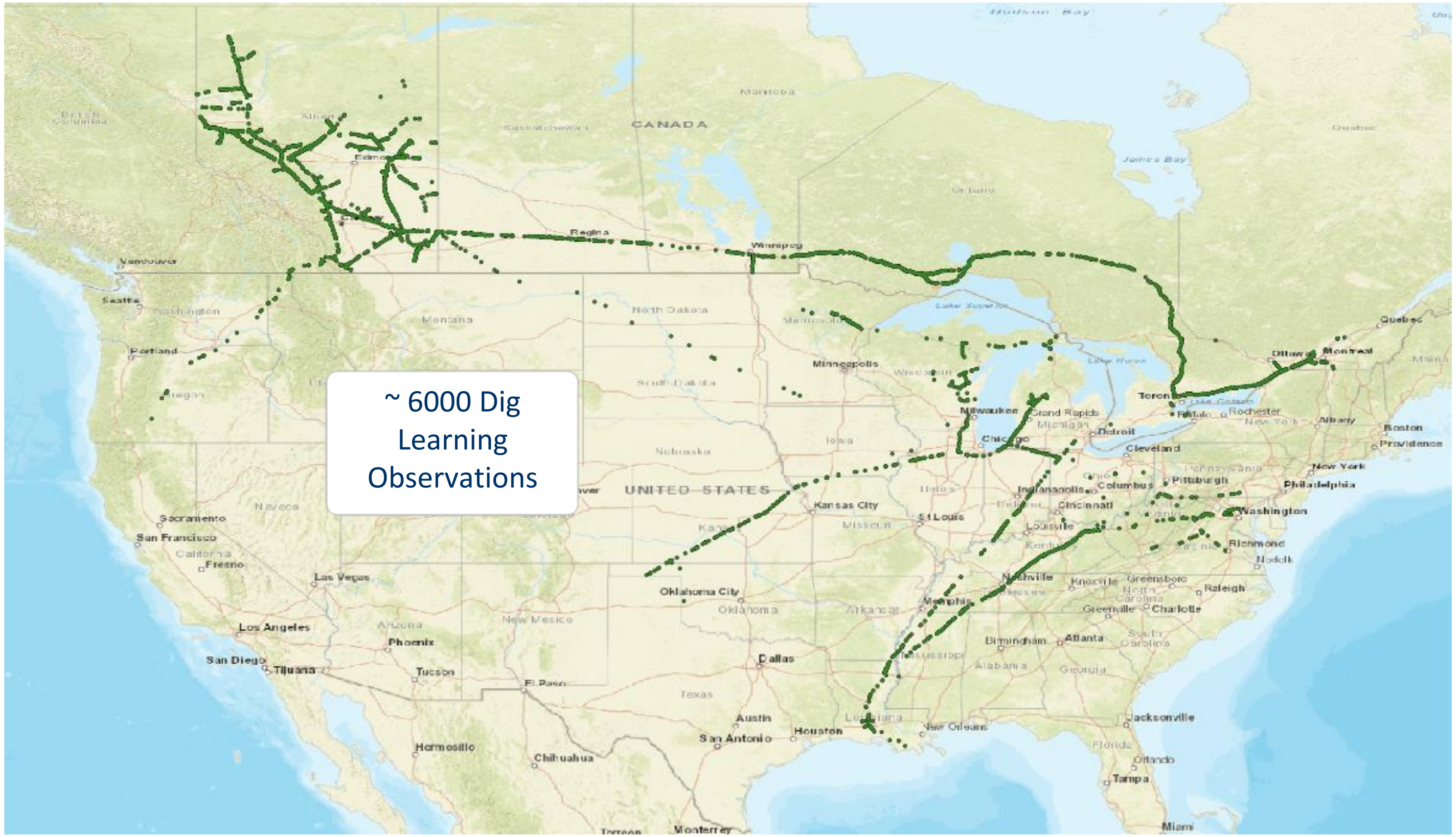
From Digs to Data: Integrating ILI and Environmental Insights for SCC Predictive Modelling

Michael Gloven, PLR

Co-authors
Syed Aijaz, TC Energy

Objectives

- Leverage ~6000 Dig Observations Augmented with MFL & Environmental Data to Improve the Assessment of SCC
- Use Machine Learning & Advanced Analytics to Support this Assessment
- Use the Results to Plan Inspections, Select DA Dig Sites, Perform Sensitivity Analysis and Augment Determinsitic Models
- Present Key Elements & Learning of this Process



Machine Learning & Advanced Analytics Process

Learning Target
(SCC True\False)

- 6000 Learning Observations (True\False)

Training Data
(Observations)

Learn Data

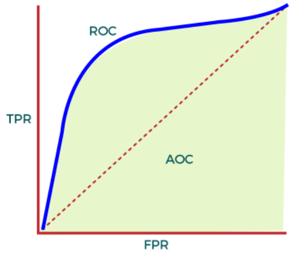
Test Data

- Pipe Inspection Data
- MFL Data
- Soils Data
- Weather Data

Learned Model
(Find Patterns)

- Learned Model is Basis of Advanced Analytics

Performance & Analytics
(Validation & Explanations)

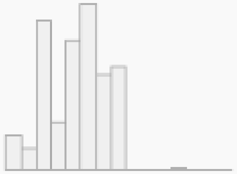
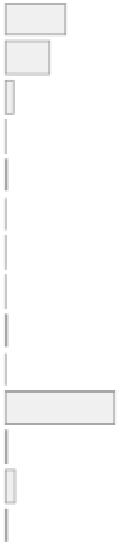
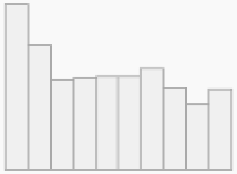


- Predictor Influence
- Predictor Directionality
- Prediction Breakdowns
- Validation
- Clustering & PCA
- Similarity Testing
- Application

Training Data – Pipe Inspections

Pipe Inspections

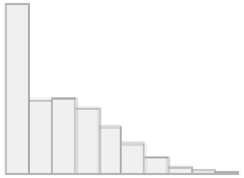
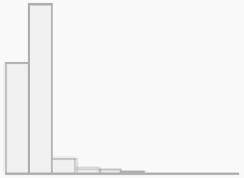
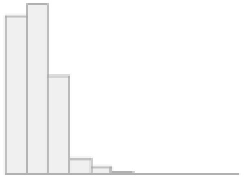
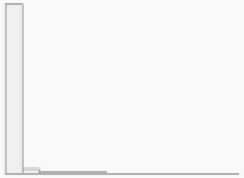
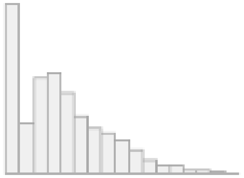
- Install Date
- Coating Types
- Normalized Distance from Launcher
- SCC Indications

Joint_Install_Date [numeric]	Mean (sd) : 1968.856 (9.821) min ≤ med ≤ max: 1946 ≤ 1970 ≤ 2017 IQR (CV) : 18 (0.005)	56 distinct values		6154 (94.9%)	332 (5.1%)
af_coating_type [factor]	1. Asphalt 2. Coal Tar 3. Epoxy (general) 4. Extruded Poly. 5. FBE 6. Liquid Epoxy 7. Mastic 8. Multi Liquid 9. No_data 10. Paint 11. Poly Tape 12. PVC Tape 13. Wax 14. Wax_Dearborn	1620 (25.0%) 1191 (18.4%) 246 (3.8%) 27 (0.4%) 39 (0.6%) 16 (0.2%) 3 (0.0%) 6 (0.1%) 50 (0.8%) 4 (0.1%) 2897 (44.7%) 69 (1.1%) 266 (4.1%) 52 (0.8%)		6486 (100.0%)	0 (0.0%)
dist_from_launcher_normalized [numeric]	Mean (sd) : 0.444 (0.296) min ≤ med ≤ max: 0 ≤ 0.43 ≤ 1 IQR (CV) : 0.52 (0.668)	101 distinct values		6412 (98.9%)	74 (1.1%)
joint_has_SCC [factor]	1. FALSE 2. TRUE	4022 (62.0%) 2464 (38.0%)		6486 (100.0%)	0 (0.0%)

Training Data – MFL

Engineered MFL Corrosion Features

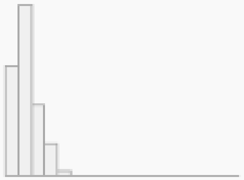
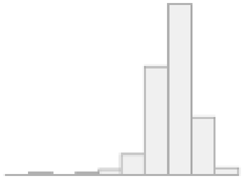
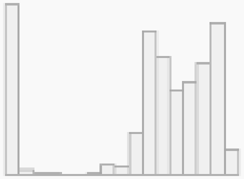
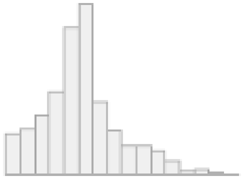
- Local (L0) vs. Generalized (L2) Corrosion
- Local (L0) vs. Generalized (L2) Severity

L0_ml_feature_count [numeric]	Mean (sd) : 54.866 (145.819) min ≤ med ≤ max: 0 ≤ 5 ≤ 1954 IQR (CV) : 40 (2.658)	494 distinct values		6485 (100.0%)	1 (0.0%)
L0_ml_max_depth [numeric]	Mean (sd) : 23.13 (21.624) min ≤ med ≤ max: 0 ≤ 21 ≤ 100 IQR (CV) : 38 (0.935)	742 distinct values		6467 (99.7%)	19 (0.3%)
L0_ml_median_depth [numeric]	Mean (sd) : 10.673 (10.008) min ≤ med ≤ max: 0 ≤ 12 ≤ 100 IQR (CV) : 14.45 (0.938)	623 distinct values		6467 (99.7%)	19 (0.3%)
L2_median_ml_depth [numeric]	Mean (sd) : 7.17 (5.526) min ≤ med ≤ max: 0 ≤ 7.51 ≤ 51.684 IQR (CV) : 7.364 (0.771)	1878 distinct values		5499 (84.8%)	987 (15.2%)
L2_ml_feature_count [numeric]	Mean (sd) : 669.357 (2351.287) min ≤ med ≤ max: 0 ≤ 28 ≤ 26504 IQR (CV) : 219 (3.513)	1140 distinct values		5499 (84.8%)	987 (15.2%)
L2_ml_max_depth [numeric]	Mean (sd) : 20.05 (16.75) min ≤ med ≤ max: 0 ≤ 17.674 ≤ 85 IQR (CV) : 22.645 (0.835)	2009 distinct values		5499 (84.8%)	987 (15.2%)

Training Data - Soils

Normalized Soils Data

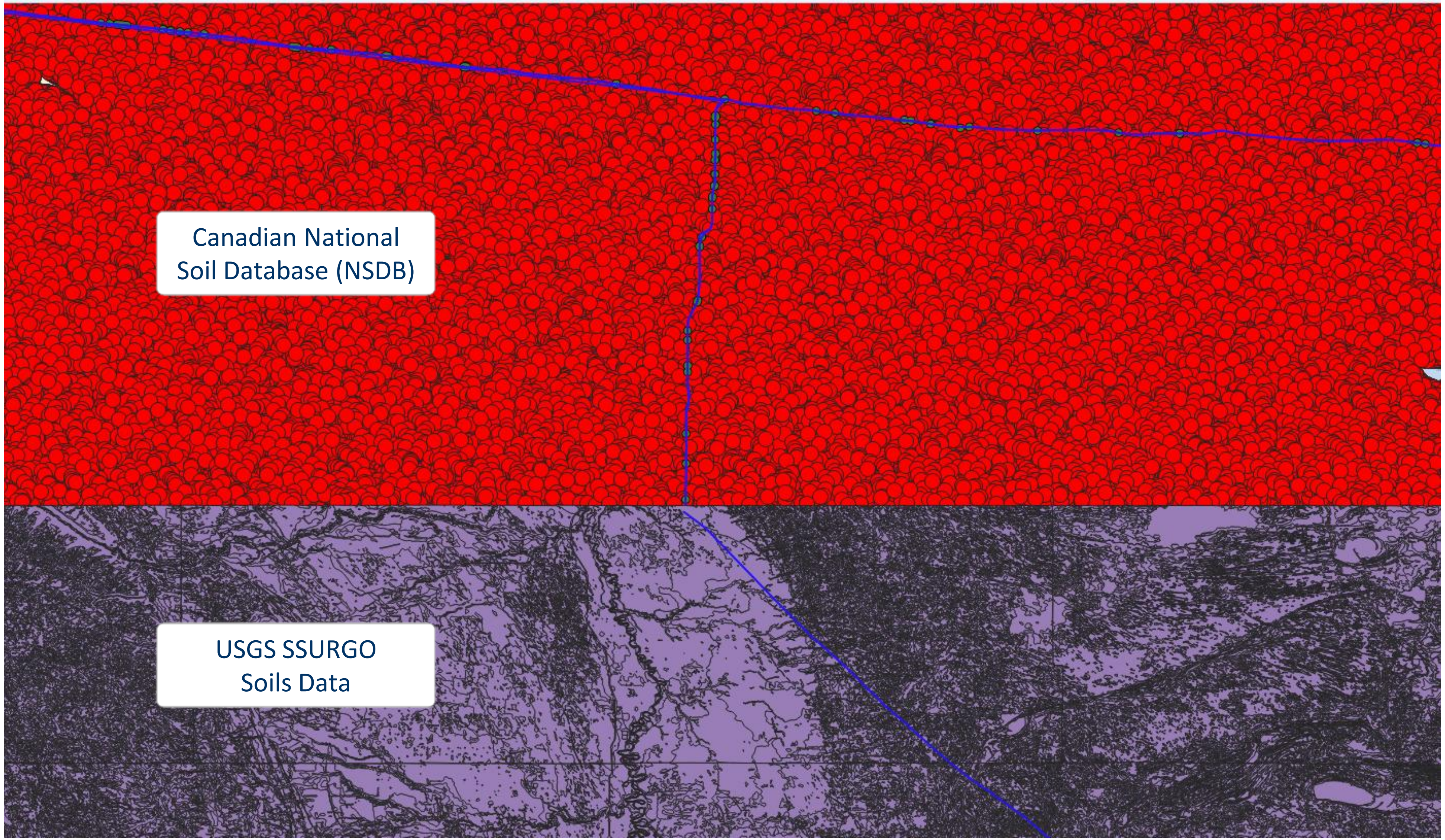
- Soil Landscape Grids of Canada (100 m resolution)
- SSURGO - Avg of Horizons (1-10 acres, ~1000ft resolution)
- Required Normalization of Values

Soil_Cations [numeric]	Mean (sd) : 16.401 (12.994) min ≤ med ≤ max: 0 ≤ 15.026 ≤ 174.533 IQR (CV) : 12.101 (0.792)	3955 distinct values		6166 (95.1%)	320 (4.9%)
Soil_Density [numeric]	Mean (sd) : 1.435 (0.204) min ≤ med ≤ max: 0 ≤ 1.447 ≤ 1.962 IQR (CV) : 0.196 (0.142)	3909 distinct values		6166 (95.1%)	320 (4.9%)
Soil_pH [numeric]	Mean (sd) : 5.227 (2.655) min ≤ med ≤ max: 0 ≤ 5.873 ≤ 8.325 IQR (CV) : 2.218 (0.508)	3619 distinct values		6166 (95.1%)	320 (4.9%)
Soil_percClay [numeric]	Mean (sd) : 26.072 (13.359) min ≤ med ≤ max: 0 ≤ 25.183 ≤ 76.7 IQR (CV) : 13.485 (0.512)	3867 distinct values		6166 (95.1%)	320 (4.9%)

Training Data - Soils

Canada

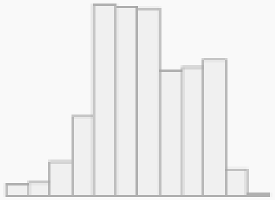
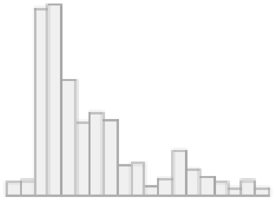
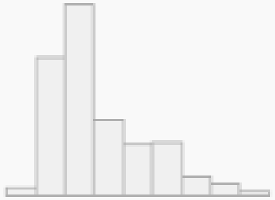
United States



Training Data - Weather

Normalized Weather Data

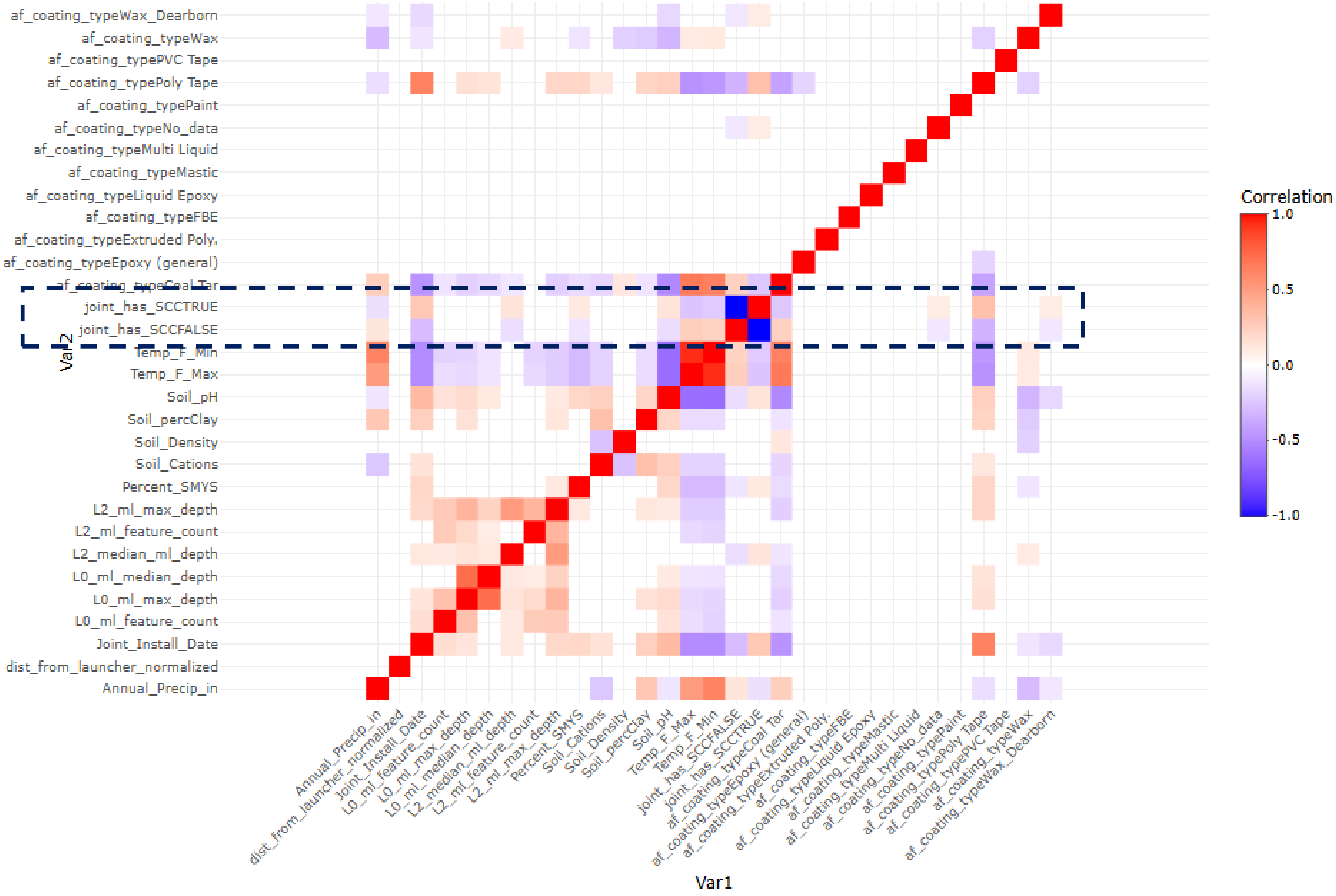
- Canadian Centre for Climate Services (DegC, mm)
- US PRISM Weather Data (DegF, Inches)
- Required Normalization of Values

Annual_Precip_in [numeric]	Mean (sd) : 42.045 (10.739) min ≤ med ≤ max: 10 ≤ 43 ≤ 67 IQR (CV) : 18 (0.255)	74 distinct values		1868 (28.8%)	4618 (71.2%)
Temp_F_Max [numeric]	Mean (sd) : 52.407 (8.209) min ≤ med ≤ max: 41.54 ≤ 49.28 ≤ 78 IQR (CV) : 8.82 (0.157)	104 distinct values		5393 (83.1%)	1093 (16.9%)
Temp_F_Min [numeric]	Mean (sd) : 30.703 (8.16) min ≤ med ≤ max: 19.04 ≤ 27.5 ≤ 59 IQR (CV) : 10.62 (0.266)	125 distinct values		5427 (83.7%)	1059 (16.3%)

Exploratory Data Analysis - Correlation

Correlation

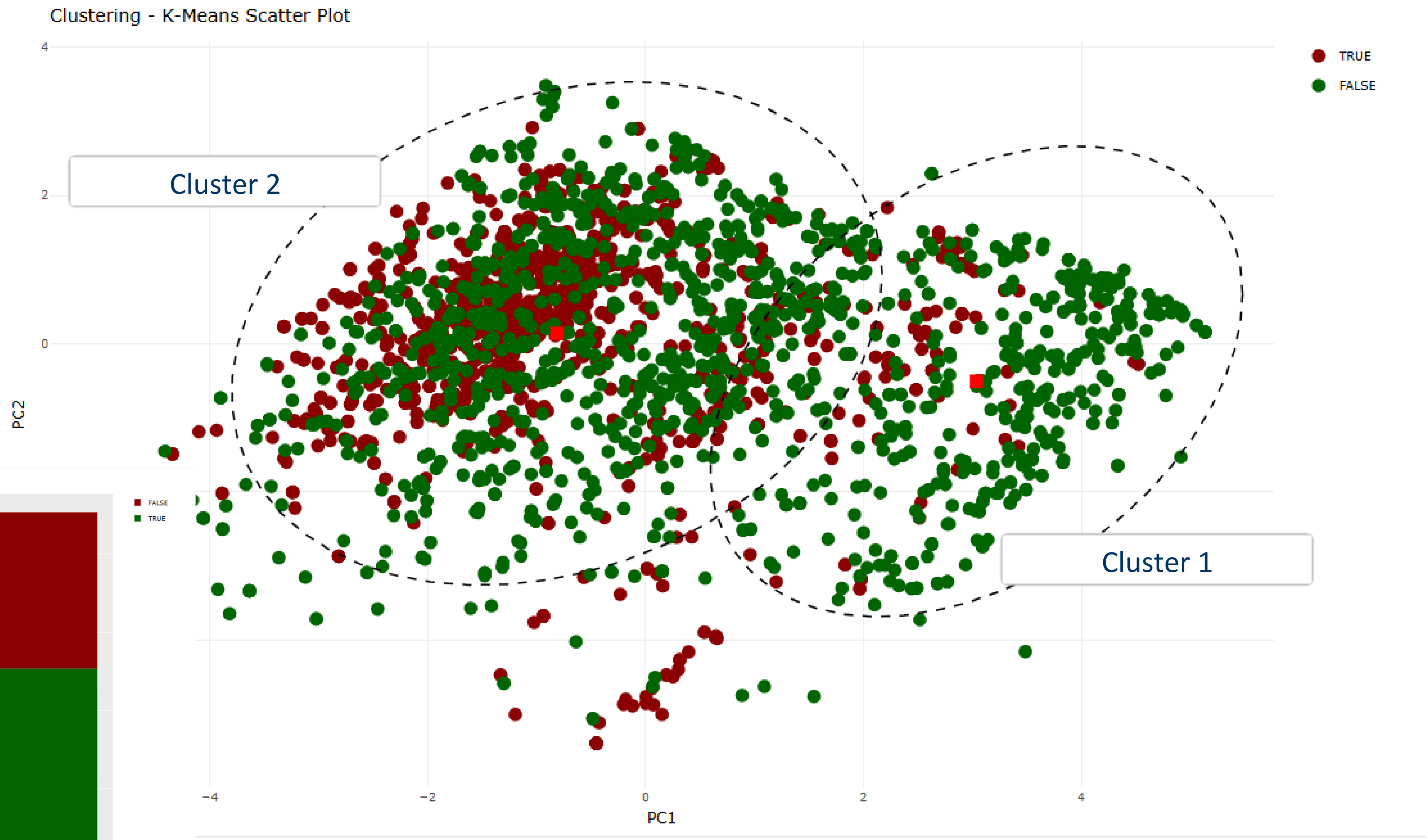
- Are there any Significant Pair-Wise Relationships with the Presence or Non-Presence of SCC?
- Correlations are Weak, therefore, Use Machine Learning to Reveal Multi-Variate Relationships and Non-Linearities

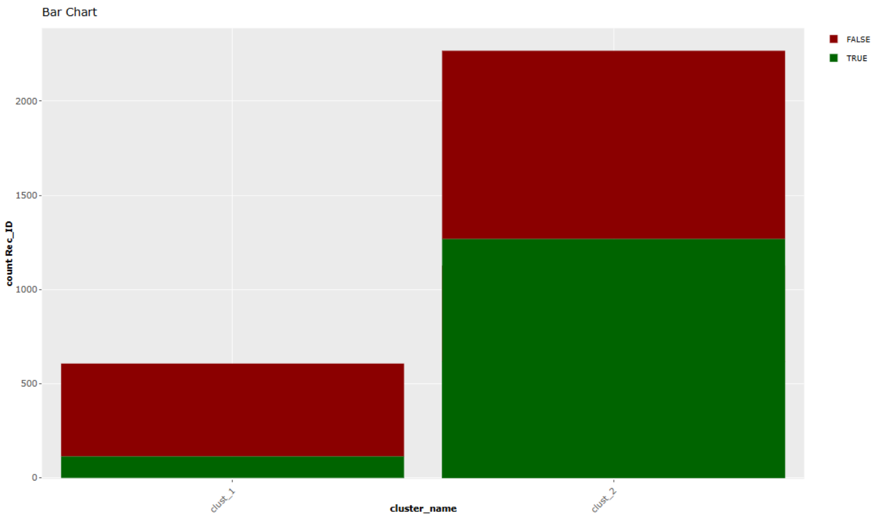


Unsupervised Learning – Clustering (Advanced Analytic Prior to Model Learning)

Clustering

- How Does the Training Data Naturally Cluster?
- Strong Delineations can Support a new Classification Model for Rare Threats

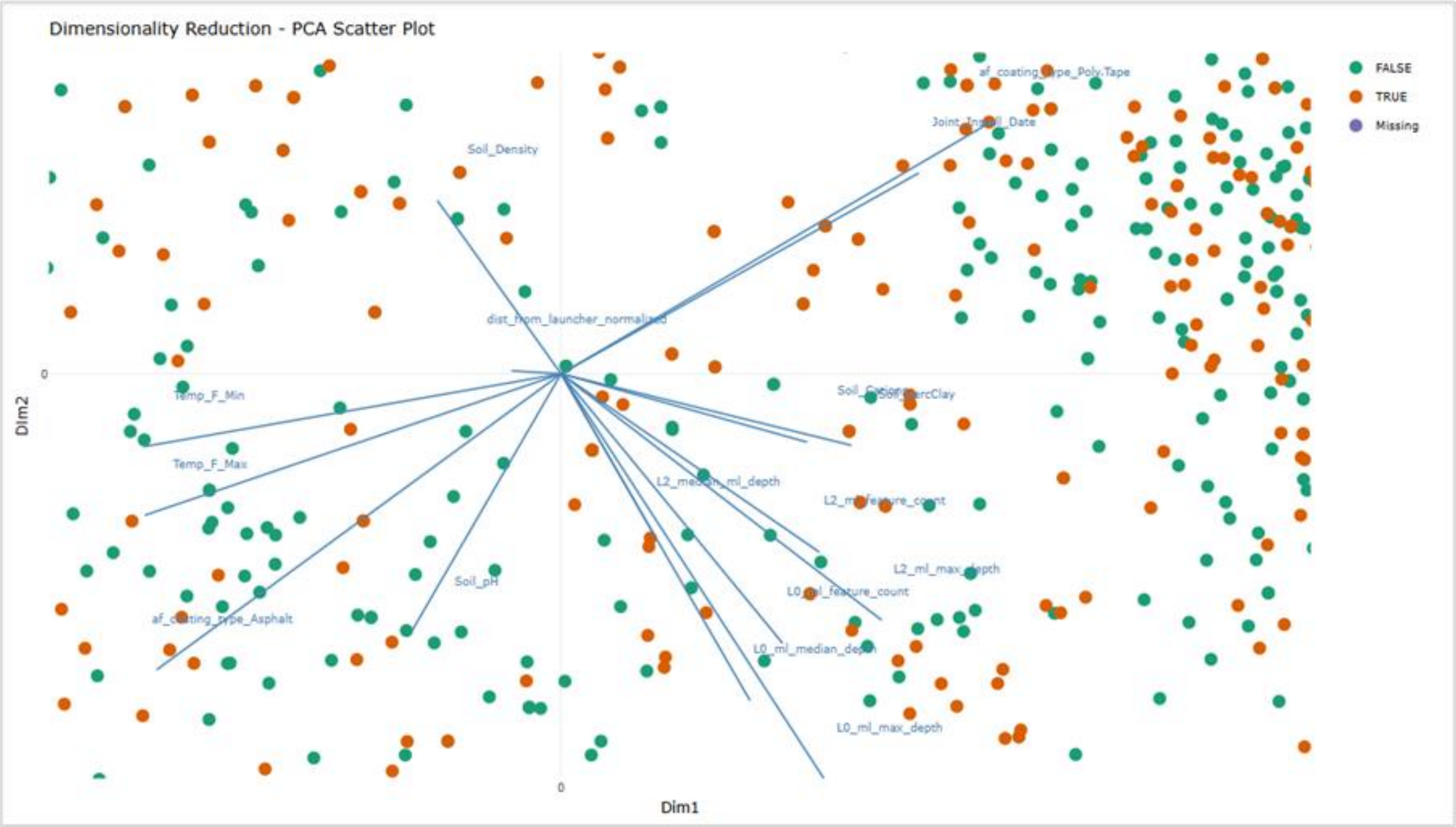




Unsupervised Learning – PCA (Advanced Analytic Prior to Model Learning)

Principal Component Analysis

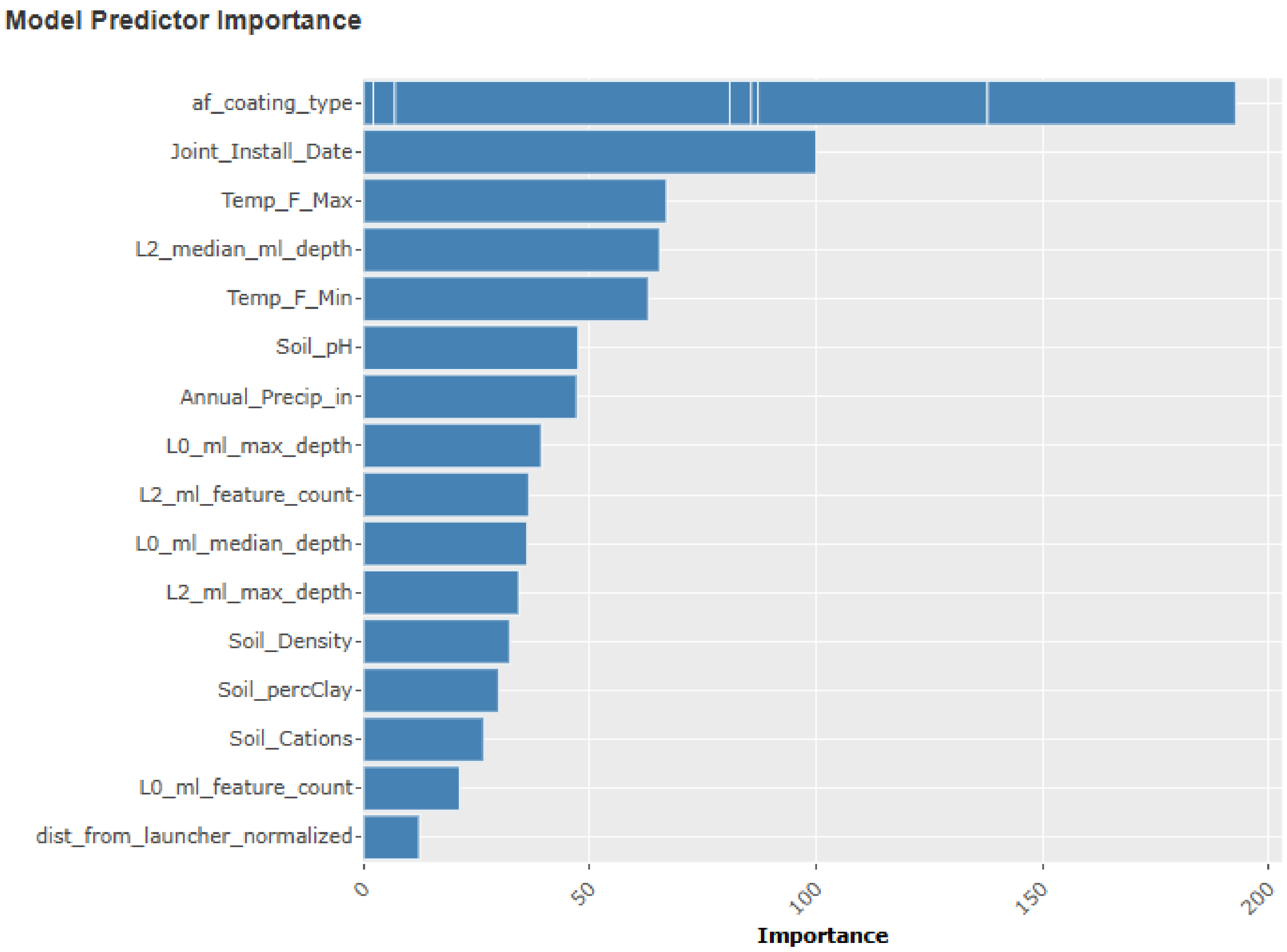
- Is there a Relationship between the Strength of Predictor Variation and the Presence or Non-Presence of SCC?
- PCA is Useful in Selection of Predictor Data



Learned Model – Predictor Influence

Predictor Importance

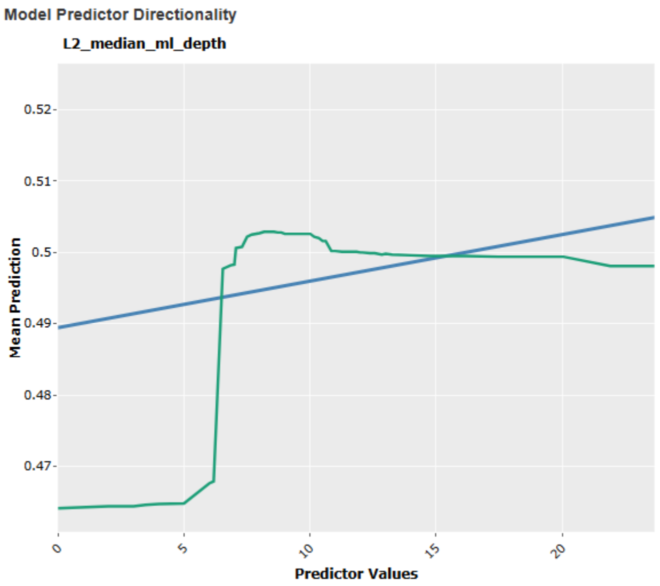
- Learn Model based on xgboost Method (Tune & Validate)
- Importance Values Derived from Model, Importance is Measured by the Ability of Predictor to Reduce the Entropy or Separate the Training Data in Consideration of the SCC True\False Target
- Results may be Proxy for “Value” of Data

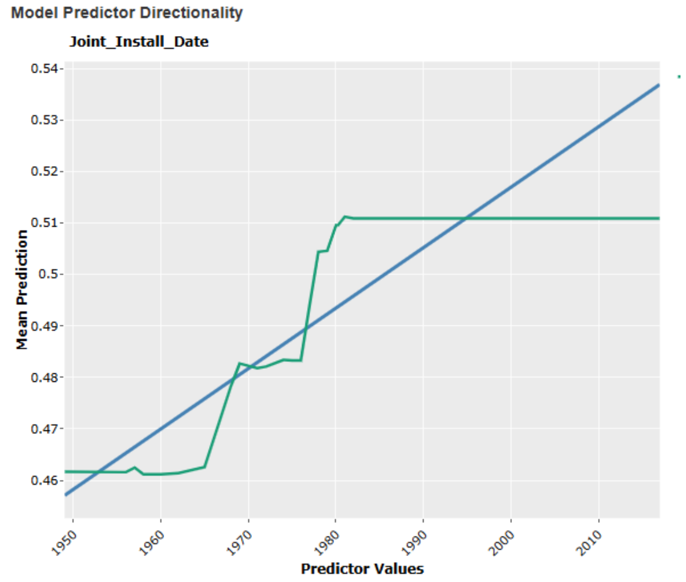


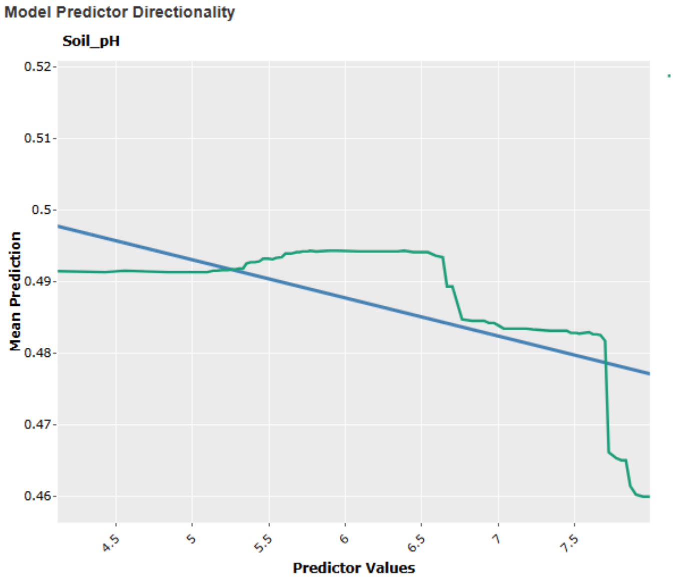
Learned Model – Partial Dependency Plots

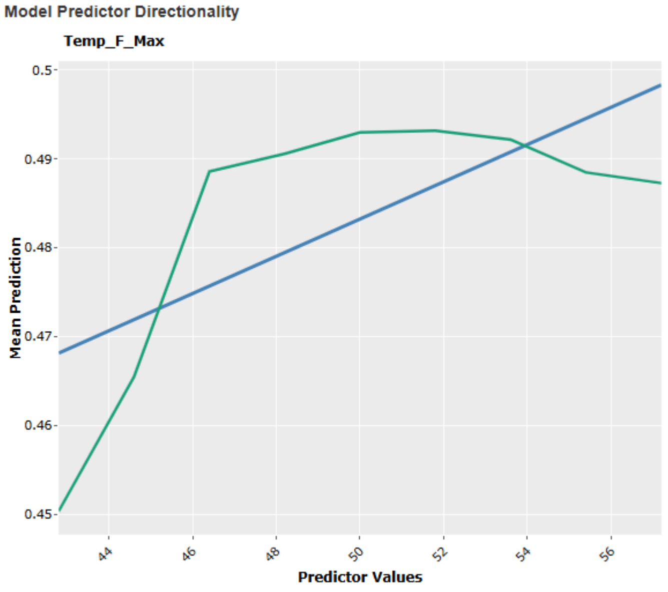
Predictor Dependency

- PDP's Plot the Variation of Predictions to Actual Predictor Values (ref. examples)
- Plots may be Used to Visualize Predictor Interactions and Non-Linearities
- Blue Line is Fit Line whereas Green Line is Average of Sampled Observations





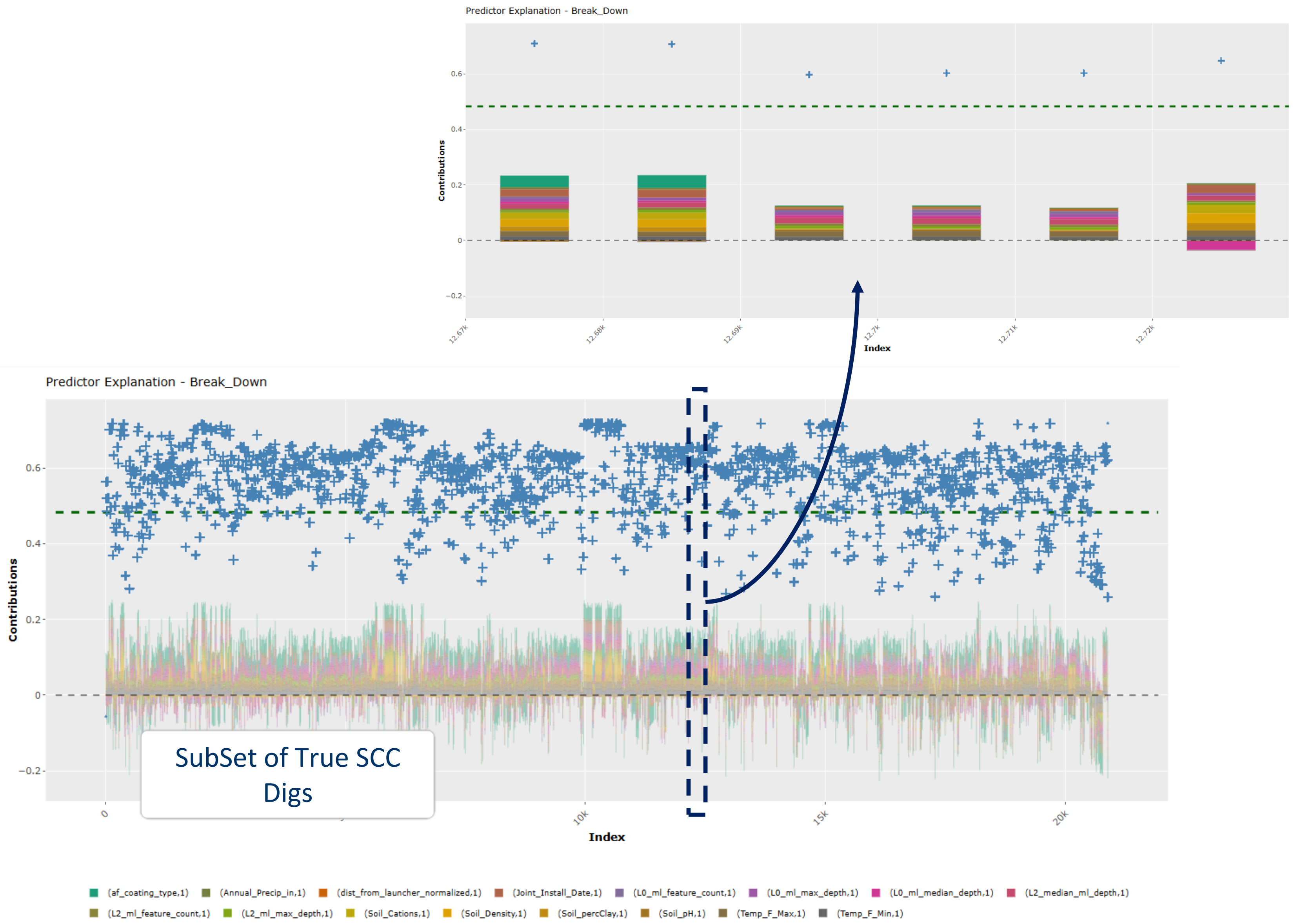




Learned Model – Explanatory

Prediction Explanations

- Predictions may be Deconstructed at a Local Level to Reveal the Contribution of Each Predictor
- Analysis Provides Transparency and Practical Humna Readable Explanations of Results



Learned Model – Validation Prior to Deployment

Task = Binary Classification
Target = joint_has_SCC
Truth Observations = joint_has_SCC
Positive Class = TRUE
Available Truth Observations = 814
Model Learning Records = 0
Analysis Records = 814

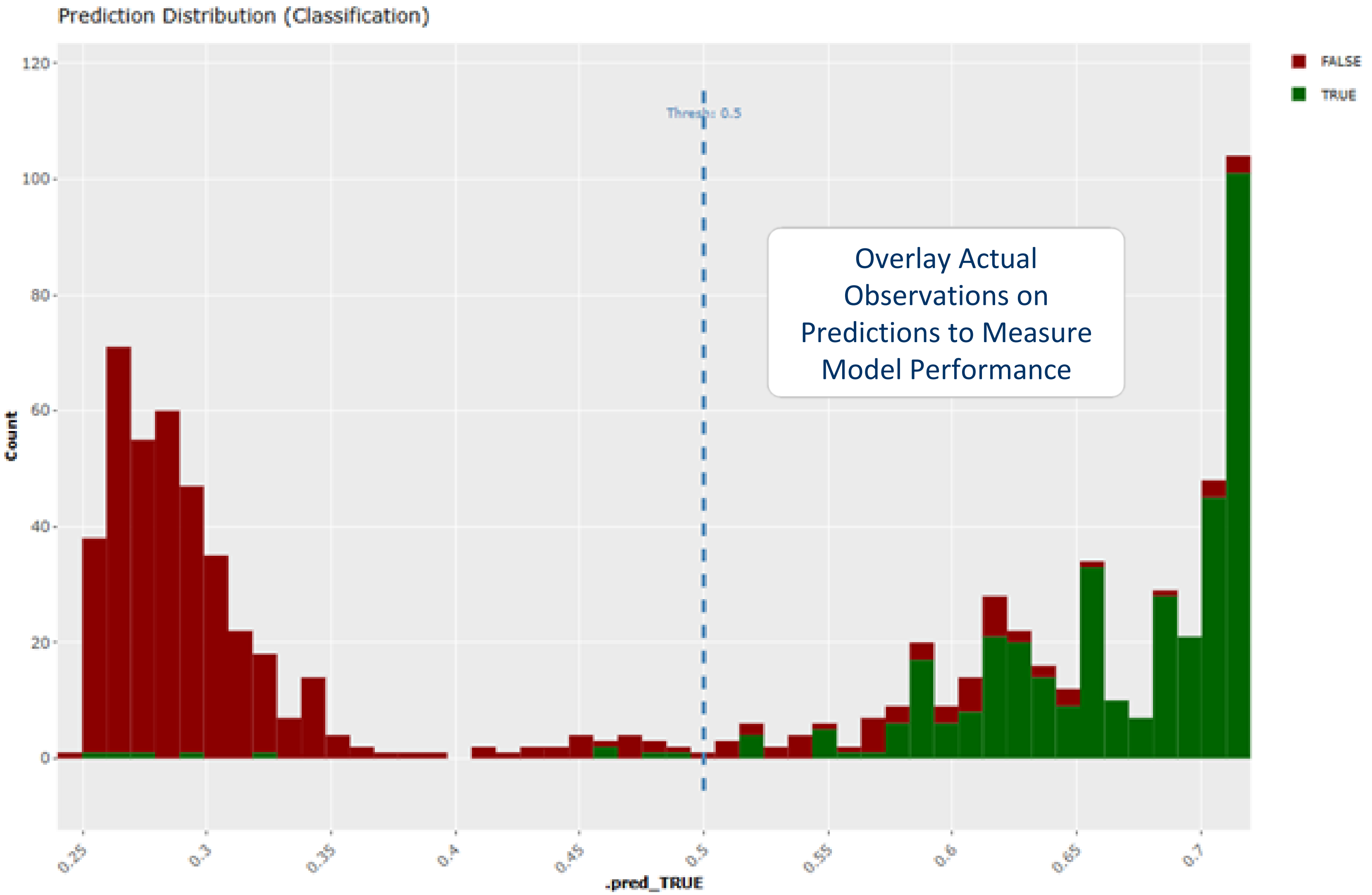
Threshold

0.5

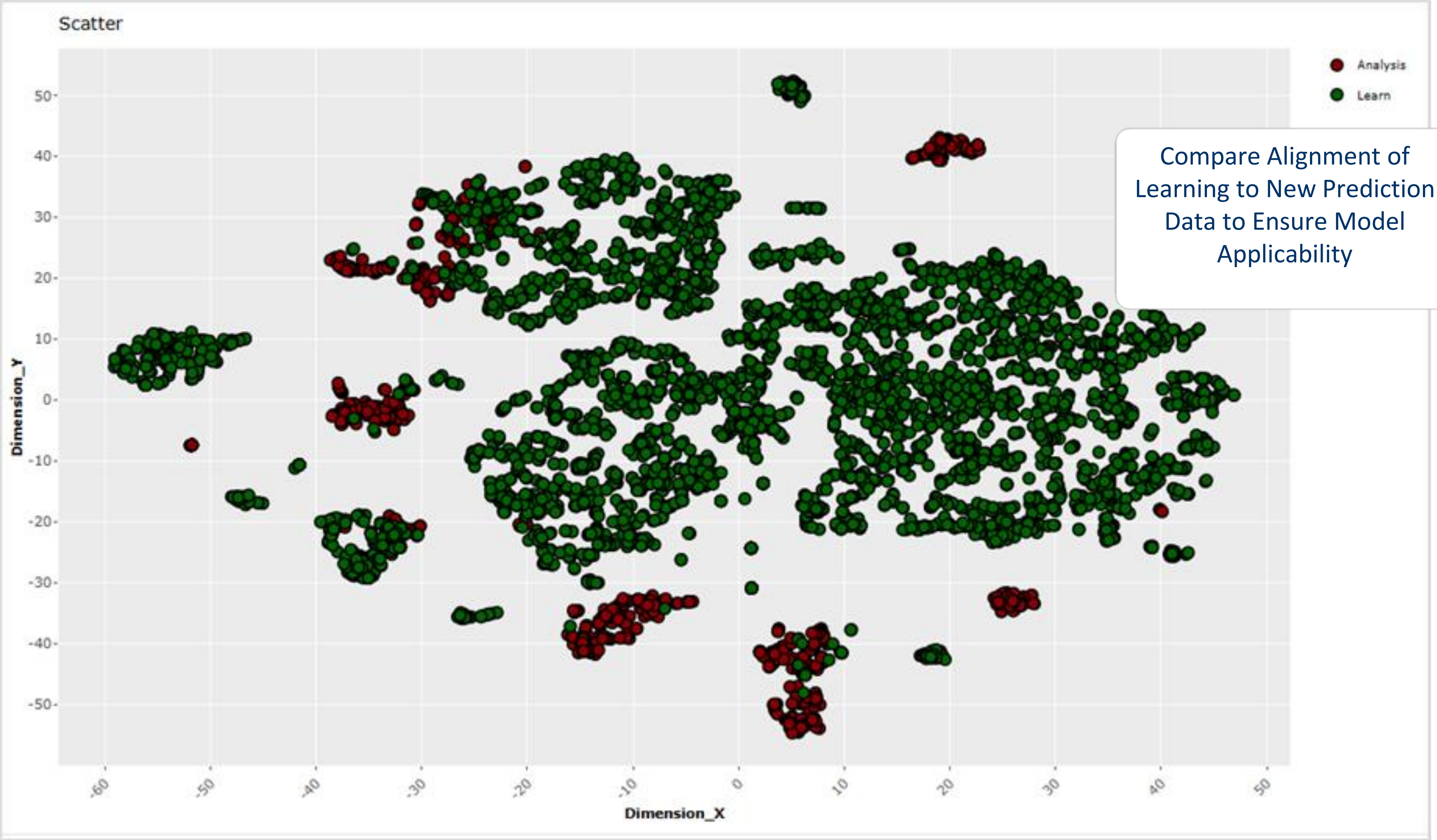
	AN	AP
PN	392	9
PP	56	357

Metric	Value
Accuracy	0.92
Sensitivity	0.97
Specificity	0.88
Balanced Accuracy	0.92

PN = Predicted Negative
PP = Predicted Positive
AN = Actual Negative
AP = Actual Positive



Learned Model – Similarity Check Prior to Deployment

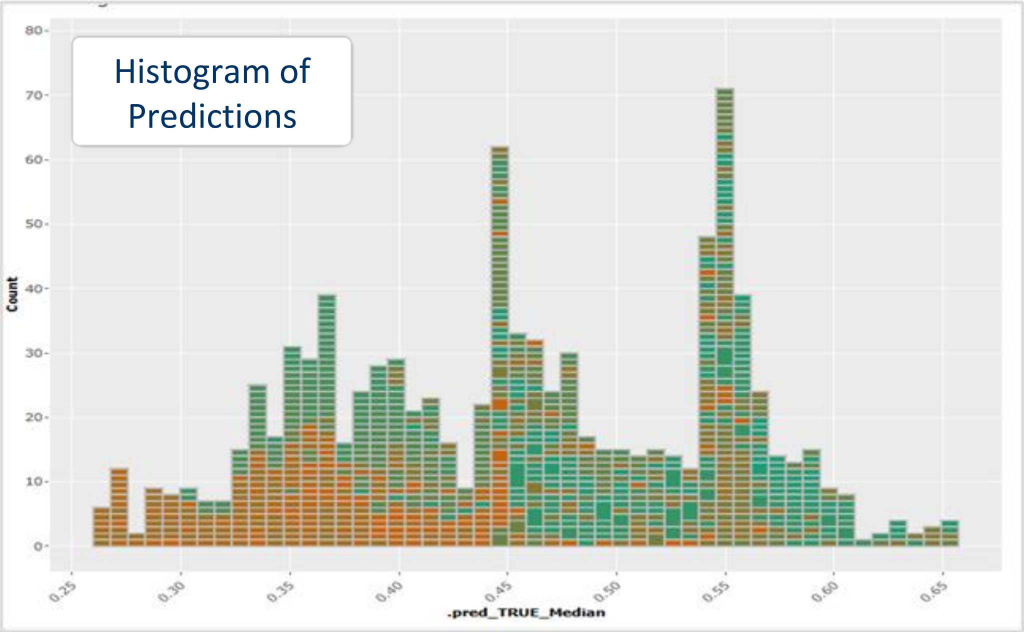


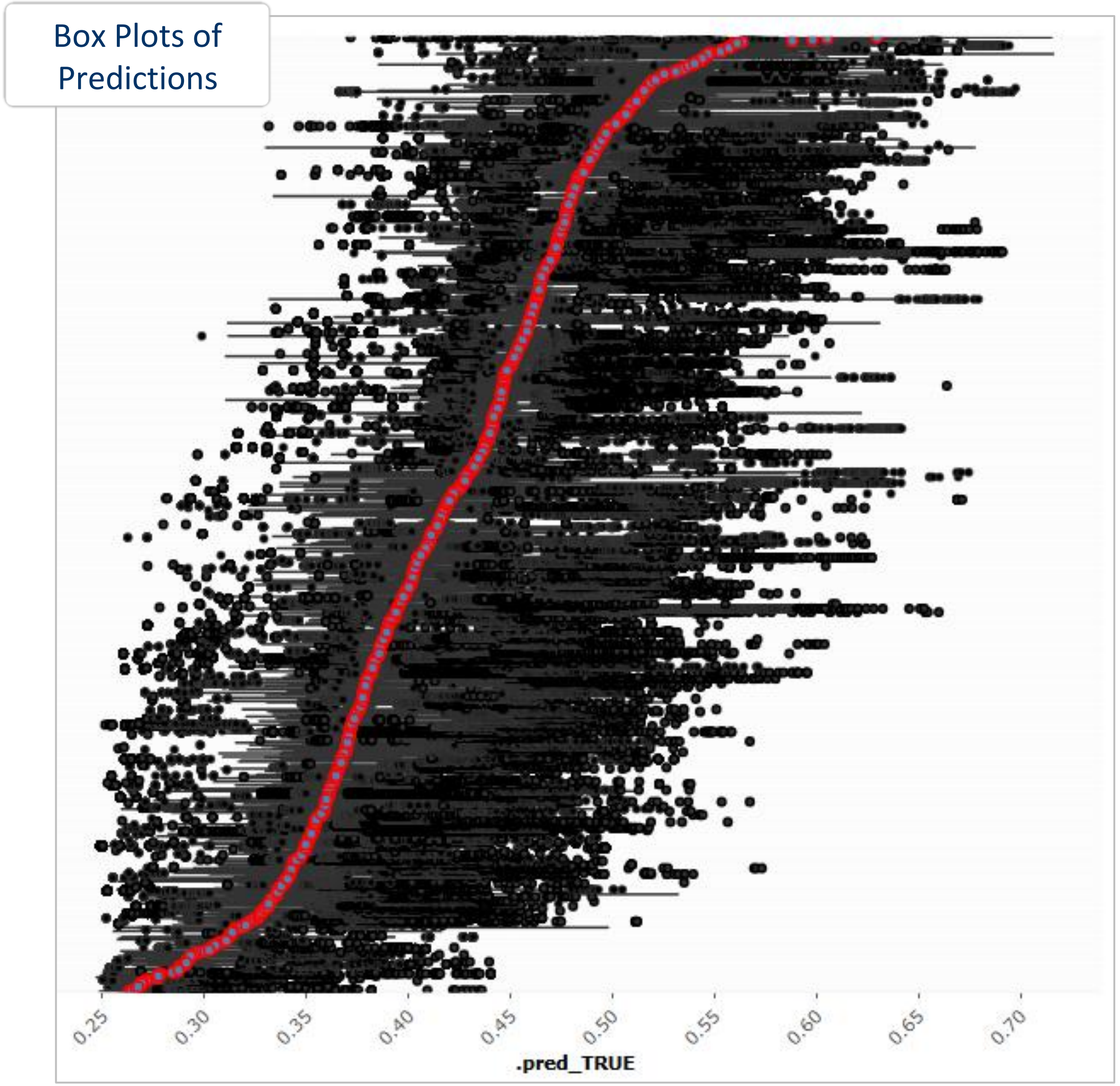
Compare Alignment of Learning to New Prediction Data to Ensure Model Applicability

Learned Model – Application

Apply ML Model

- Apply Model to Each Joint of Each Pipeline (Predict Probability of SCC)
- Box-Plots Show Prediction Variability within Pipelines
- Histograms Show Overall Prediction Profile of All Pipelines





Summary

- Augmented MFL & Environmental Data Proved Valuable in the Analysis of SCC Susceptibility
- Machine Learning & Advanced Analytics Provided Useful Outputs to the Practitioner and Domain Expert
- Results may be Used to Optimize the Planning of Inspections, Selection of DA Dig Sites, Sensitivity Analysis and Augmentation of Existing Determinsitic Models