Machine Learning & Risk (TU9)

2025 BANFF PIPELINE WORKSHOP

Michael Gloven, PE
michaelgloven@pipeline-risk.com

Land Acknowledgement

The Banff Centre for Arts and Creativity is located on the side of Sacred Buffalo Guardian Mountain.

- The Banff area, known as "Minhrpa" (translated in Stoney Nakoda as "the waterfalls") is part of the Treaty 7 territory where we recognize oral practices of the Îyârhe Nakoda (Stoney Nakoda) – comprised of the Bearspaw, Chiniki, and Goodstoney First Nations – as well as the Tsuut'ina First Nation and the Blackfoot Confederacy comprised of the Siksika, Piikani, Kainai.

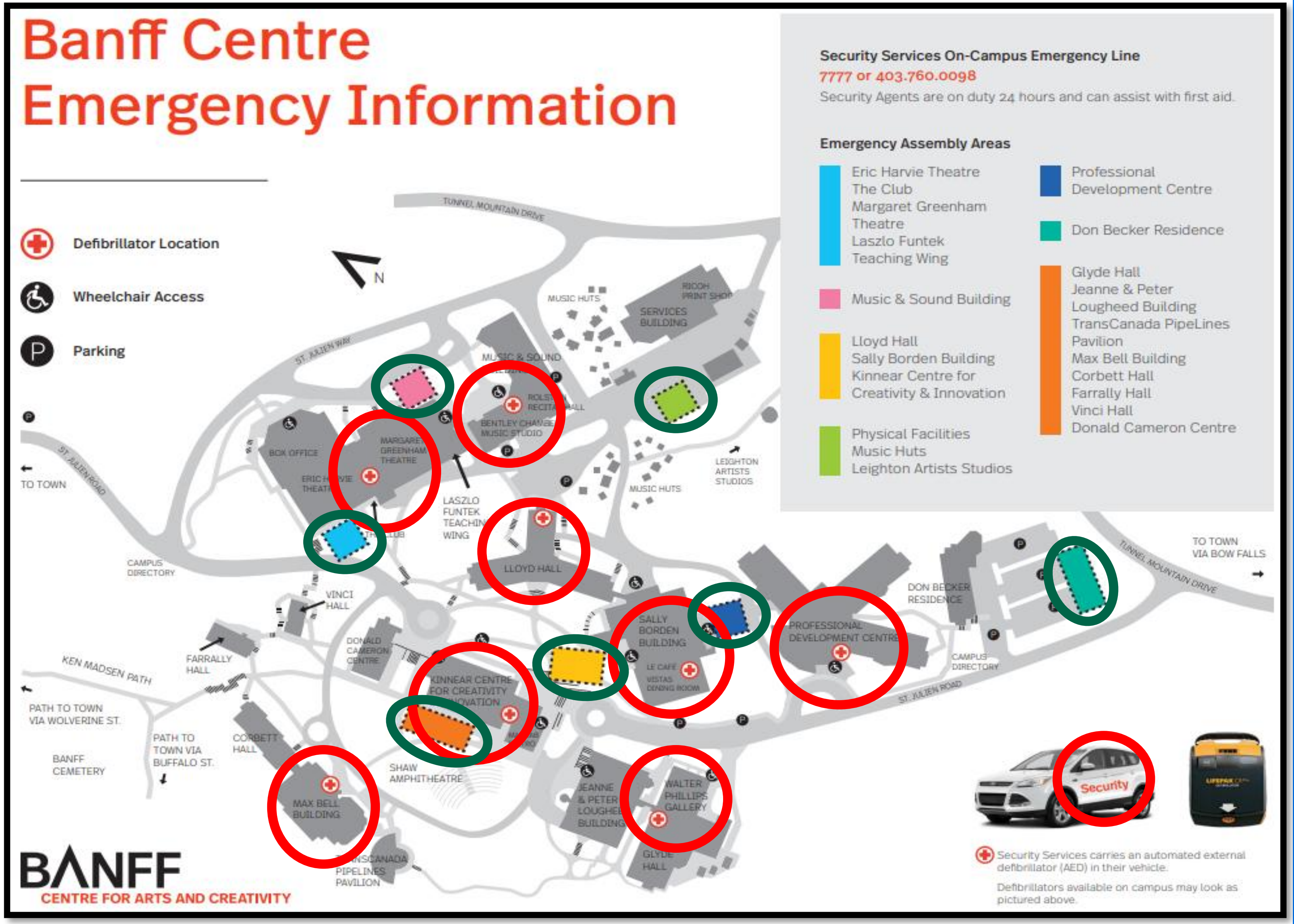
- This territory is home to the Shuswap Nations, Ktunaxa Nations, and Métis Nation of Alberta, Region 3. We acknowledge all Nations who live, work, and play here, help us steward this land, and honour and celebrate this place.

Site Safety

Site Security Cellphone:

403-760-0098

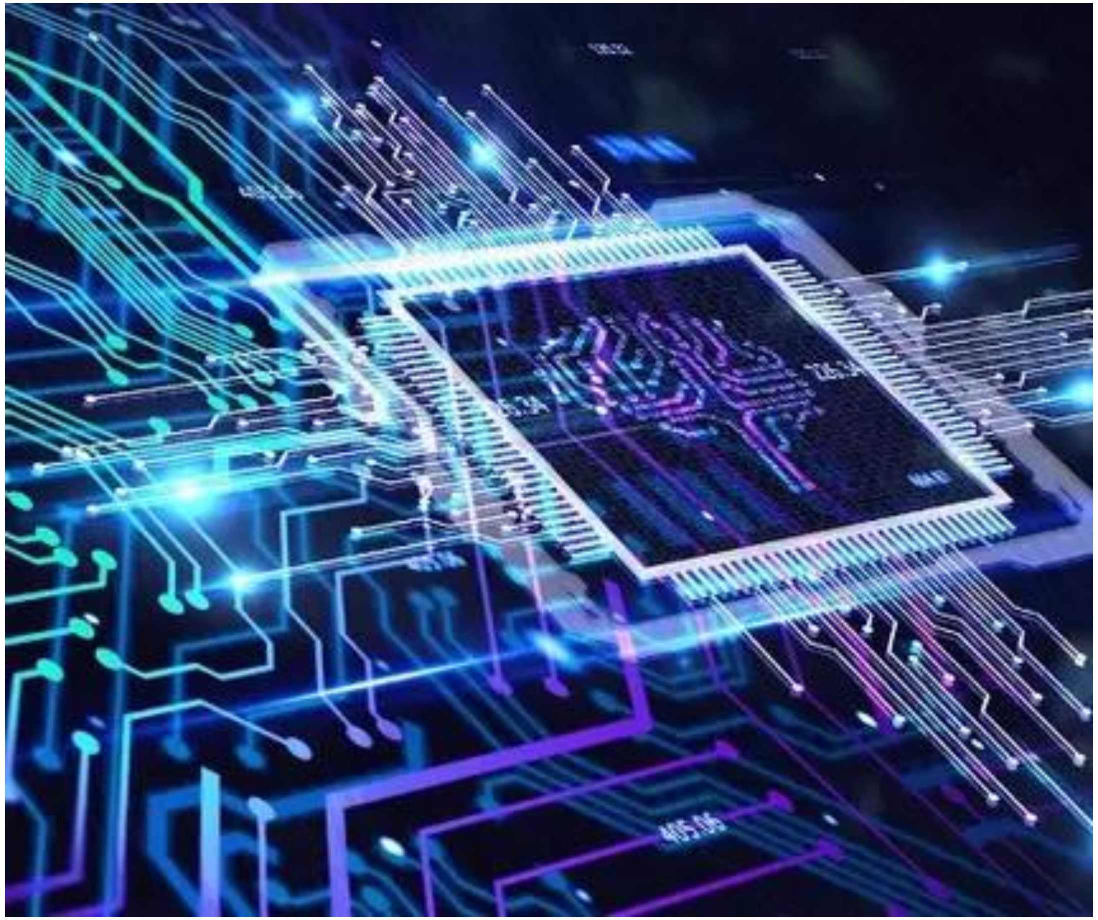
- **AED: 8 Locations + Security Cars**
- **7777: House Phone**
- **911: Emergencies** Be as specific as possible – building, room, and issue
- **30 Second Alarm:** Proceed to **Muster** Points



2025
BANFF PIPELINE WORKSHOP

Machine Learning & Risk

- Big Picture & Common Questions
- Use Cases (Learning Tasks)
  - □ Classification
  - □ Regression
  - □ Time-Series
  - □ Model Validation
- Summary & Wrap-Up



Big Picture - The Business Case

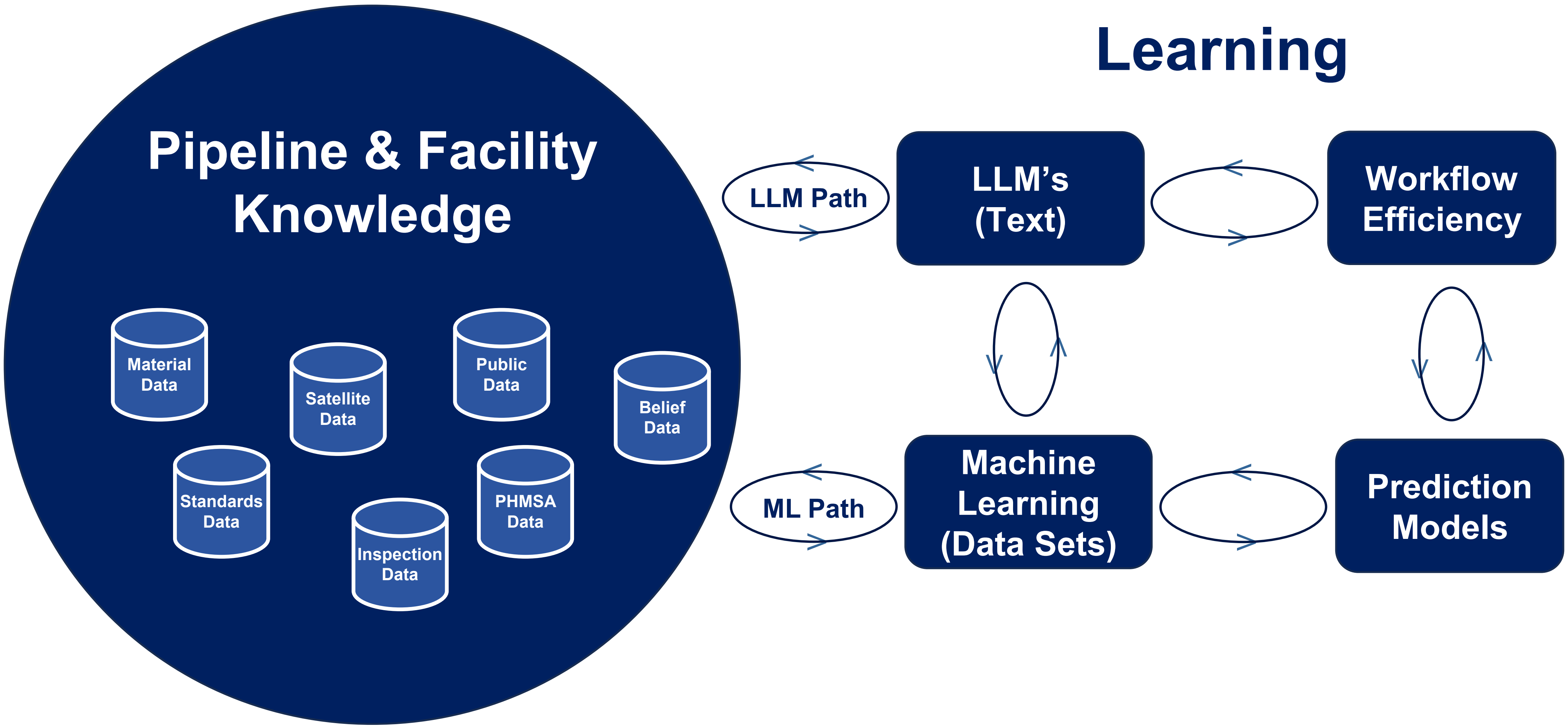
- **Reduce Over-Inspections:** Pinpoint exactly which assets need attention, reducing unnecessary inspections while ensuring critical risks aren't overlooked

- **Uncover Hidden Risks:** Detect subtle patterns, interactions and anomalies that deterministic models might miss

- **Speed Up Decision-Making:** Automate complex analysis to deliver actionable insights faster

- **Maximize Existing Investments:** ML integrates with your current systems, enhancing accuracy and efficiency without requiring a complete overhaul.

Big Picture



Learning
Pipeline & Facility Knowledge
Material Data
Satellite Data
Public Data
Belief Data
Standards Data
Inspection Data
PHMSA Data
LLM Path
LLM's (Text)
Workflow Efficiency
ML Path
Machine Learning (Data Sets)
Prediction Models

Big Picture – LLM’s

The equation for pipeline risk is probability of failure (PoF) x Consequence of Failure (CoF)

Human Enters “Truth”

Token count

33

Truth Converted to Tokens

```
<|im_start|>system<|im_sep|>The equation for pipeline risk
is probability of failure (PoF) x Consequence of Failure (C
oF)<|im_end|><|im_start|>user<|im_sep|><|im_end|><|im_start
|>assistant<|im_sep|>
```

Training Data Learns (Weights) this Sequence

```
200264, 17360, 200266, 976, 42006, 395, 29412, 7031, 382, 3
4933, 328, 14090, 350, 25010, 37, 8, 1215, 25365, 6975, 32
8, 59925, 350, 8037, 37, 8, 200265, 200264, 1428, 200266, 2
00265, 200264, 173781, 200266
```

Big Picture – LLM's

what is the equation for pipeline risk?

DeepSearch Think Attach

Human Asks Question

LLM Returns "Best" Response based on Context & Weights

Risk = Probability of Failure (PoF) × Consequence of Failure (CoF)

Breakdown:

1. **Probability of Failure (PoF):**
   - This represents the likelihood that the pipeline will fail due to factors like corrosion, external damage, material defects, or operational errors.
   - It's often calculated using historical data, inspection results, or statistical models.
   - Example factors: pipe age, wall thickness, pressure, environmental conditions.
2. **Consequence of Failure (CoF):**
   - This measures the potential impact if a failure occurs, such as environmental damage, economic loss, or human safety risks.
   - It's often quantified in terms of cost (e.g., dollars), area affected, or population at risk.

Big Picture – LLM's & Data Stores

write code to calculate pipeline risk against an MS SQL table

DeepSearch Think Attach

Human Requests Code to Apply to Data Store since Data Stores are not Tokenized for Learning

```
R

# Load required libraries
library(DBI)
library(odbc)

# Establish connection to MS SQL Server
con <- dbConnect(odbc::odbc(),
                 Driver = "ODBC Driver 17 for SQL Server",  # Adjust driver name if different
                 Server = "your_server",                    # e.g., "localhost" or IP address
                 Database = "your_db",                      # e.g., "PipelineDB"
                 UID = "your_user",                         # e.g., "sa"
                 PWD = "your_password",                     # Your SQL Server password
                 Port = 1433)                               # Default MS SQL port

# Check if connection is successful
if (dbIsValid(con)) {
  cat("Connected to MS SQL Server database successfully!\n")
} else {
  stop("Failed to connect to the database.")
}

# Function to calculate pipeline risk
calculate_pipeline_risk <- function(pof, cof) {
  return(pof * cof)
}

# Query to fetch pipeline data
query <- "SELECT segment_id, pof, cof FROM dbo.pipeline_data"
pipeline_data <- dbGetQuery(con, query)

# Calculate risk for each segment
pipeline_data$risk <- mapply(calculate_pipeline_risk, pipeline_data$pof, pipeline_data$cof)
```

LLM Returns Code to Apply to Data Sets

Code Requires Function Inputs, Review and Verification by Human or System

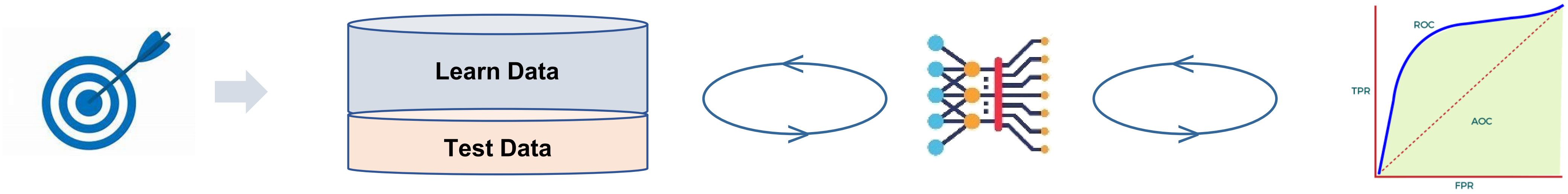
Machine Learning & Risk

Learning Target
(Threats, Consequence, Risk)

Training Data
(Observations)

Learned Model
(Methods, Tuning)

Performance & Insights
(Validation & Acceptance)



"Use LLM's to Create Code to Support Machine Learning Process"

Use LLM

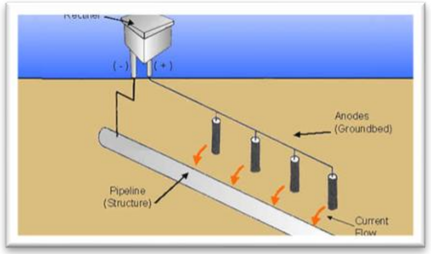
Machine Learning Tasks

Classification

- Probability of Cracking
- Probability of Third-Party Damage
- Probability of Pipe Manufacturer

Regression

- Prediction of Corrosion Growth Rates
- Prediction of Inspection Costs
- Simulation of Deterministic Results

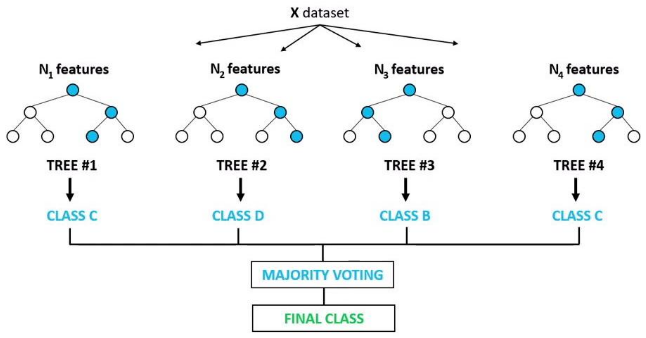


Time Series

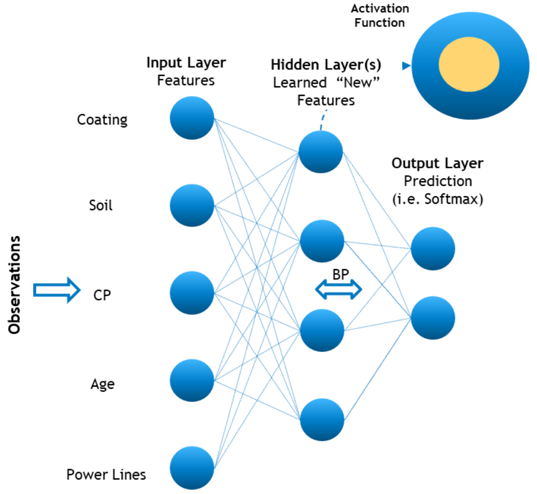
- Prediction of CP Readings
- Prediction of Ground Bed Life
- Simulation of Seasonal Patterns

Learning Methods

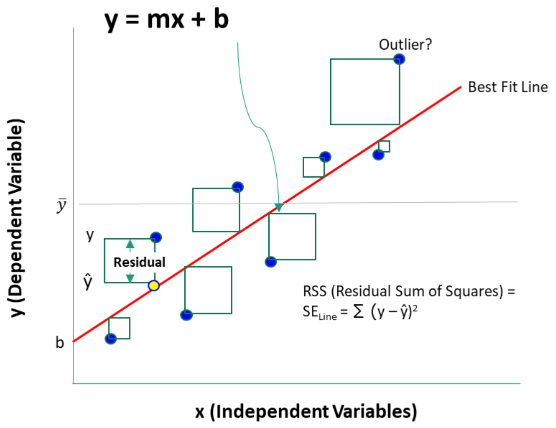
Tree Bagging



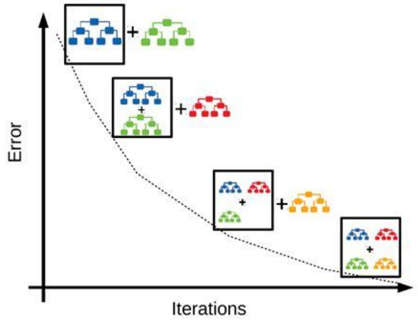
Neural Net



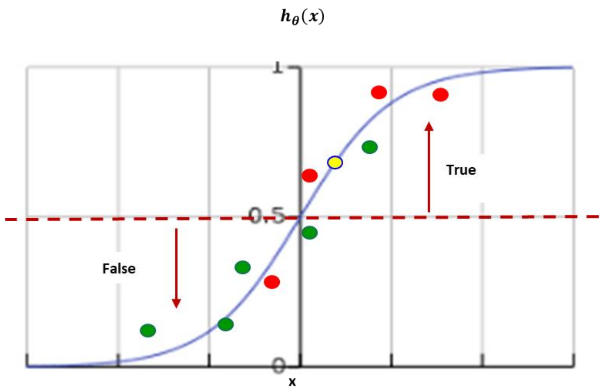
Linear Regression



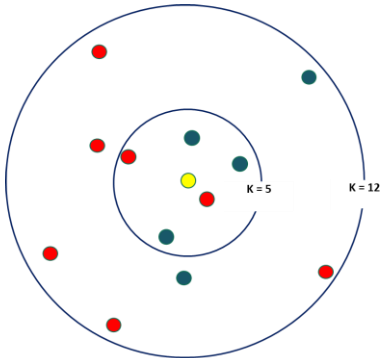
Tree Boosting



Logistic Regression



KNN

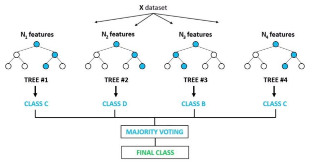


Hundreds of Methods are Available to Practitioner

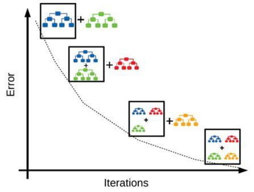
2025 BANFF PIPELINE WORKSHOP

Method Tuning

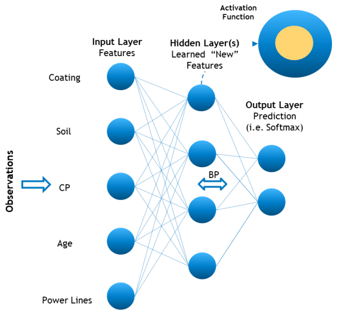
Open-Source Methods & Hyper-Parameters



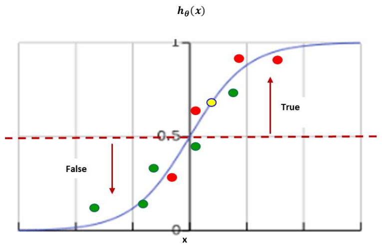
- Number Trees
- Tree Depth
- Min Observations
- Stop Criteria



- Number Trees
- Tree Depth
- Min Observations
- Stop Criteria



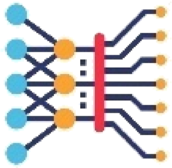
- Layers
- Activation Function
- Learning Rate



- Solver
- Regularization

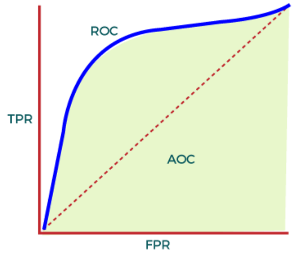
Iterations

Model



Iterations

Performance



Final Model uses One Method with Tuned Parameters & Acceptable Performance

2025 BANFF PIPELINE WORKSHOP

Domain Experts FAQ

- Are Machine Learned Models an Improvement Over Deterministic Models?
- Do I have Enough of the Right Data?
- Are Patterns Inferential or Predictive? What's the Difference?
- Does the Model Meet Domain Expert Review?
- What Assets can I Apply the Learned Model?
- Is Performance Acceptable for Production Use?

Stress Corrosion Cracking

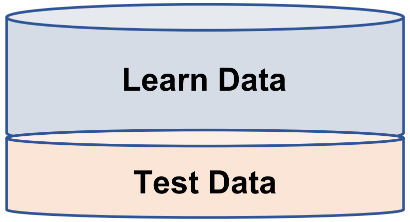
Classification Model Example

2025
BANFF PIPELINE WORKSHOP

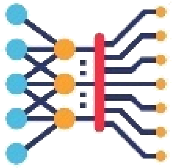
Machine Learning Process

Learning Target
(SCC Found)

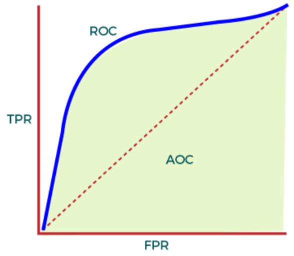
Training Data
(Observations)



Learned Model
(Methods, Tuning)



Performance & Insights
(Validation & Acceptance)



Classification – Find\Learn Pattern to Predict Categorical Values

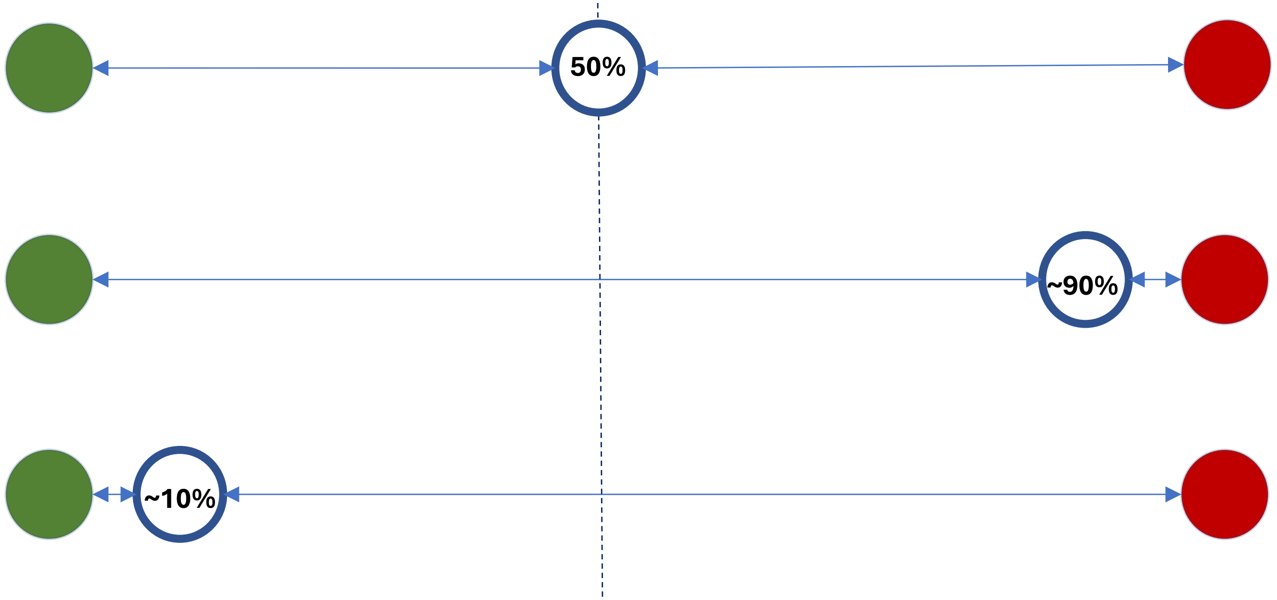
Classification Intuition

Learning Observations are Points in n-Dimensional Vector Space

Learning Observation
No SCC

New Points
What is Probability of Corrosion?

Learning Observation
Yes SCC



No Corrosion <---- Threshold ----> Yes Corrosion

How do you make the class call?

Model Performance

Two-Class Performance Learning Data Example:

- Joints of Pipe = 100
- Joints with Defects = 10
- Joints without Defects = 90

Overall Accuracy 89%	Actual (No Defects = 90)	Actual (Defects = 10)	
Prediction (No Defects = 81)	**80** (TN = true negatives)	**1** (FN = false negatives)	
Prediction (Defects = 19)	**10** (FP = false positives)	**9** (TP = true positives)	**47%** (precision)
	89% (specificity)	**90%** (sensitivity or recall)	

2025 BANFF PIPELINE WORKSHOP

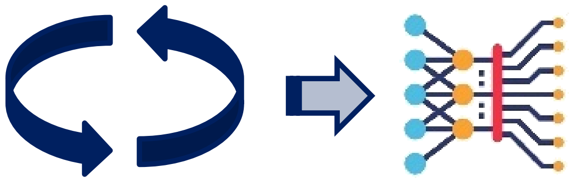
Training Data

Learning Target

Predictors

SCC	Coating_Condition	Coating_Type_Body_rev	cp_mv_mean	Depth_of_Cover_ft	pr_Precip_Annual	s_Avg_wthirdbar_r
	All	All	All	All	All	All
No	Poor	No Data	-1,578.00	2.30	32.00	16.32
No	Poor	Wax_Unknown	-1,253.00	6.00	31.00	39.73
No	Well	Asphalt	-1,654.00		32.00	19.30
No	Poor	Wax_Unknown	-1,385.00	5.00	37.00	32.20
No	Poor	Wax_Unknown	-1,383.00	2.67	37.00	32.20
Yes	Well	Wax_Unknown	-1,240.00		31.00	22.42
Yes	Well	Wax_Unknown	-1,318.00		31.00	54.50
No	Well	Wax_Unknown	-1,860.00	3.00	32.00	20.80
No	Well	Wax_Unknown	-1,728.00	5.00	32.00	23.55
No	Excellent	Asphalt	-1,520.00	5.00	20.00	12.20
No	Well	Wax_Unknown	-1,725.00	5.00	32.00	18.73
No	Well	Wax_Unknown	-1,546.00	4.00	33.00	37.50
No	Well	Wax_Unknown	-1,717.50	5.00	32.00	18.73
No	Excellent	Epoxy	-1,395.00	9.20	29.00	24.12
No	Fair	Wax_Unknown	-1,176.00	5.00	31.00	12.52
No	Fair	Wax_Unknown	-1,748.00	5.00	34.00	11.58
No	Fair	Coal Tar	-1,064.00		44.00	13.50

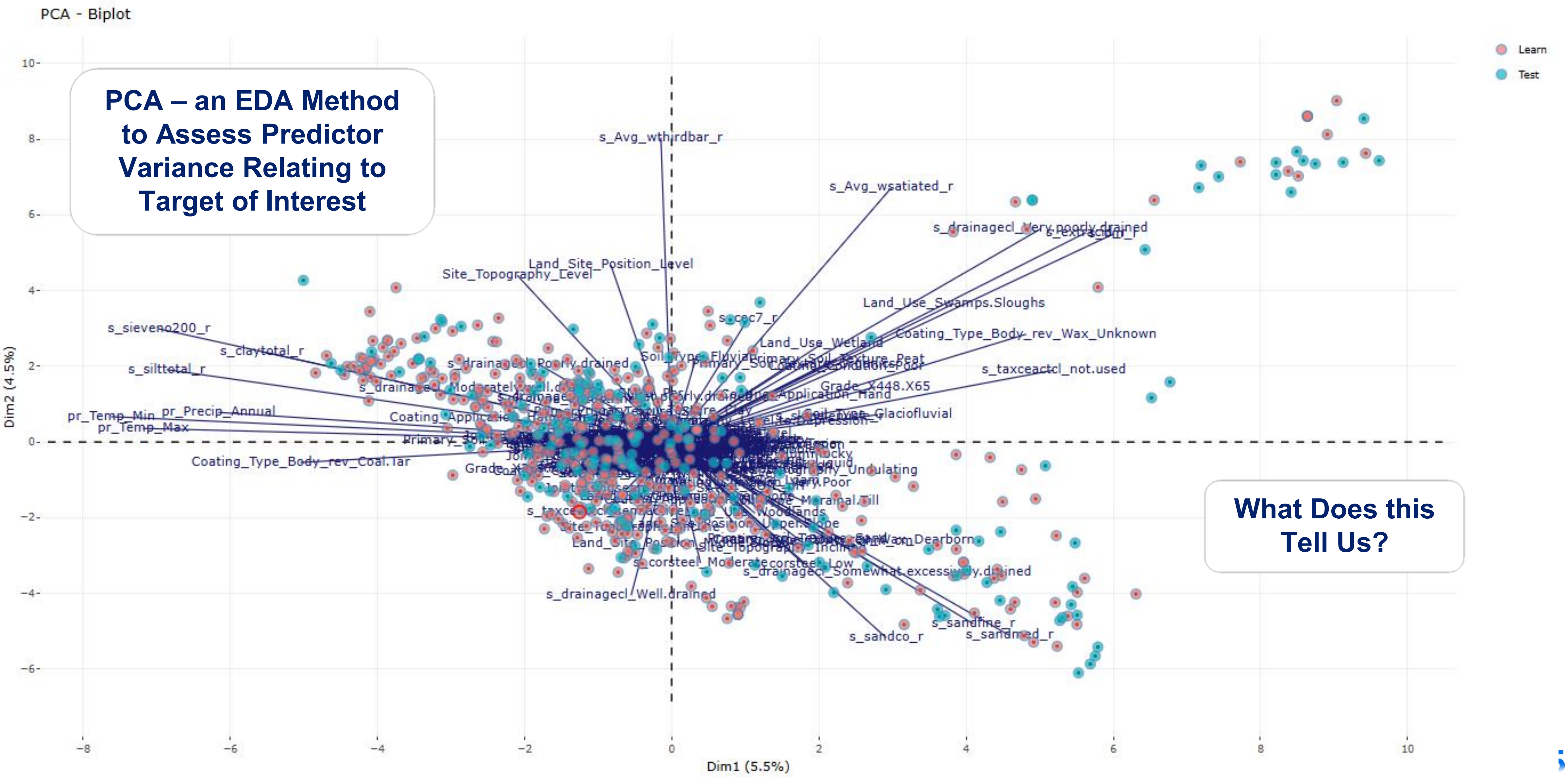
Training Data



Machine Learning Process

Learned Model

Principal Component Analysis

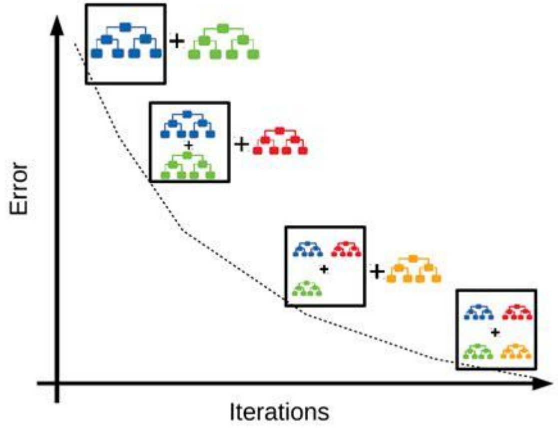


PCA – an EDA Method to Assess Predictor Variance Relating to Target of Interest

What Does this Tell Us?

Learned Model

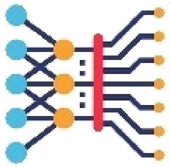
Gradient Boosted Tree



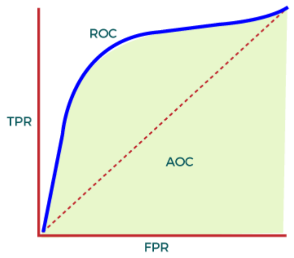
Tuned Hyper-Parameters

mtry	26
trees	960
min_n	4
tree_depth	14
learn_rate	0.0013
loss_reduction	0
sample_size	0.2247
stop_iter	5

Model



Performance



Test Data

Candidate Model Performance

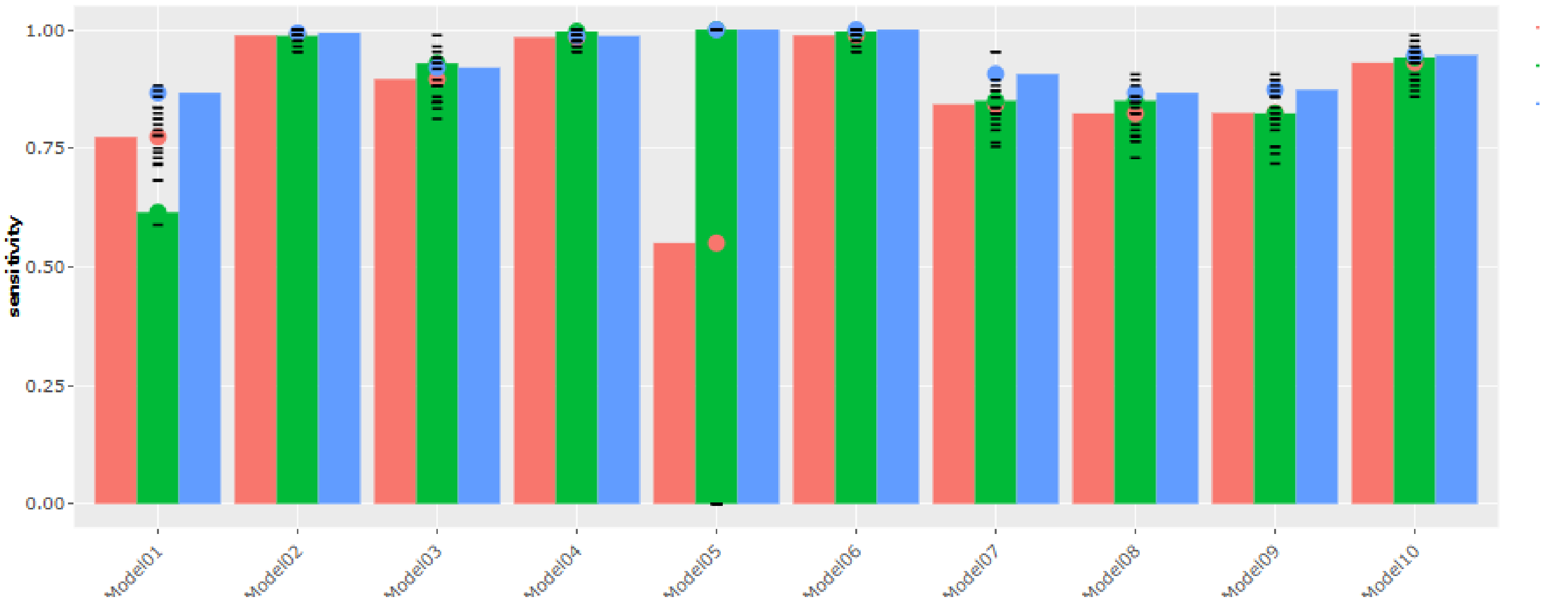
Metrics

- Accuracy
- Sensitivity
- Specificity
- AUC

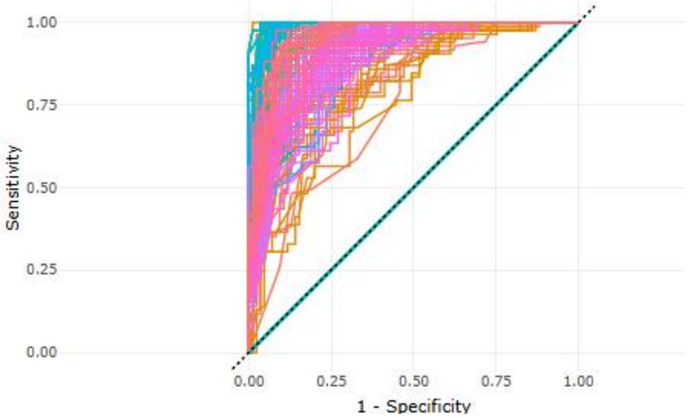
Selected Model

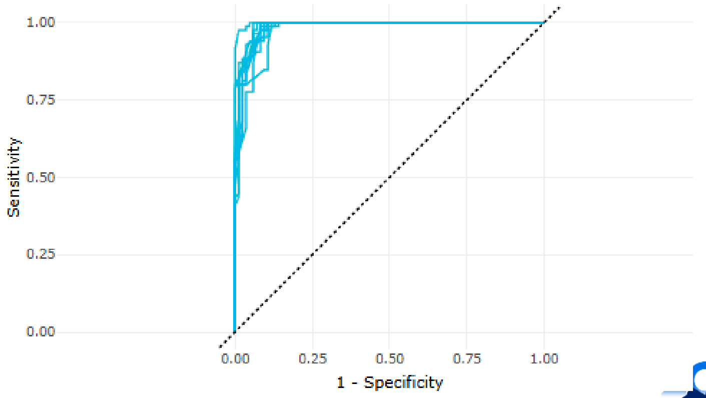
- Xgboost Method
- 2000 Trees
- 5 Depth
- 2 Min Obs
- .0001 Loss

Candidate Model Performance (Resample, Learn, Test)



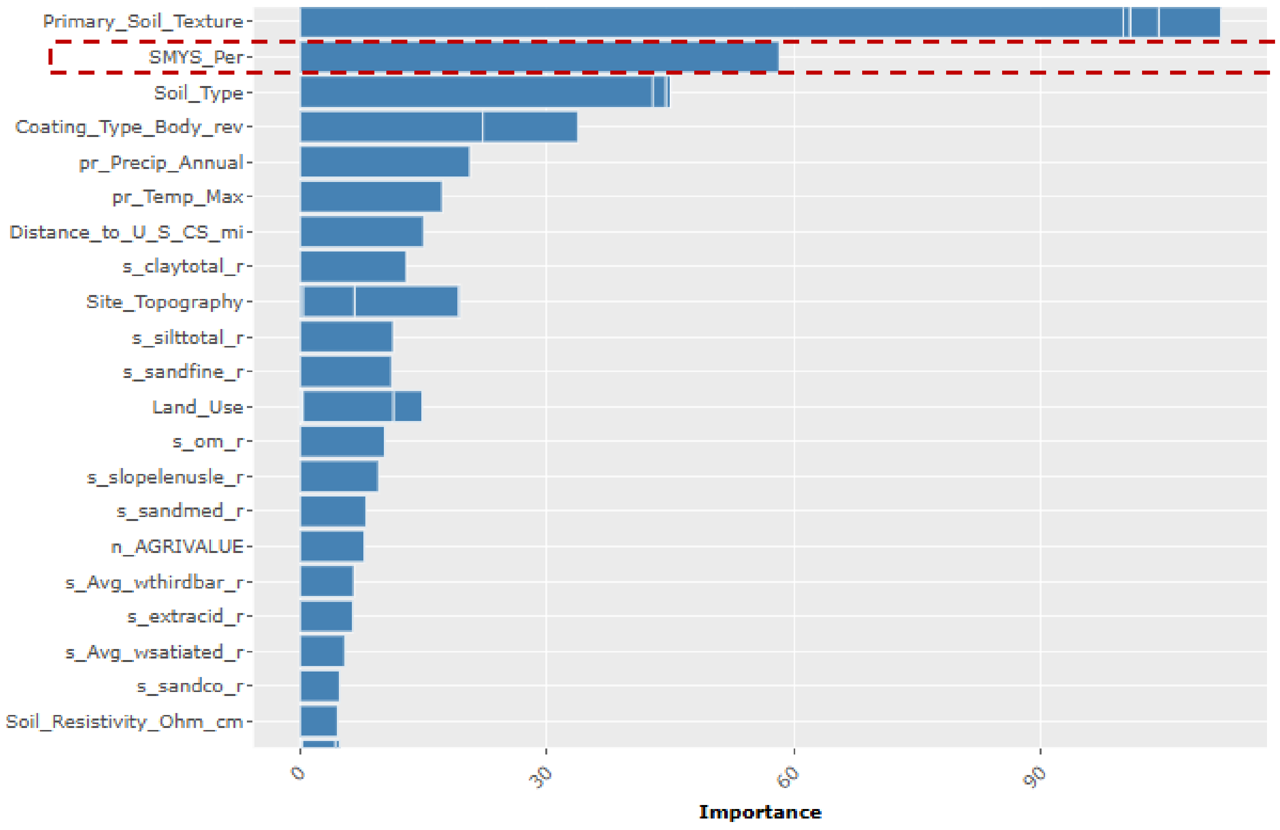
Candidate Model ROC's



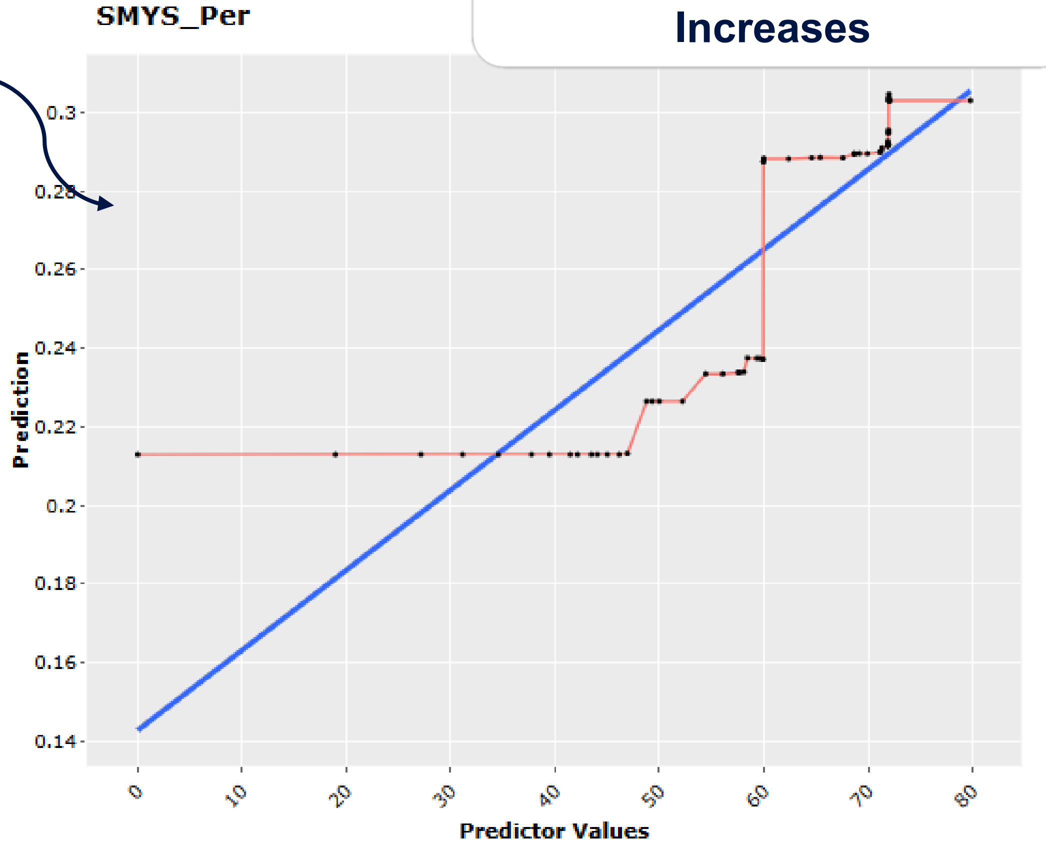


Model Weights

Model Predictor Importance



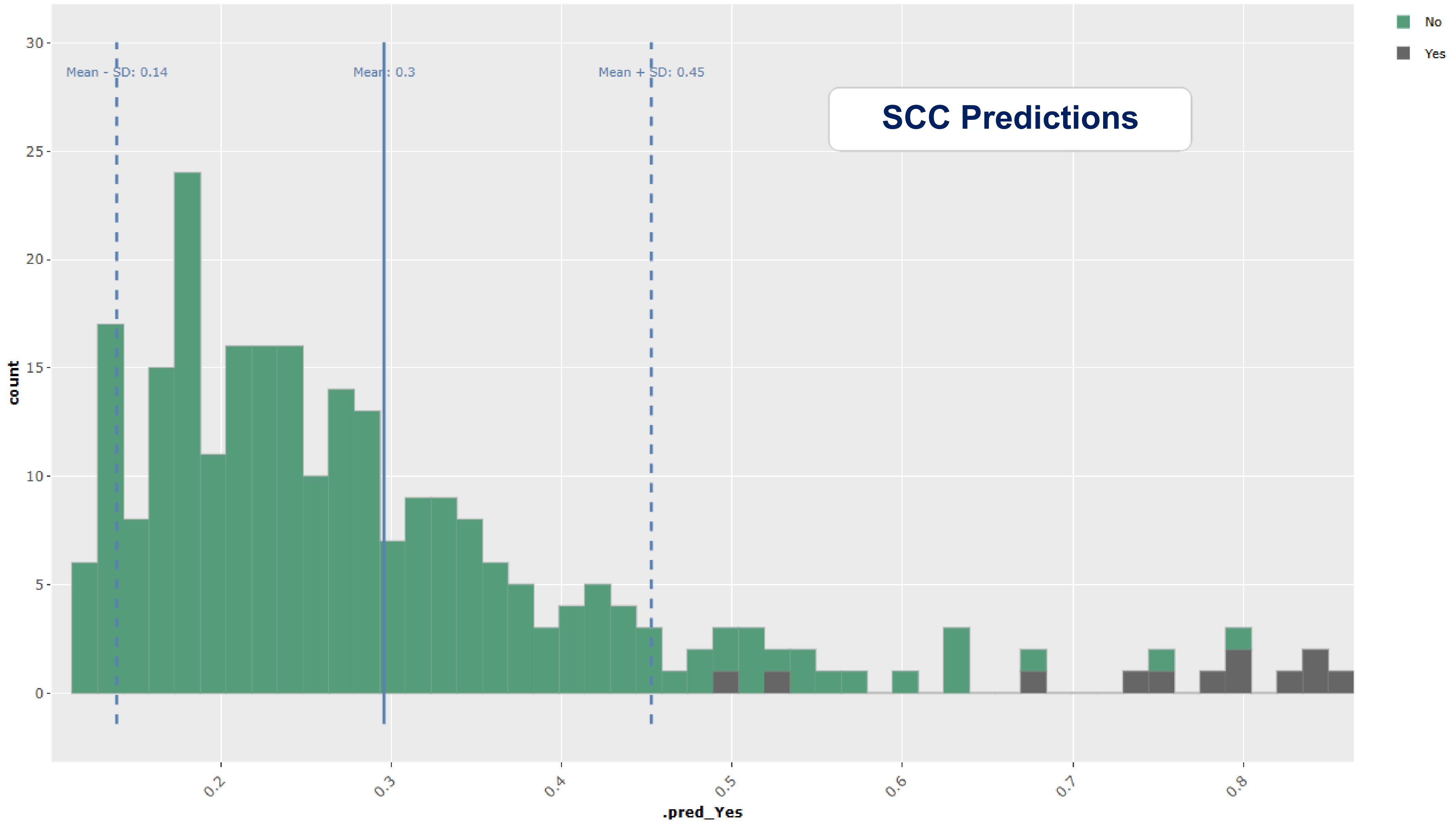
Model Predictor Directionality



Probability of SCC Increases as SMYS Increases

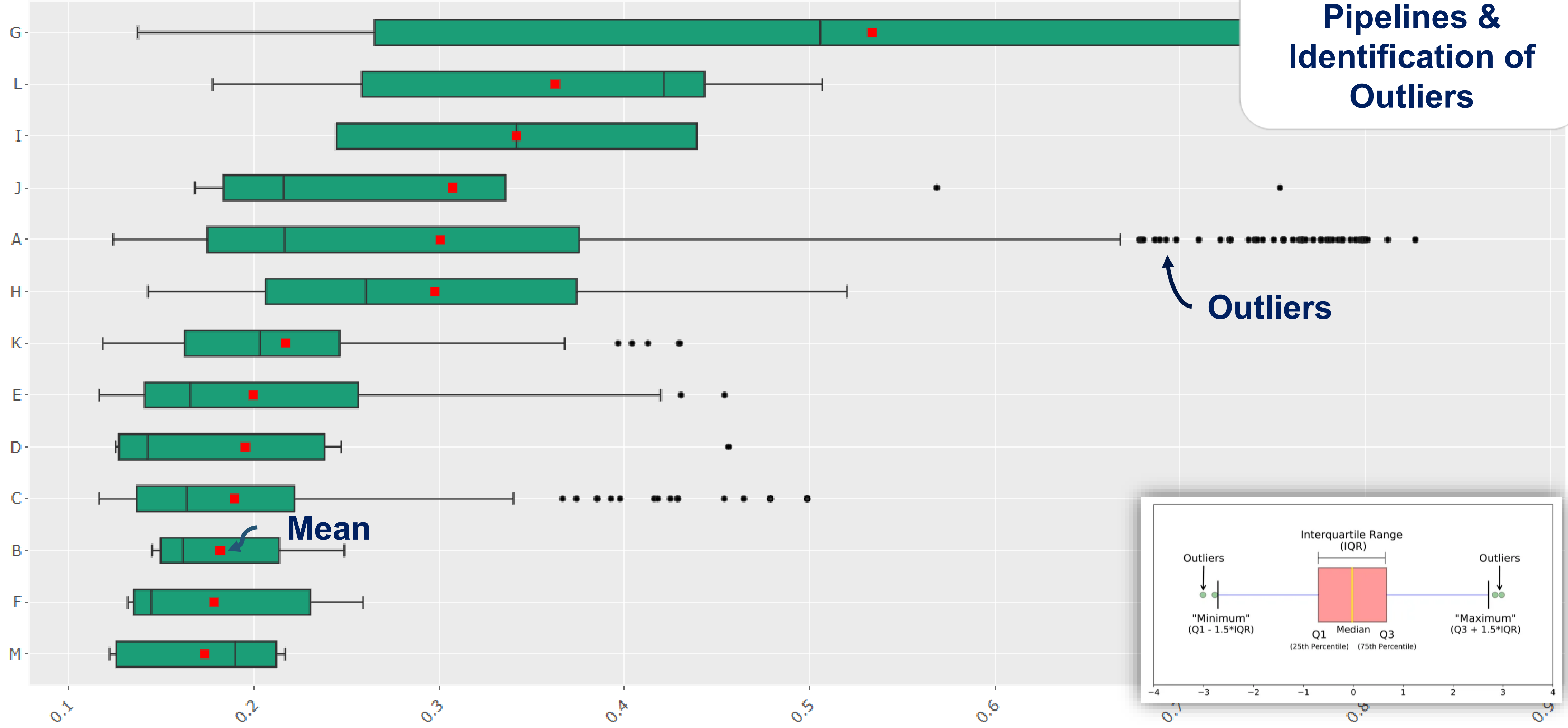
2025 BANFF PIPELINE WORKSHOP

Typical ML Output



2025 BANFF PIPELINE WORKSHOP

Model Application



SCC Predictions by Pipeline

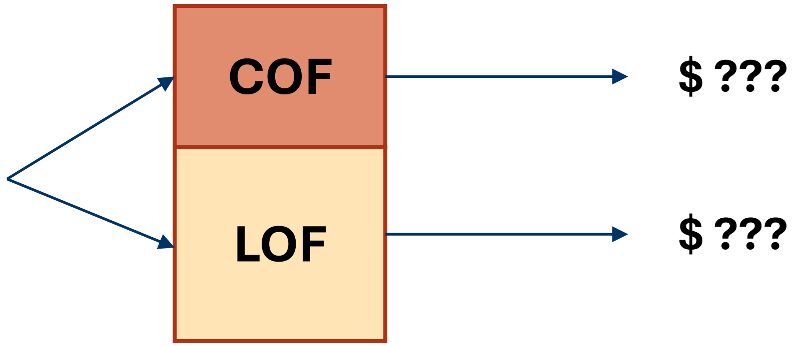
Prioritization of Pipelines & Identification of Outliers

Model Explainability

ROF = LOF x COF

Example:

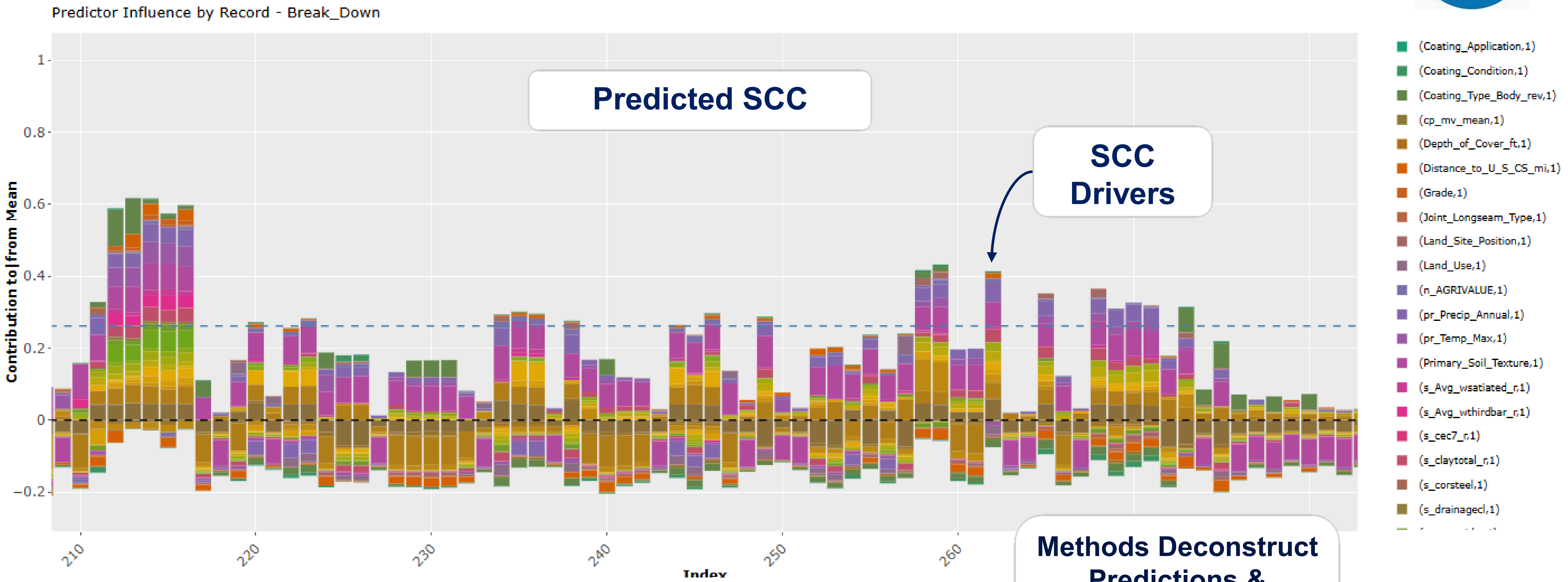
$120 = 60% x $200



- What is the Contribution of LOF to ROF in absolute terms?
- What is the Contribution of COF to ROF in absolute terms?

- Is it useful to know these contributions?
- What if you have a risk algorithm with 100 predictors & non-linearities?
- How do you know what each factor contributes?

Results Explanation



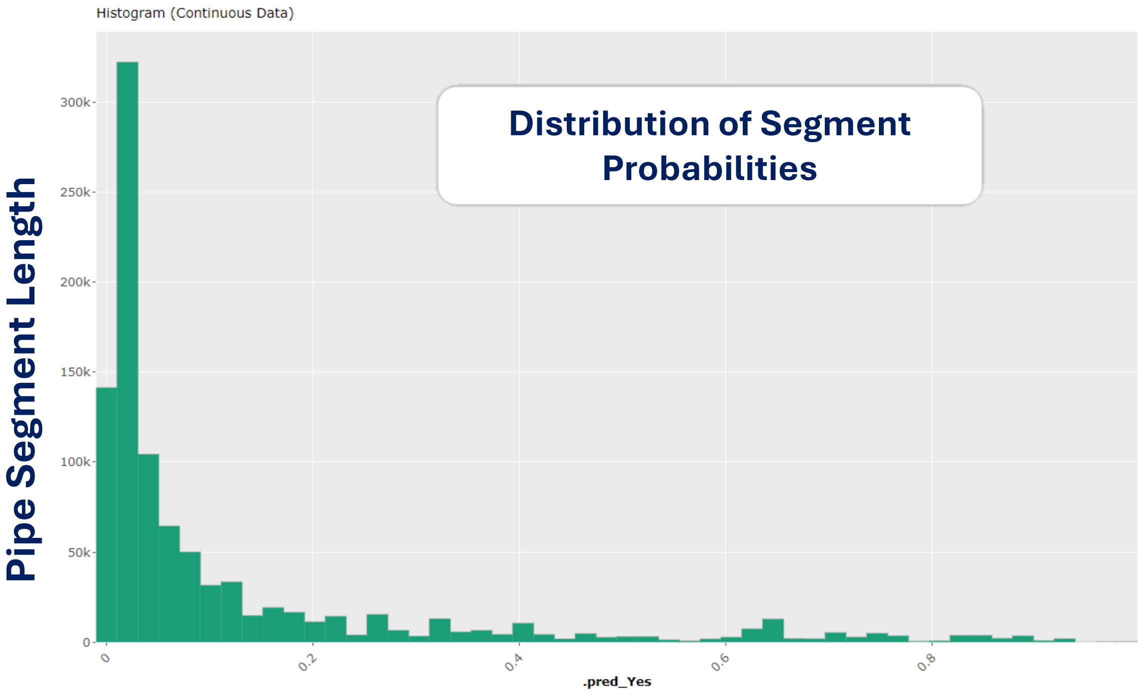
Pipeline Segments

Methods Deconstruct Predictions & Consider Interactions & Non-Linearities

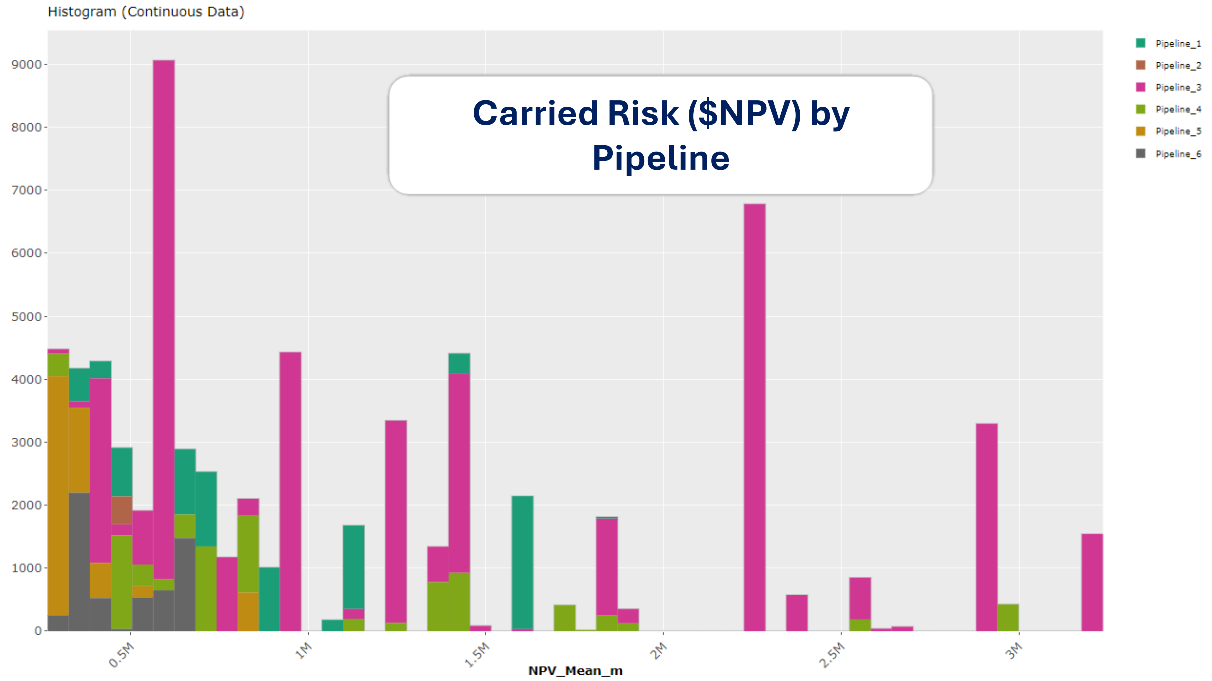
QRA – Classification Time Independent

NPV(Σ [Probability x Event Rate (#/length-yrs) x Event Cost ($/Event)], Yrs)

- Consider Resistance (pipe WT, toughness)
- Normalize to Incident & Consequence Event Distributions (P50\P99)



Probability of Third-Party Damage



Monetized Carried Risk by Pipeline

Machine Learned Results

QRA Monetized Risk

2025 BANFF PIPELINE WORKSHOP

Questions?

BANFF PIPELINE WORKSHOP

Corrosion Growth Rate

Regression Model Example

Machine Learning Process

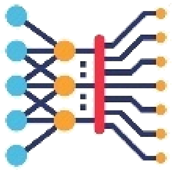
Learning Target
(EC CGR)

Training Data
(Observations)

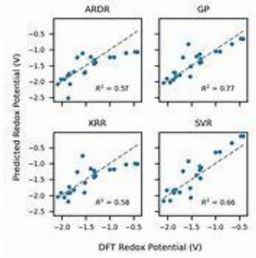
Learn Data

Test Data

Learned Model
(Methods, Tuning)



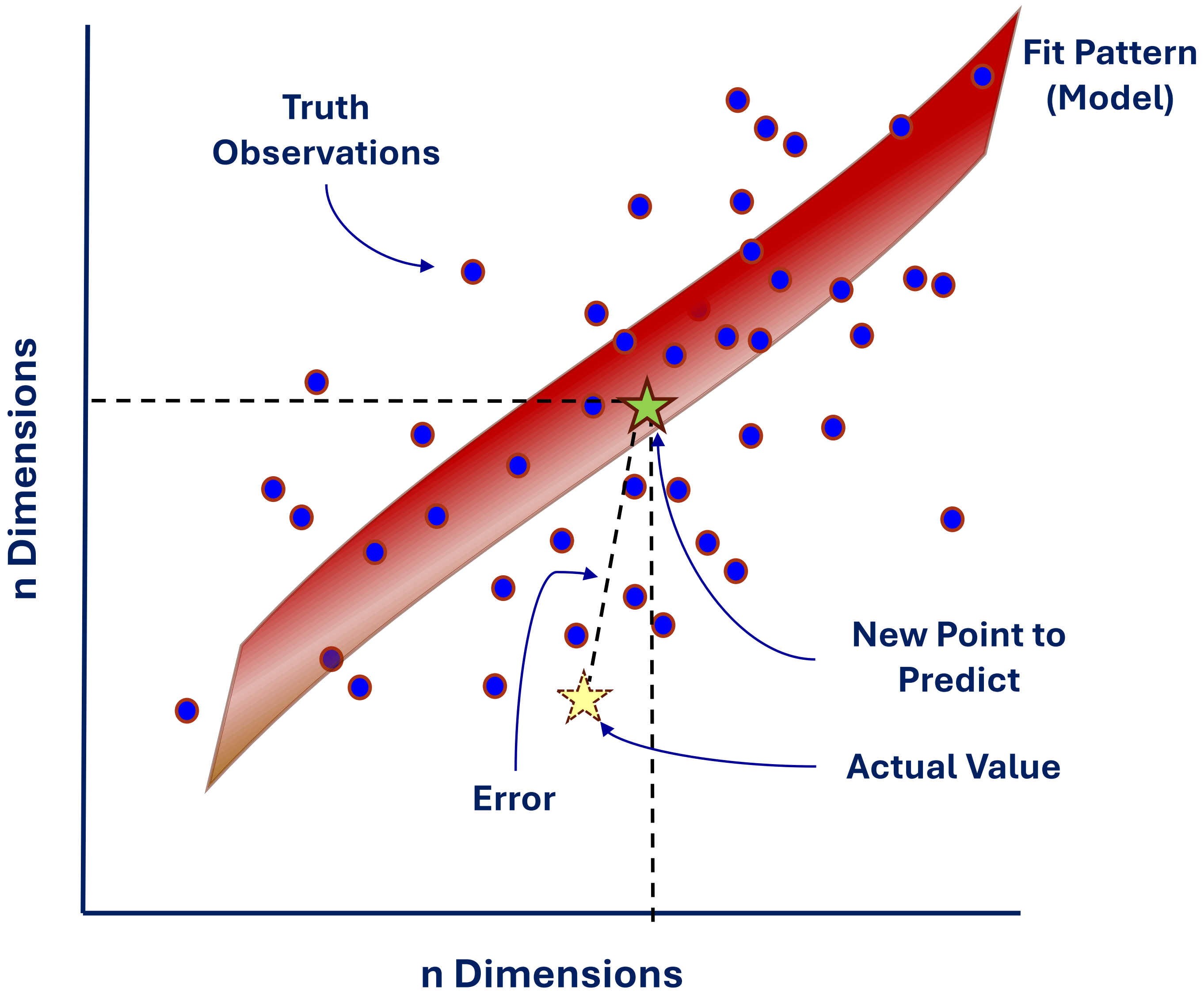
Performance & Insights
(Validation & Acceptance)



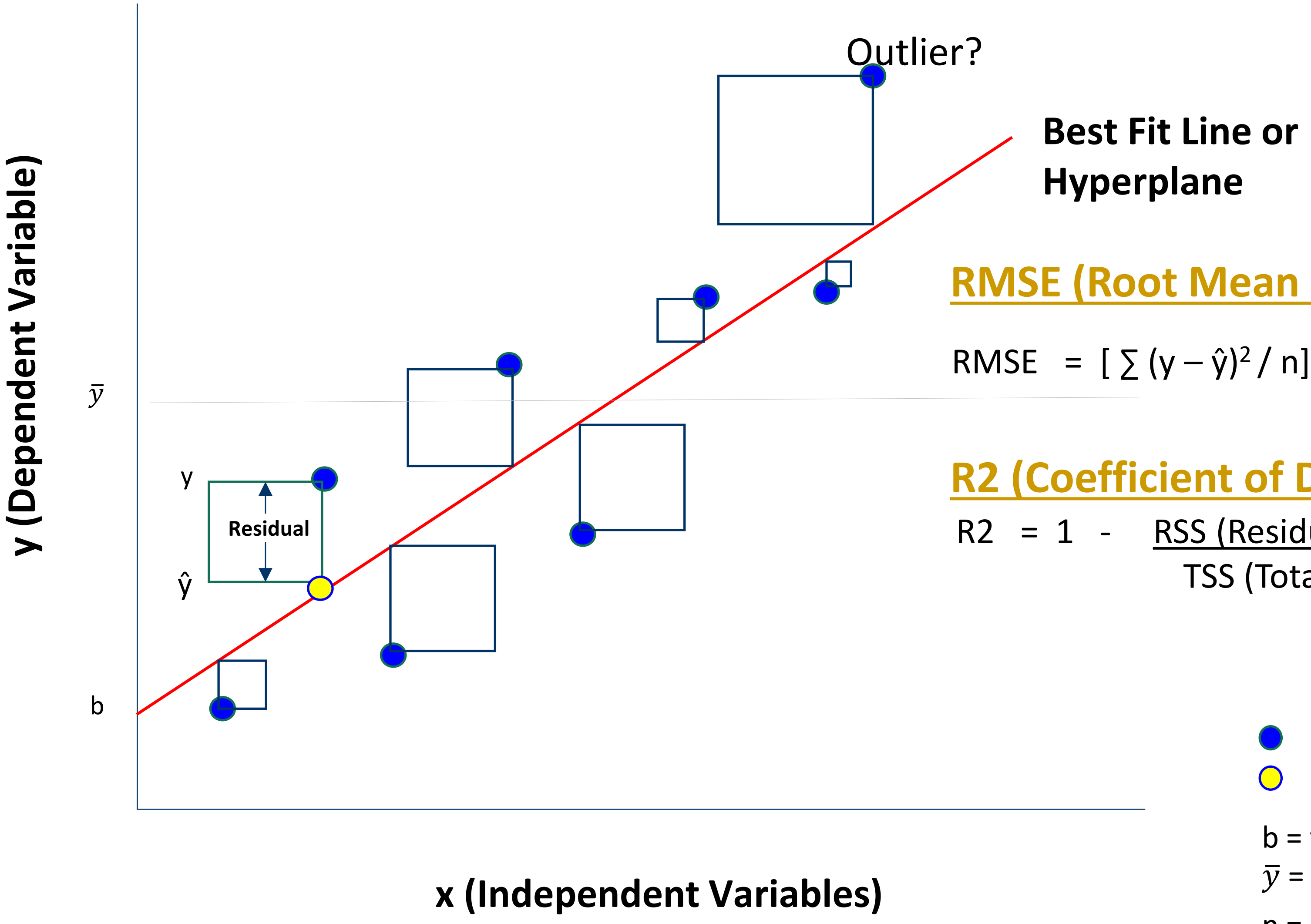
Regression – Find\Learn Pattern to Predict Numerical Values

Regression Intuition

Learning Observations are Points in n-Dimensional Vector Space



Model Performance



Cost Function will Minimize Error

RMSE (Root Mean Squared Error)

RMSE $= [\sum (y - \hat{y})^2 / n]^{½}$

R2 (Coefficient of Determination)

$$R2 = 1 - \frac{\text{RSS (Residual Sum of Squares)}}{\text{TSS (Total Sum of Squares)}}$$

Observation (x_n, y_n)

$\hat{y}$ = prediction, h_θ

b = y intercept, bias

$\bar{y}$ = y mean

n = number of points,

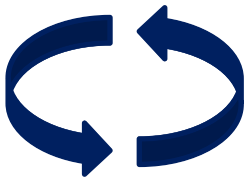
Training Data

Predictors

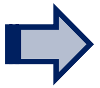
Learning Target

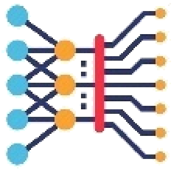
EC_mpy	CP_Off_Change	Dist_Road_Agg	DOC	n_AGRIVALUE	Nominal_OD	Pipe_Coating	Pipe_Seam	st_SOIL_cor
	All	All		All	All	All	All	All
[illegible]	-0.03	Limited_Impact	24.00	0.00	30.00	ASPHALT_ENAMEL	DSAW	High
2.05	-0.01	Out_Of_Range	30.00	0.00	30.00	TGF_E	DSAW	Moderate
2.00	-0.02	Out_Of_Range	24.00	0.00	30.00	TGF_E	DSAW	Moderate
2.00	-0.03	Limited_Impact	24.00	0.00	30.00	TGF_E	DSAW	Moderate
2.00	-0.03	Out_Of_Range	60.00	0.00	18.00	TGF_A	DSAW	High
2.00	-0.03	Out_Of_Range	63.00	0.00	18.00	TGF_A	DSAW	High
1.90	-0.07	Potential_Impact	24.00	0.00	30.00	TGF_E	DSAW	Moderate
1.85	-0.03	Out_Of_Range	18.00	0.00	30.00	TGF_E	DSAW	Moderate
1.80	0.02	Limited_Impact	24.00	0.00	18.00	FBE	DSAW	High
1.75	-0.03	Limited_Impact	24.00	0.00	30.00	TGF_H	DSAW	Moderate
1.70	-0.07	Limited_Impact	24.00	0.00	30.00	TGF_E	DSAW	Moderate
1.70	-0.07	Limited_Impact	30.00	0.00	30.00	TGF_E	DSAW	Moderate
1.70	-0.08	Potential_Impact	24.00	0.00	30.00	TGF_E	DSAW	Moderate
1.70	-0.08	Out_Of_Range	24.00	0.00	30.00	TGF_E	DSAW	Moderate
1.70	-0.07	Out_Of_Range	24.00	0.00	30.00	TGF_E	DSAW	Moderate

Training Data



Machine Learning Process



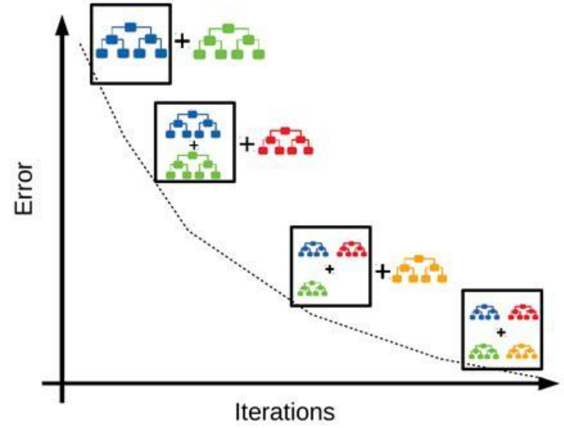


Learned Model

2025 BANFF PIPELINE WORKSHOP

Learned Model

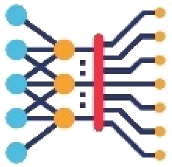
Gradient Boosted Tree



Tuned Hyper-Parameters

mtry	26
trees	960
min_n	4
tree_depth	14
learn_rate	0.0013
loss_reduction	0
sample_size	0.2247
stop_iter	5

Model



Performance

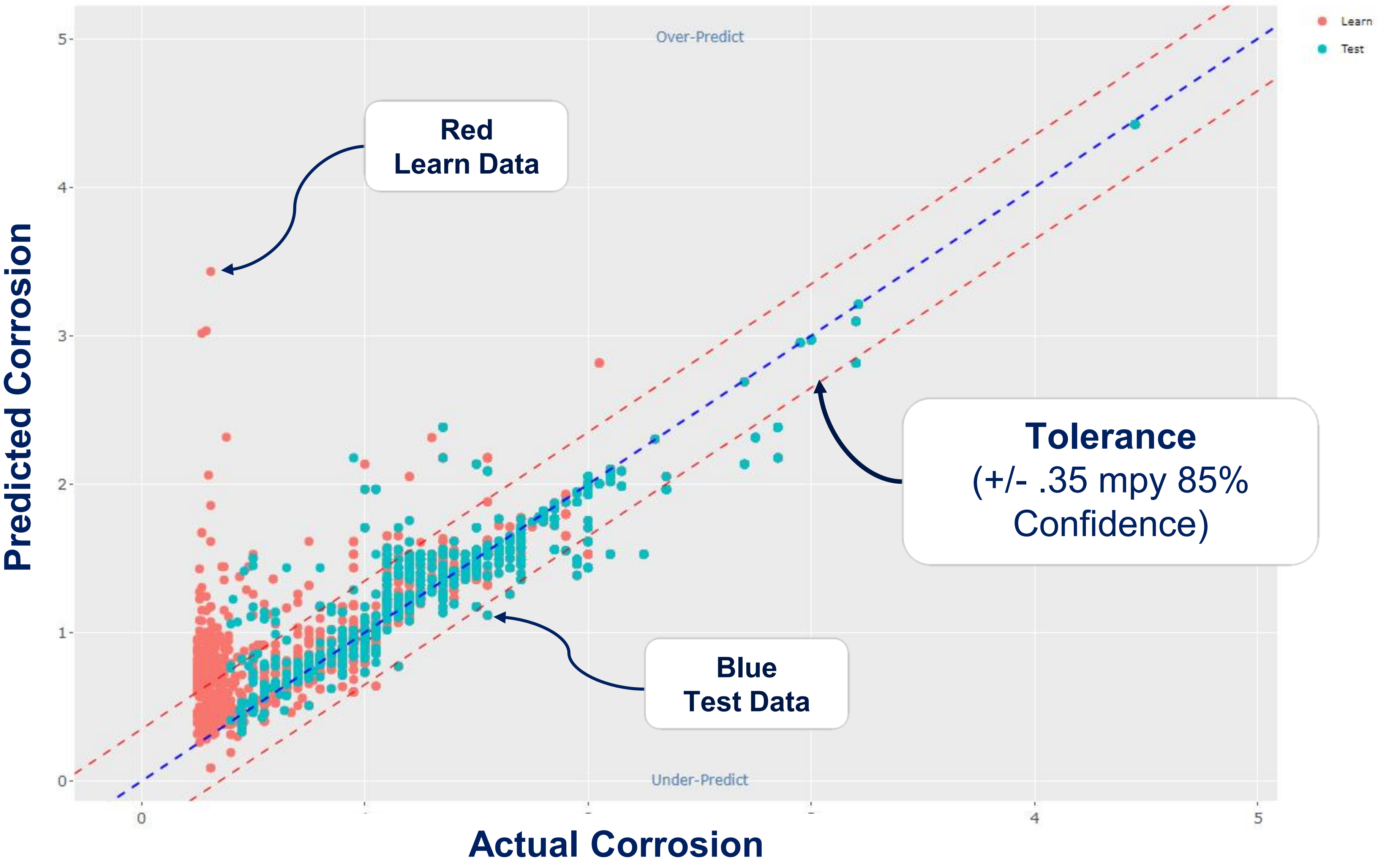
rmse	0.4634
rsq	0.8903
mae	0.3338

Test Data

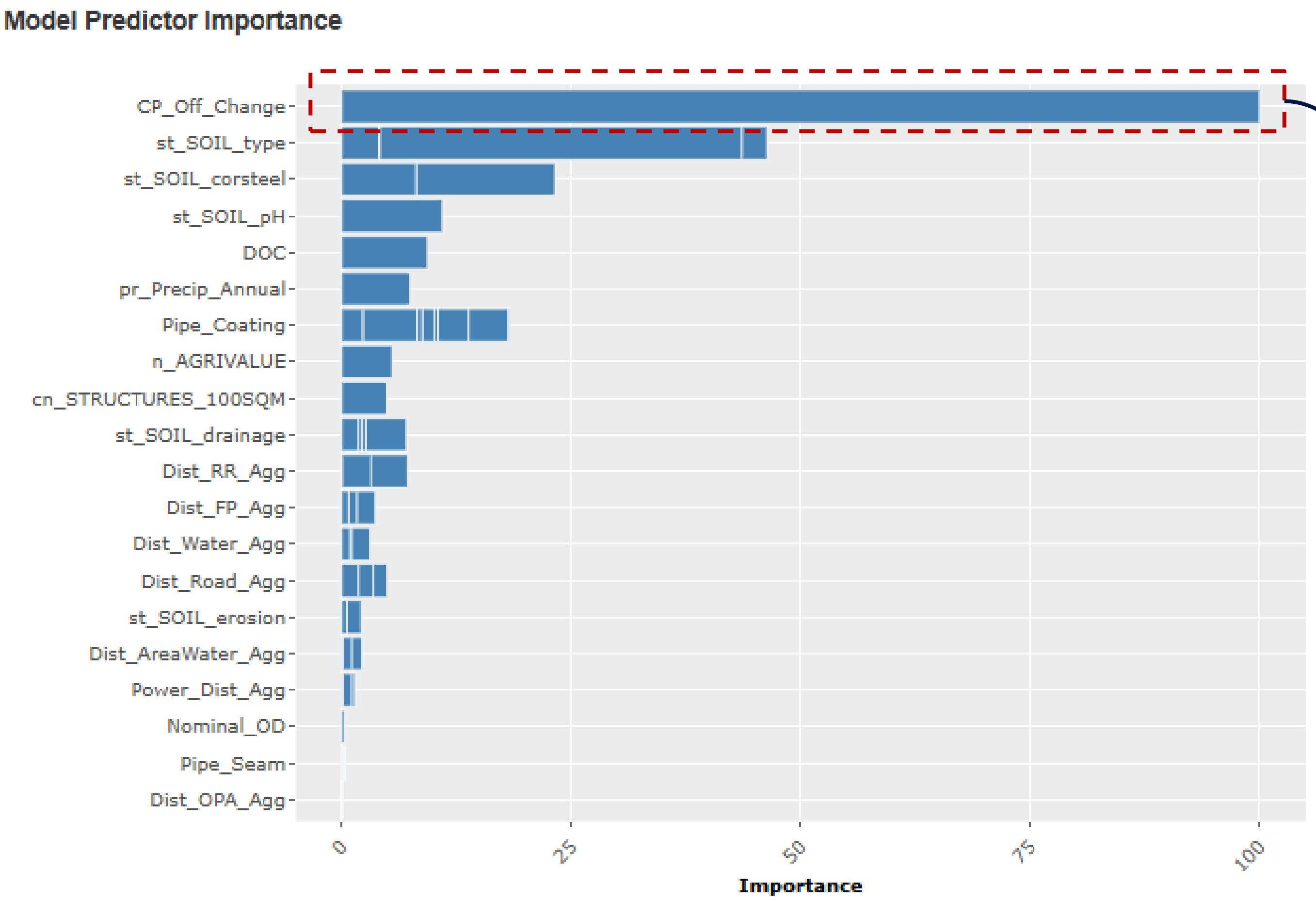
Model Performance

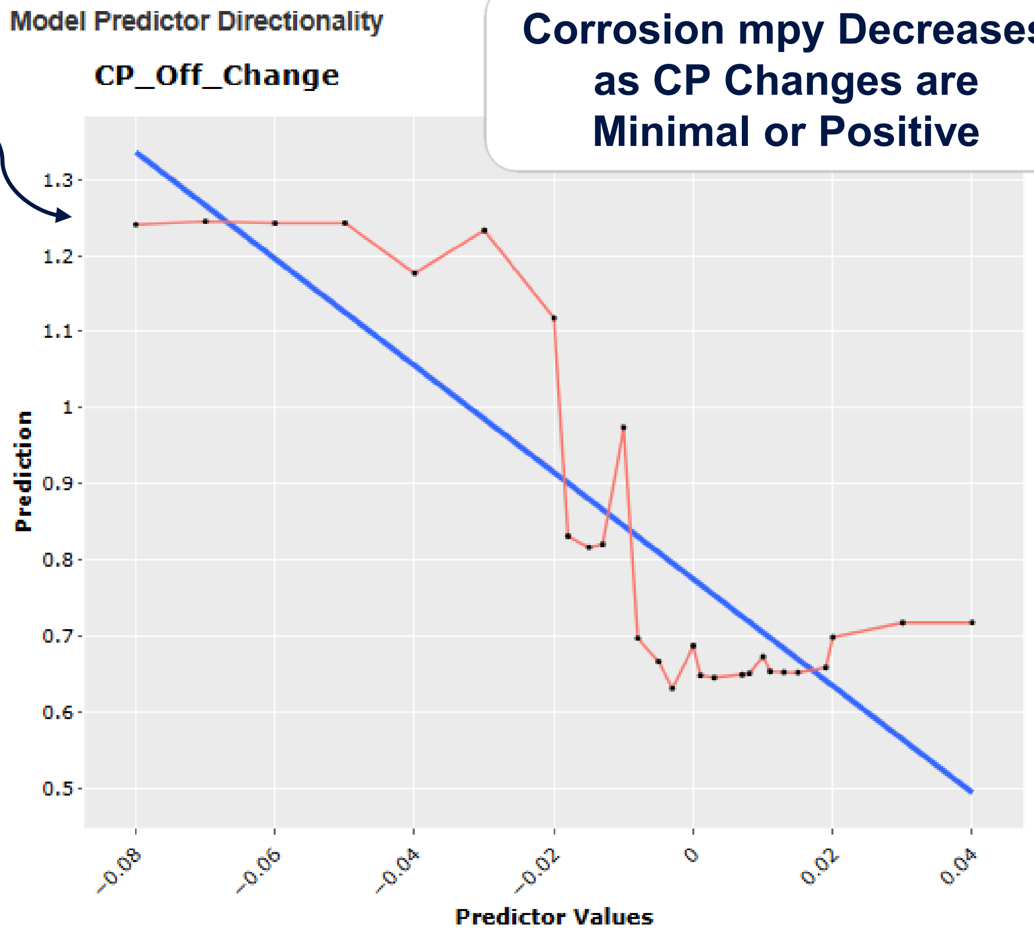
Metrics

- RMSE
- R2
- MAE



Model Weights

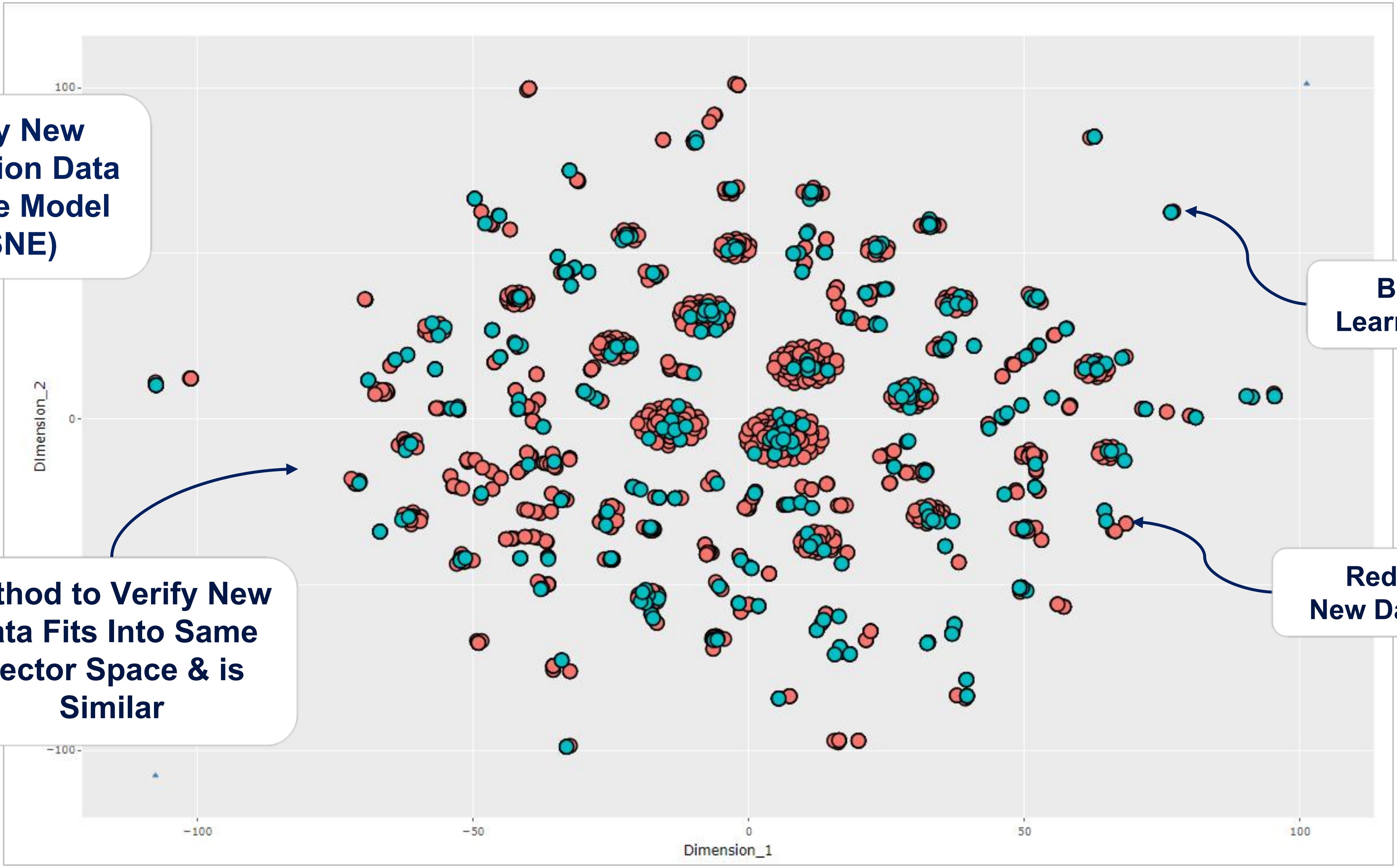




Corrosion mpy Decreases as CP Changes are Minimal or Positive

Model Applicability

Verify New Prediction Data Fits the Model (t-SNE)



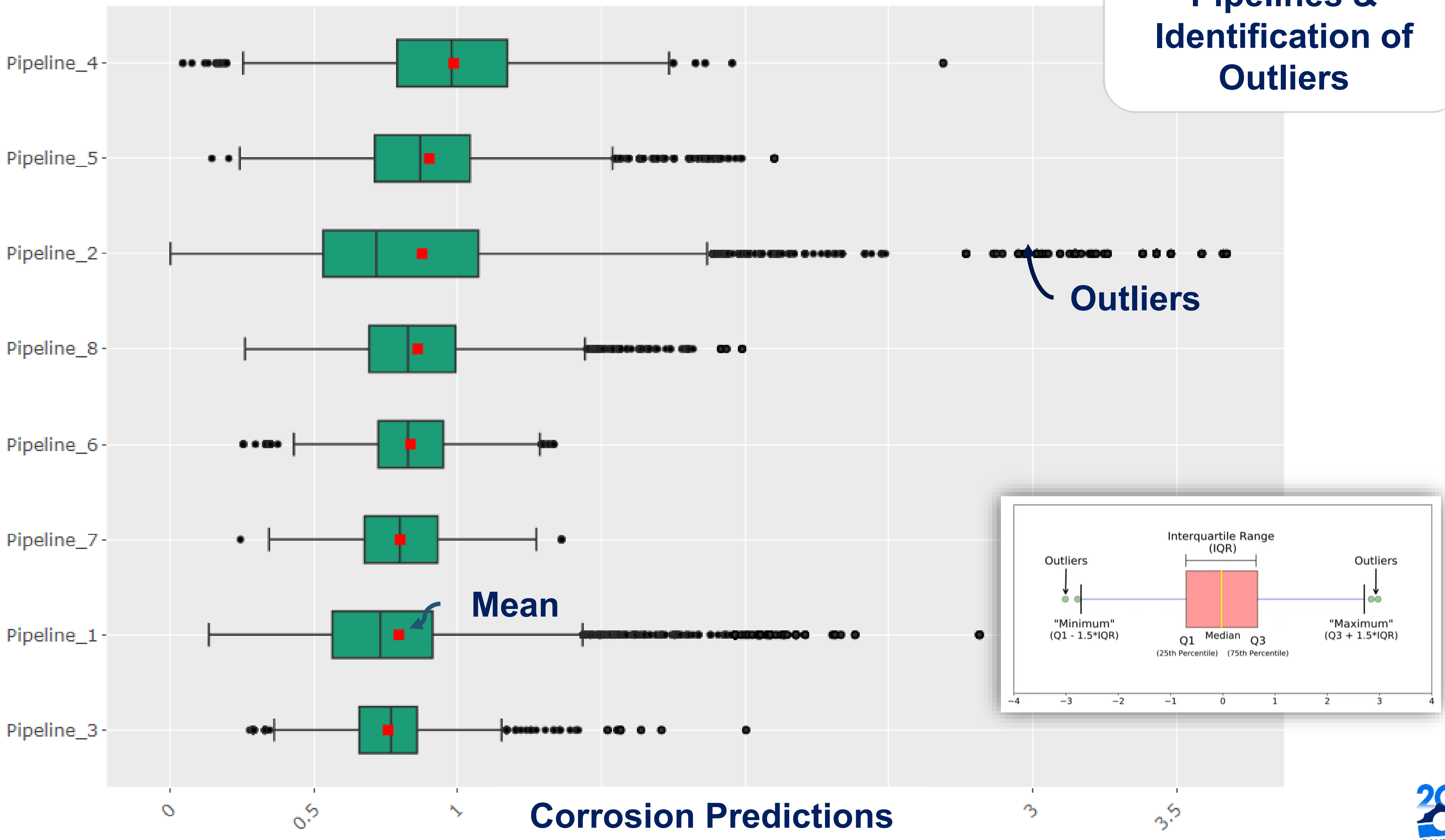
Blue Learn Data

Red New Data

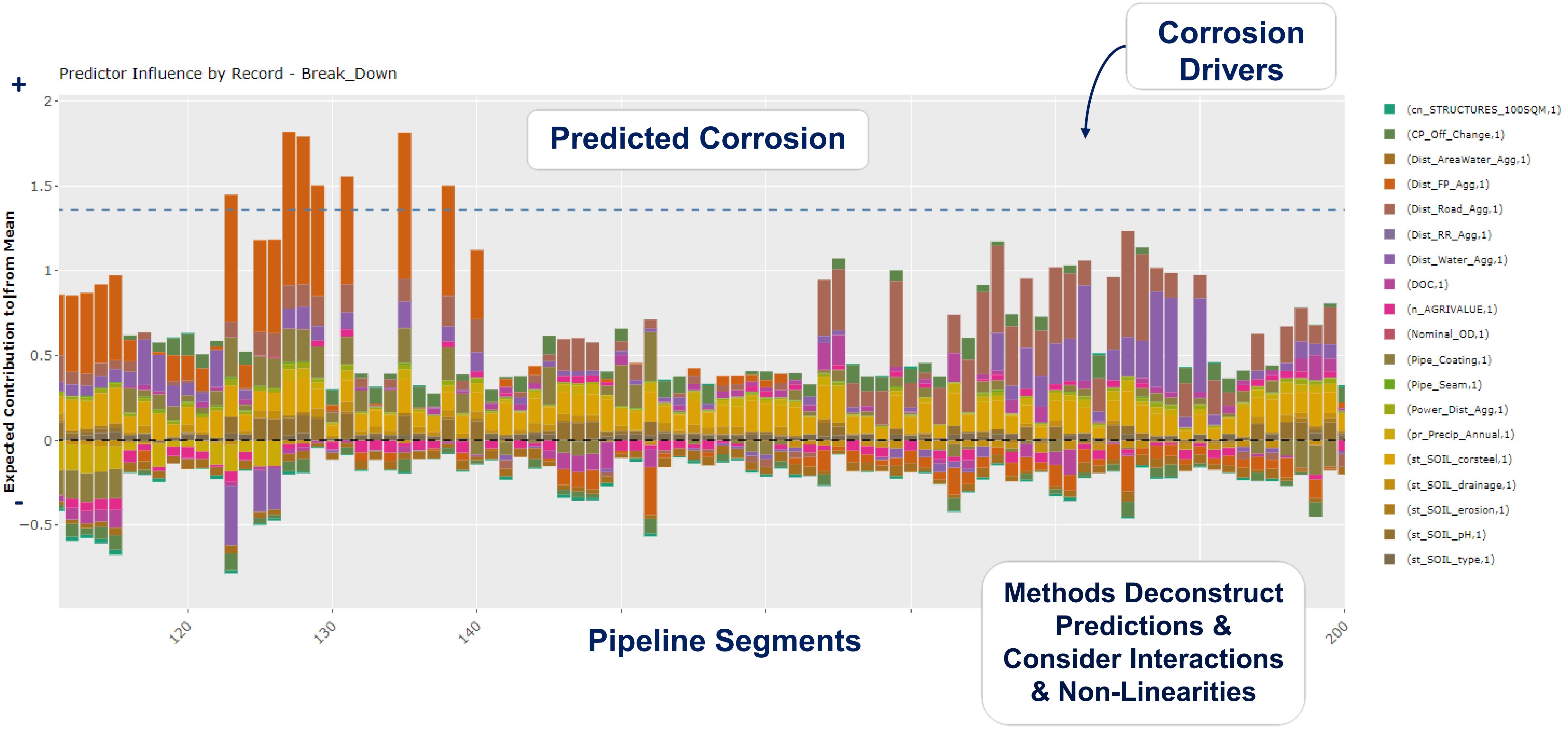
Method to Verify New Data Fits Into Same Vector Space & is Similar

Prediction Results

Prioritization of Pipelines & Identification of Outliers

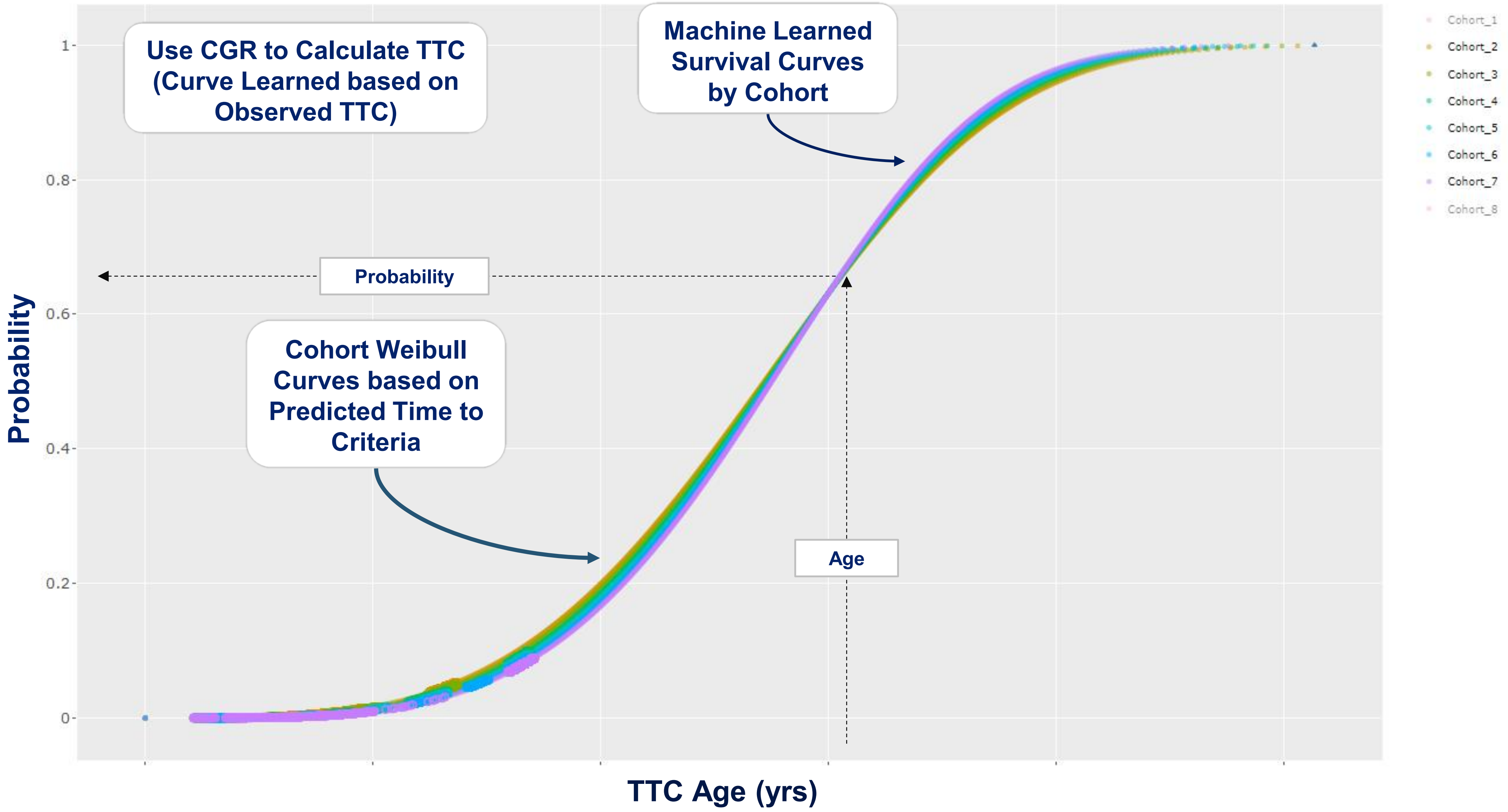


Results Explanations



CGR to Probability

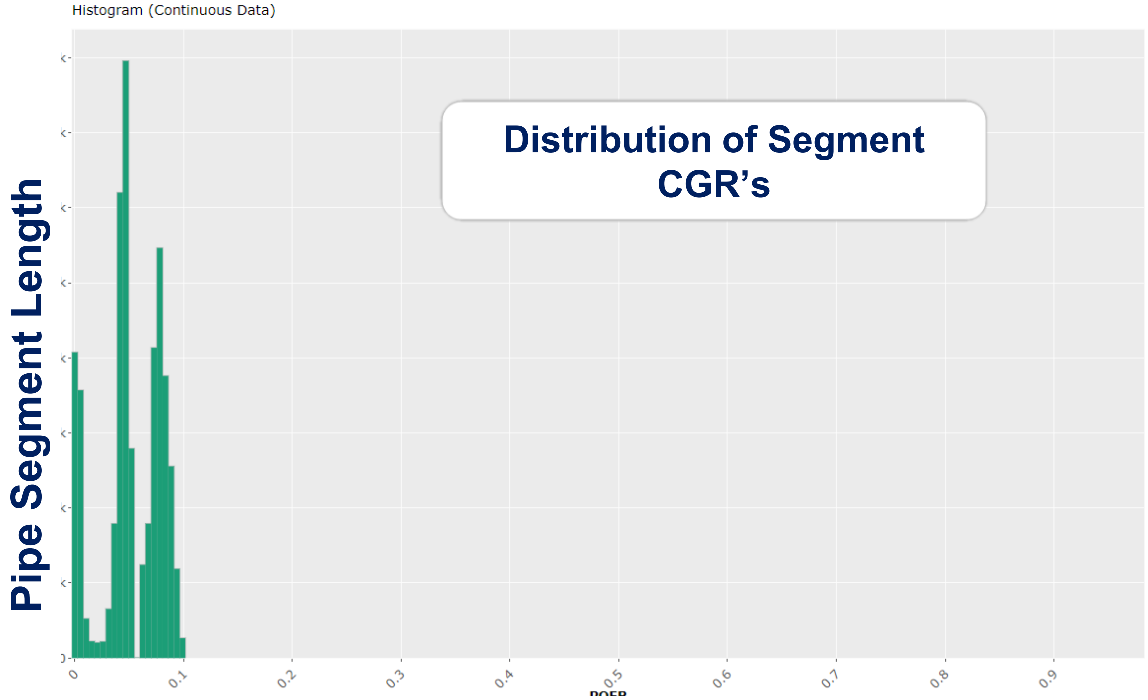
Microsoft Excel Worksheet

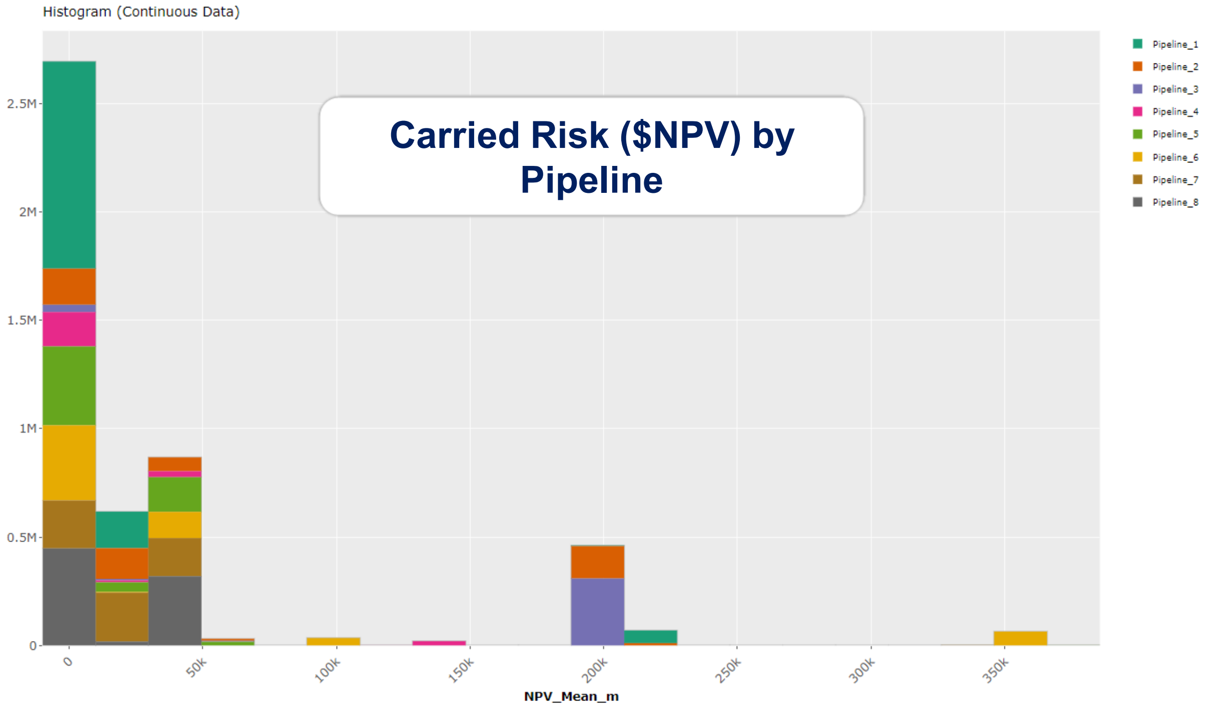


QRA – Regression Time Dependent

NPV(Σ [Probability x Event Rate (#/length-yrs) x Event Cost ($/Event)], Yrs)

- Learned Weibull Reliability Curves to convert Time to Criteria to Probability
- Normalize to Incident & Consequence Event Distributions (P50\P99)





Probability of External Corrosion →

Monetized Carried Risk by Pipeline →

Machine Learned Results

QRA Monetized Risk

2025 BANFF PIPELINE WORKSHOP

Questions?

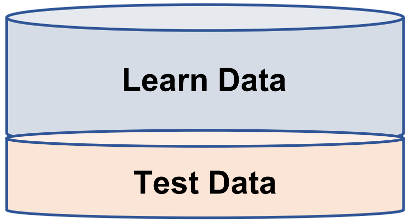
Cathodic Protection Potentials

Time Series Model Example

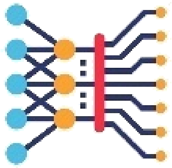
Machine Learning Process

Learning Target
(EC CGR)

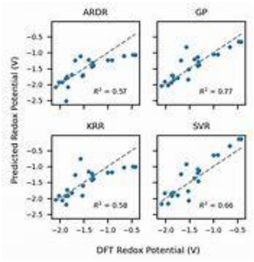
Training Data
(Observations)



Learned Model
(Methods, Tuning)



Performance & Insights
(Validation & Acceptance)



Time Series – Find\Learn Pattern to Predict Numerical Values
Considers Seasonality Effects Using Slices of Time

2025 BANFF PIPELINE WORKSHOP

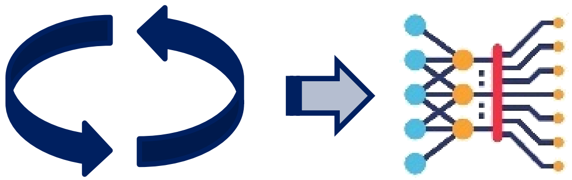
Training Data

Predictors

Learning Target

CP_Read	Date	Pipe_Age	pr_Temp_Diff	Precip	st_SOIL_corsteel	st_SOIL_type	TS
	2015-0	All	All	All	All	All	All
80	2015-09-01	60.00	23.00	30.00	High	Sand	3300
921.10	2015-09-01	77.00	23.00	30.00	High	Sand	28386
919.40	2015-09-01	77.00	23.00	30.00	High	Loam	22280
918.90	2016-09-01	77.00	22.00	30.00	High	Sand	30020
917.90	2016-09-01	59.00	23.00	30.00	High	Sand	1499
917.50	2015-09-01	60.00	23.00	30.00	High	Sand	5122
917.20	2016-09-01	77.00	22.00	30.00	High	Sand	30917
916.90	2015-09-01	77.00	22.00	30.00	High	Sand	30020
916.20	2017-09-01	21.00	22.00	30.00	High	Clay	31511
915.90	2016-09-01	77.00	22.00	30.00	High	Clay	21137
915.30	2017-09-01	77.00	23.00	30.00	High	Sand	32118
914.90	2016-09-01	33.00	23.00	30.00	High	Sand	26501
914.60	2017-09-01	59.00	23.00	30.00	High	Sand	1588
914.50	2015-09-01	59.00	23.00	30.00	High	Sand	2277
914.50	2017-09-01	77.00	23.00	30.00	High	Sand	24686
914.40	2016-09-01	77.00	23.00	30.00	High	Clay	24470

Training Data

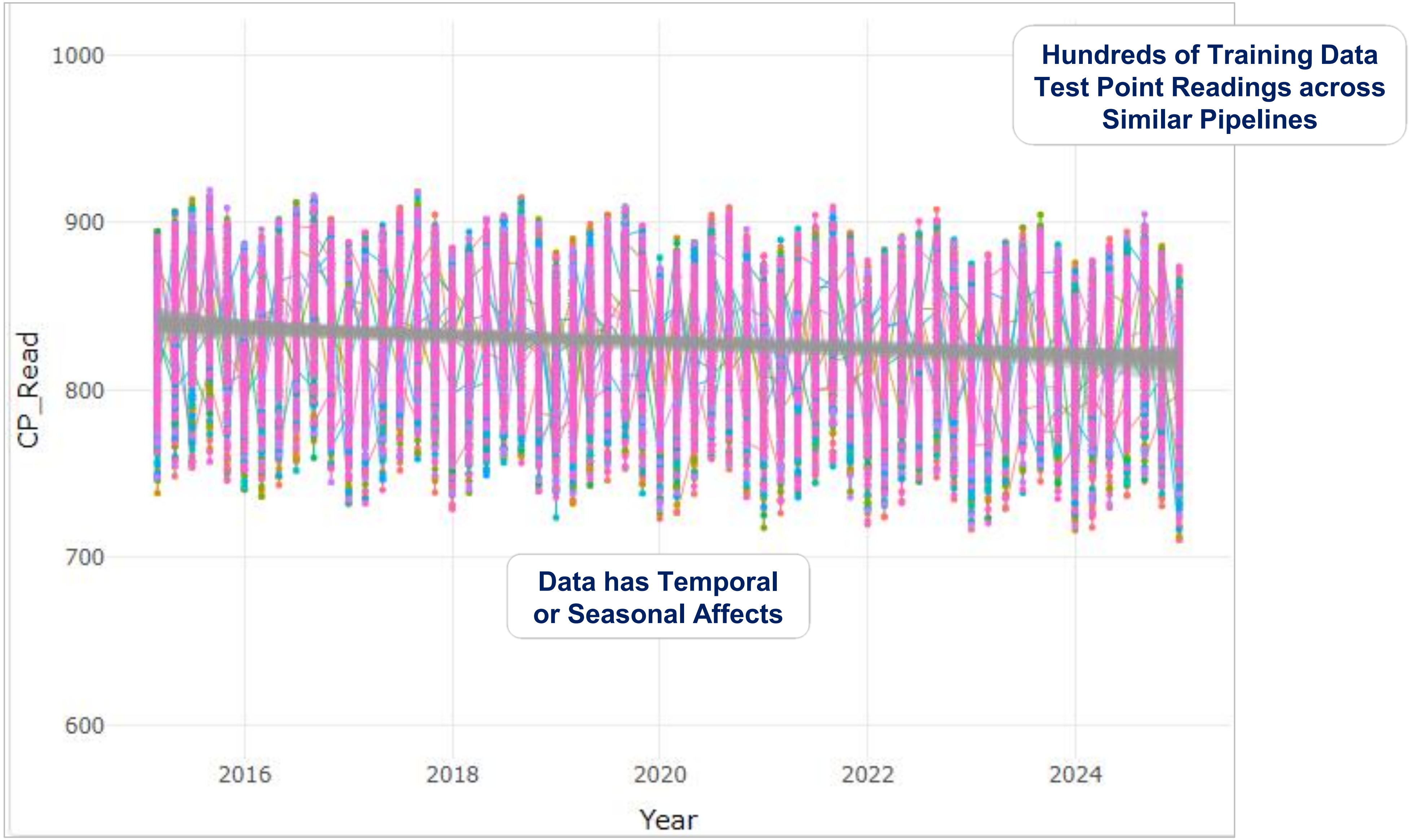


Machine Learning Process

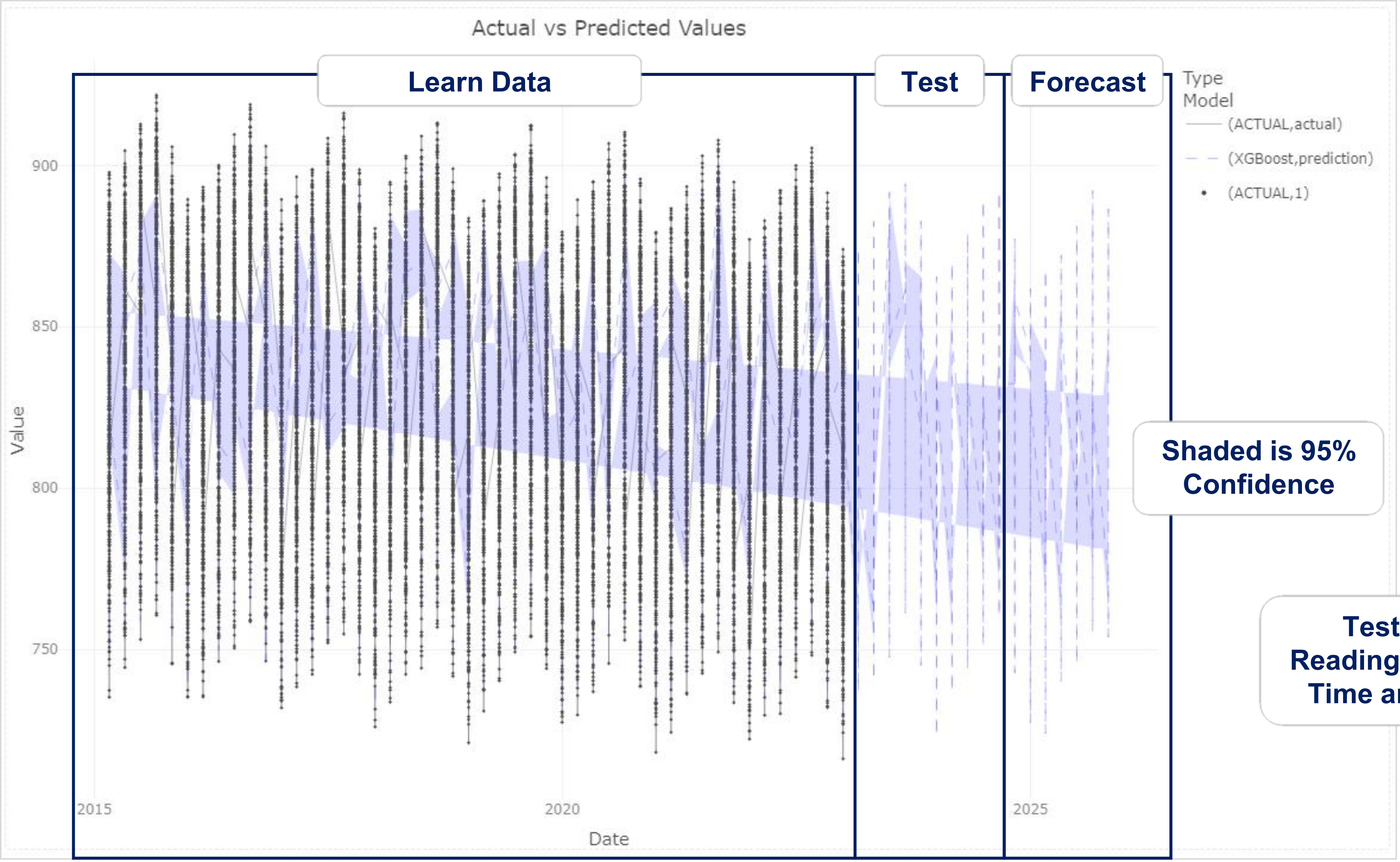
Learned Model

2025 BANFF PIPELINE WORKSHOP

Training Data



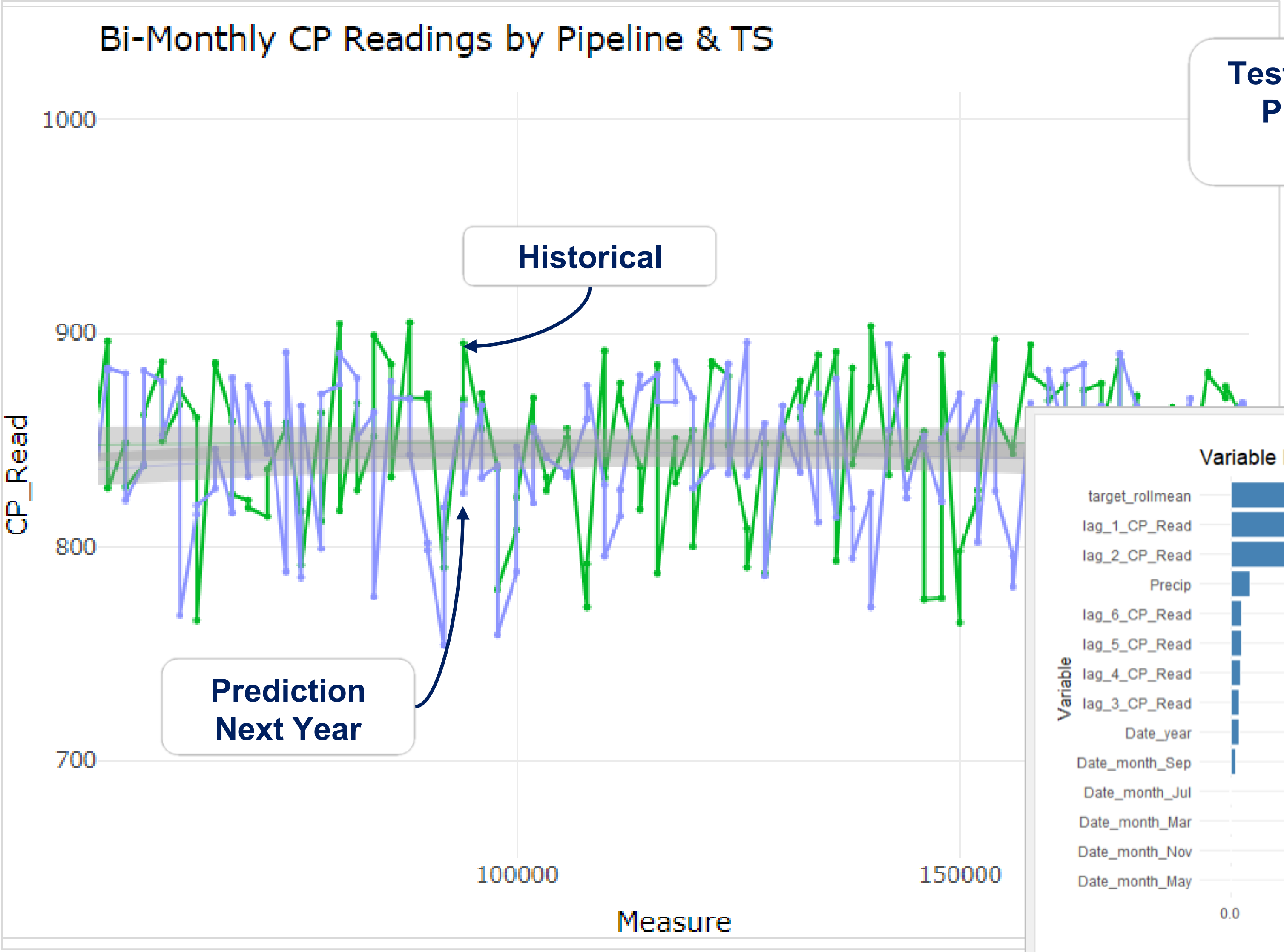
Model Application & Results – All Test Stations



Shaded is 95% Confidence

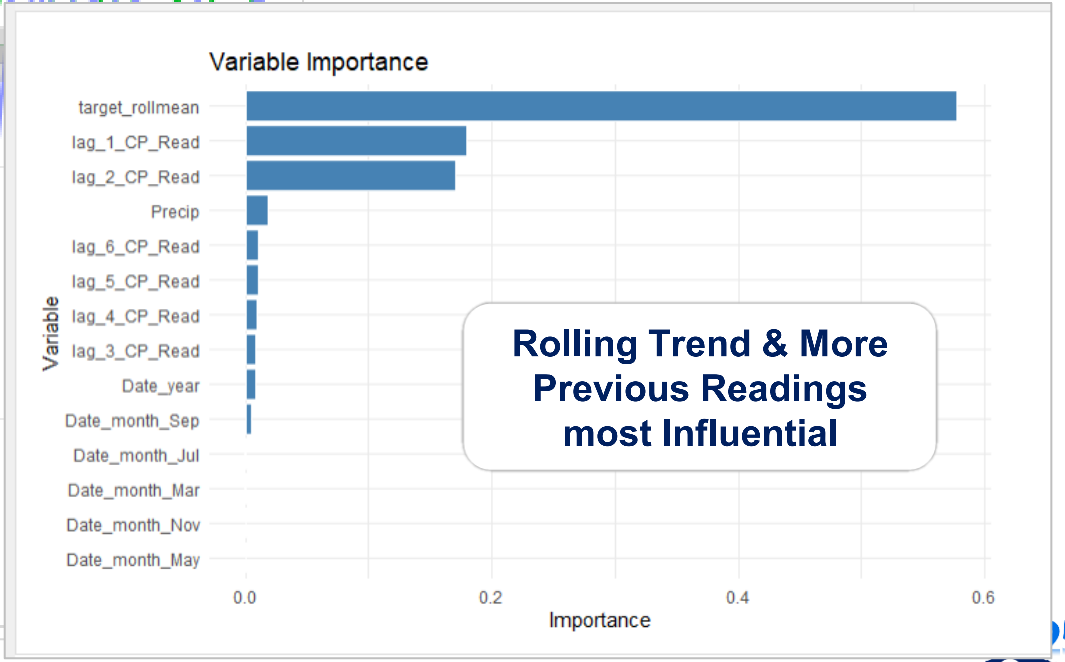
Test Station Readings Vary Over Time and Season

Model Application & Results - Pipeline



Test Point CP Read Predictions vs. Historical

Predictor Influence



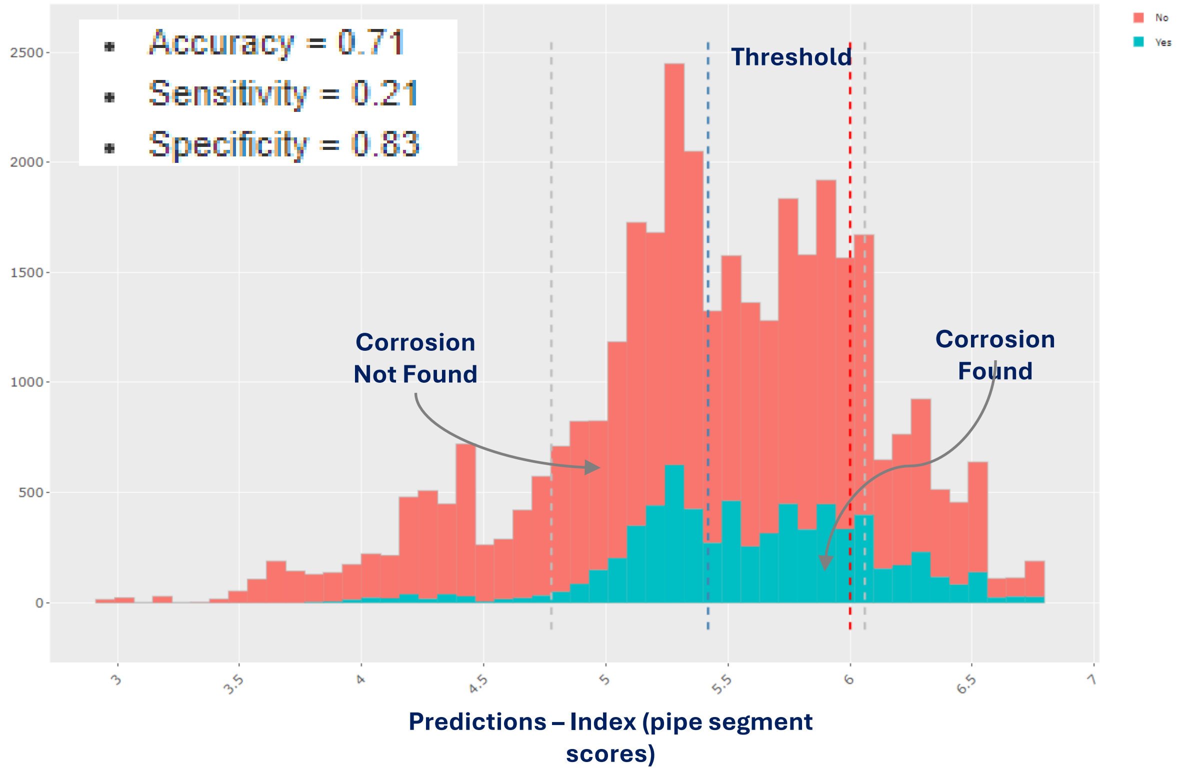
Questions?

Deterministic Model Validation

2025
BANFF PIPELINE WORKSHOP

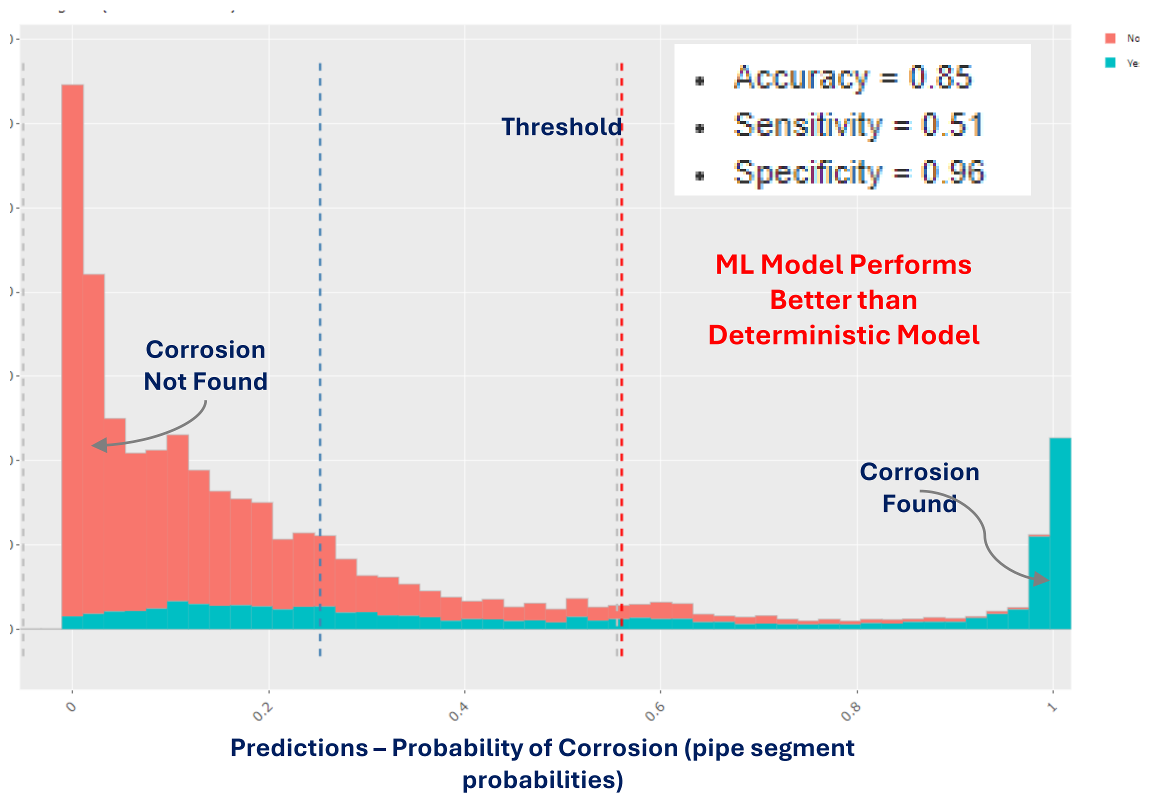
Model Validation

Deterministic Model



Model Learned based on Deterministic Structure

Machine Learned Model



Model Learned with Observational Data

Questions?

Machine Learning based Integrity & Risk Management

- **Data Driven**
- **Validated**
- **Transparent**
- **Explainable**
- **Versatile**

Machine Learning & Risk (TU9)

2025 BANFF PIPELINE WORKSHOP

Michael Gloven, PE
michaelgloven@pipeline-risk.com