGBFI (R - xgboost)
- xgboostExplainer (R - xgboost)
- DALEX (R)
- $H_2$0 Driverless AI
- Aequitas
- Themis ML (Python)

Thank you

Michael.Niemerg@milliman.com