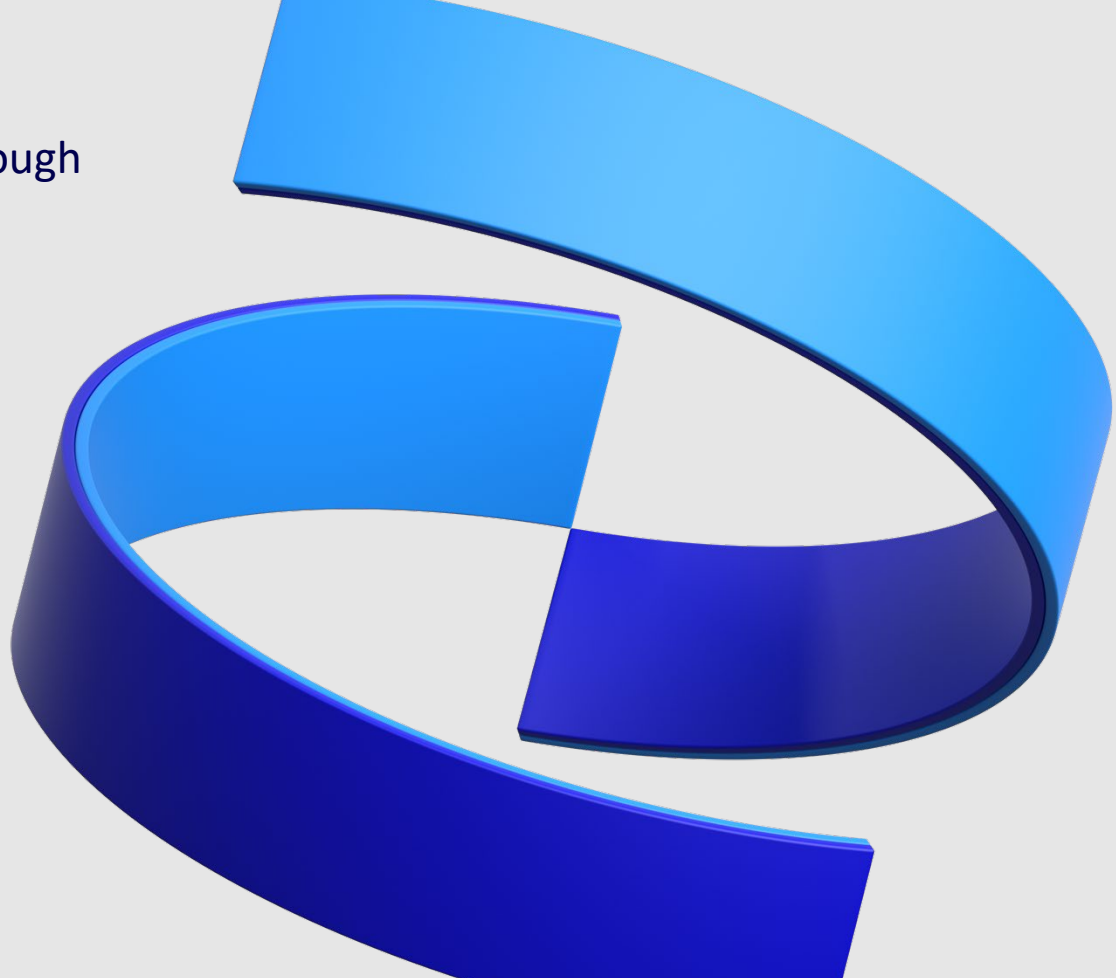
Enhancing Yield Investigations through Interpretable Machine Learning



Shu Yang
Manager, AI & Advanced Control

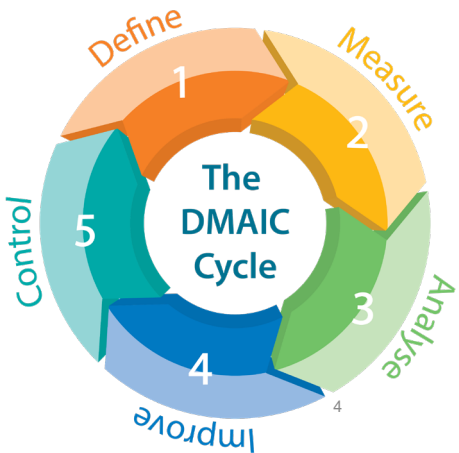
Contents

- Background: Goals and Challenges
- Methodology: Introduction to IML
- Case Study
- Discussions

Background

Six-Sigma

- Six Sigma: A Major Trend
  - Six Sigma and DMAIC methodologies are widely adopted in biopharma for process optimization and yield improvement.
  - The FDA encourages mature quality management systems and data-driven investigations to ensure high product quality and supply chain resilience.
  - Continuous improvement and root cause analysis are central to regulatory expectations and industry best practices.

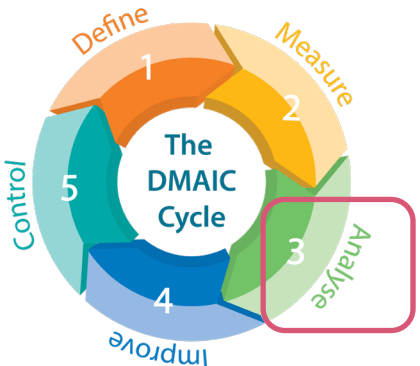


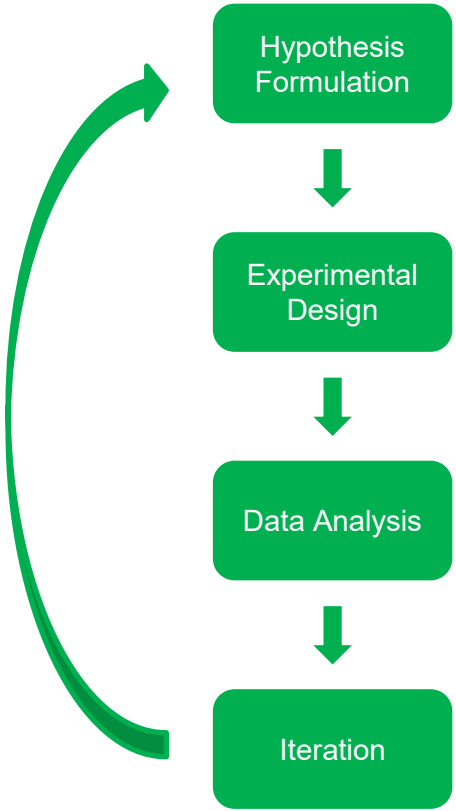
Analysis Challenge

*Analyze the data to investigate and **verify cause and effect**. Determine what the relationships are, and attempt to **ensure that all factors have been considered**. Seek out the root cause of the defect under investigation.*

Six Sigma - Wikipedia

- Massive hypothesis set: Hundreds of variables, time series
- Nonlinearity: Intrinsic nonlinear biological systems
- Multivariate : Variables interact in unpredictable ways





The Catch-22 in Hypothesis Formulation

Goal: Understand the data generating process (gain knowledge)

Knowledge: to build a structured model, we need knowledge of the data generating process

Model: To analyze the data, we need a statistical model

Structure: To model complex relationships, we need hypothesis and structured model

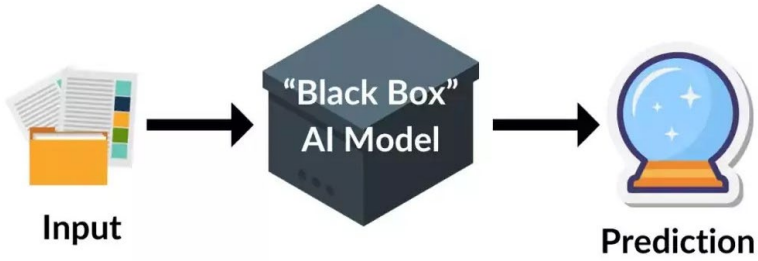
Methodology

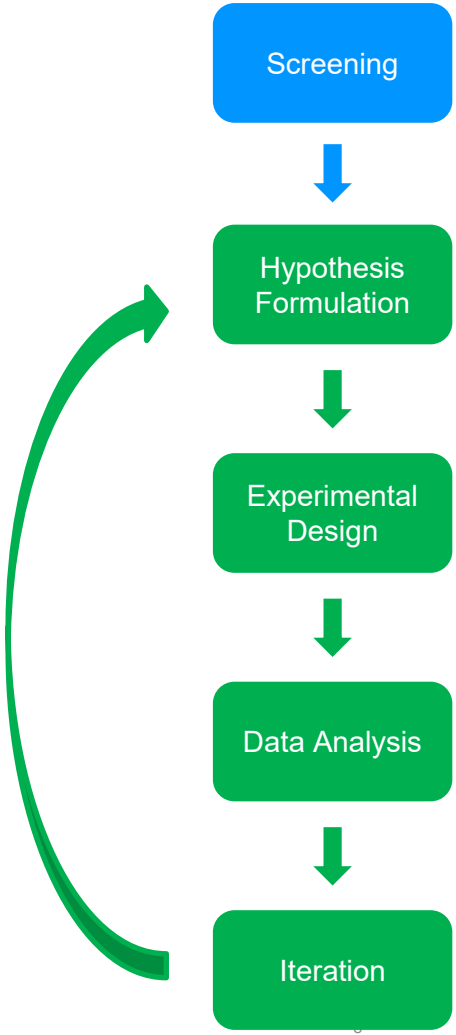
Breaking the Catch-22

Aid Hypothesis-Driven With Data-Driven

- Biostatistics excels at rigorous hypothesis testing and model building.
- Non-parametric methods can map complex correlation with less effort
- What if:
  - Throw the tedious variable screening work to non-parametric model.

  - Biostats and process SMEs focus on rigorous model development





Break the Catch-22: What's Available

Statistical Models

- High Interpretability
- Uncertainty Quantification
- Strong Theoretical Foundation
- Effective for Smaller Dataset

- Limited Flexibility
- More Effort

Interpretability + Pattern Recognition

Speed-Up Hypothesis Generation

- Non-parametric, data-driven methods: pattern recognition
  - No predefined model structure
- Generate ranked list of “potential factors”: method with interpretability
  - Scan for impactful variables from historical data
- Biostatisticians and process SMEs pick and refine
  - Avoid tedious screening process

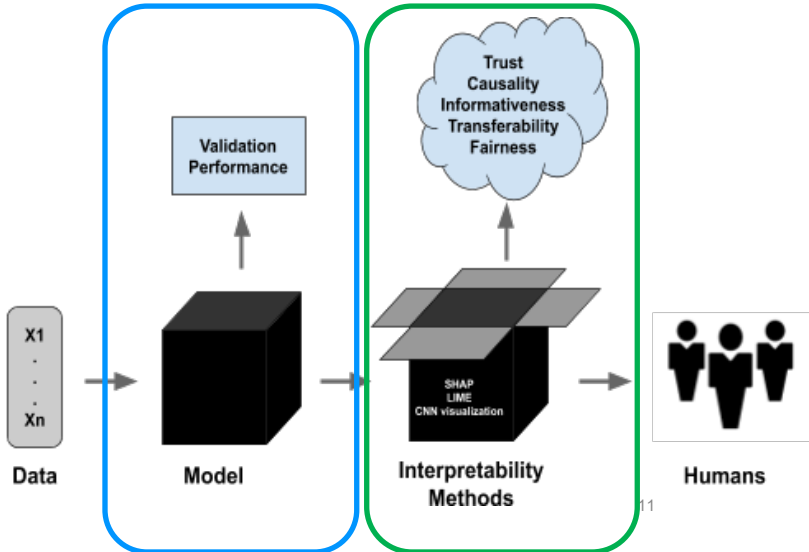
Interpretability + Pattern Recognition

Interpretable Machine Learning

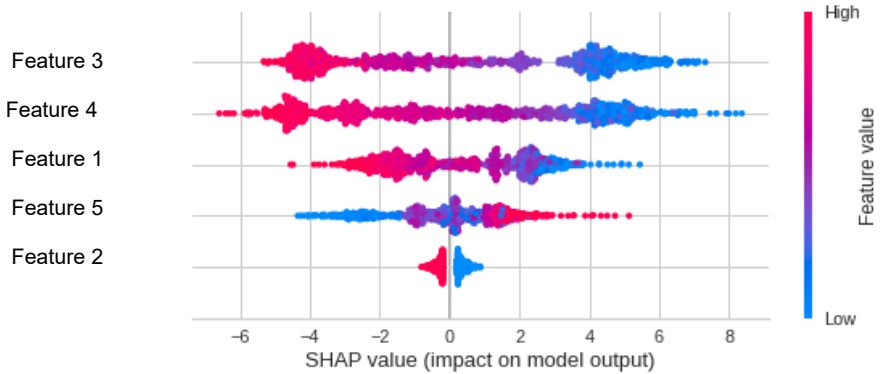
- Modular Approach

1. Mimic nonlinear patterns using arbitrary non-parametric model
   - Any off-the-shelf ML packages

2. Interrogate model for "insight"
   - Model-agnostic tools based on marginal contributions



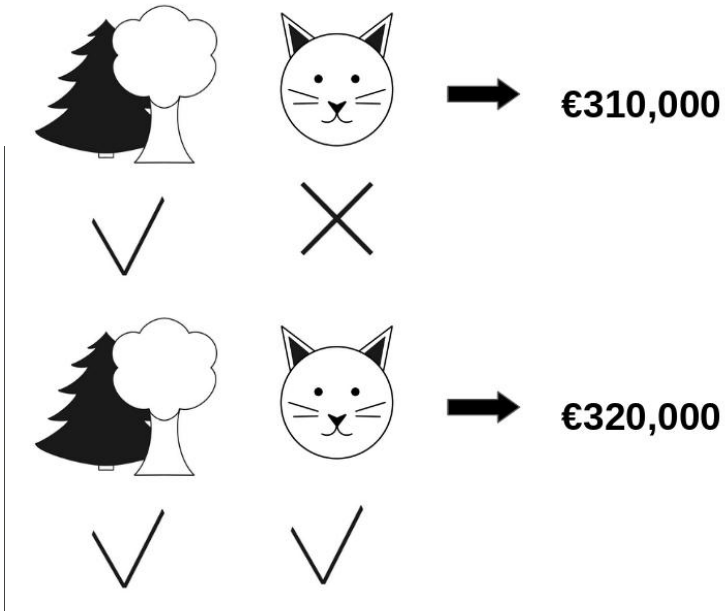
Shapley Value: Model Interrogation



Shapley Values – Interpretable Machine Learning

Marginal Contribution

- Marginal contributions in one scenario



Shapley Values – Interpretable Machine Learning

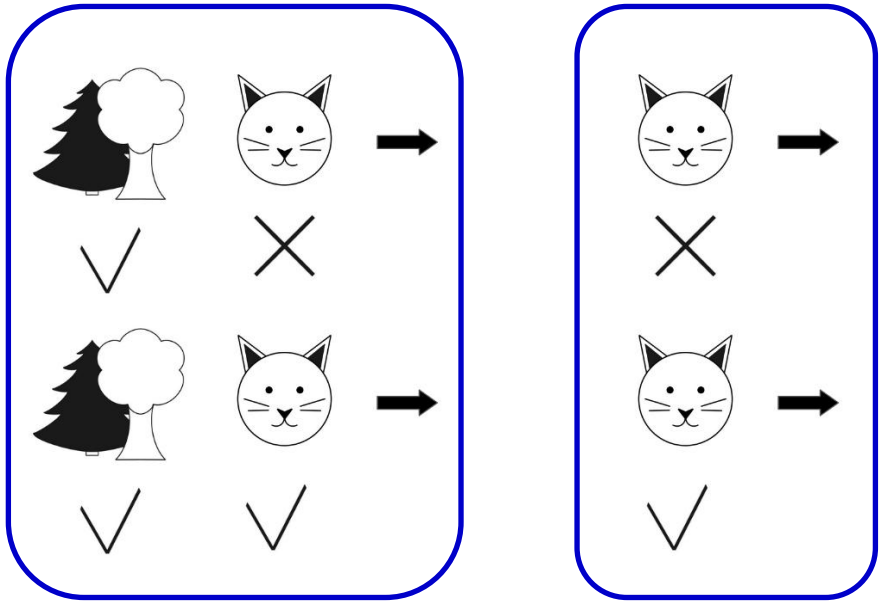
Average Marginal Contribution

- Average marginal contributions
  - For each combination of all other features
    - Calculate “cat allowed” – “cat not allowed”

$$\phi_j(val) = \sum_{S \subseteq \{1,\dots,p\} \setminus \{j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} (val(S \cup \{j\}) - val(S))$$

where

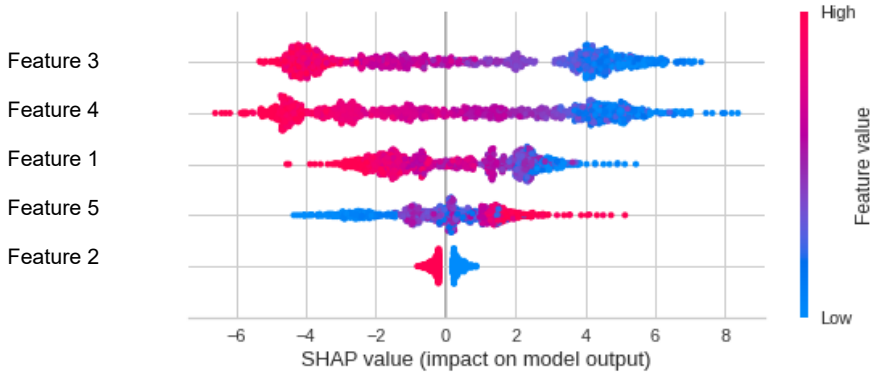
$$val_x(S) = \int \hat{f}(x_1, \dots, x_p) dP_{X_C} - E[\hat{f}(X)]$$



Calculate Shapley Value For All Samples

1. X-axis: Sample-specific impact
2. Y-axis: Variables in order of importance
3. Color: Actual feature value

d

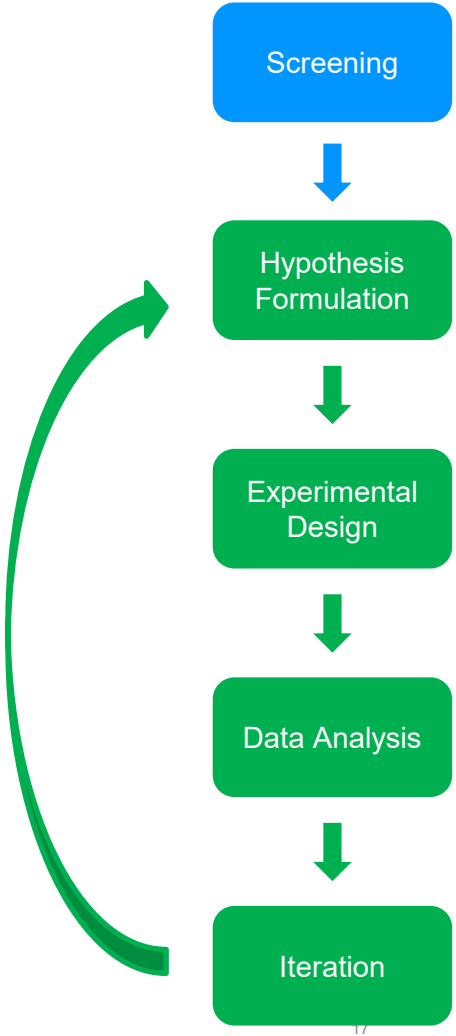


Shapley Values – Interpretable Machine Learning

Case Study

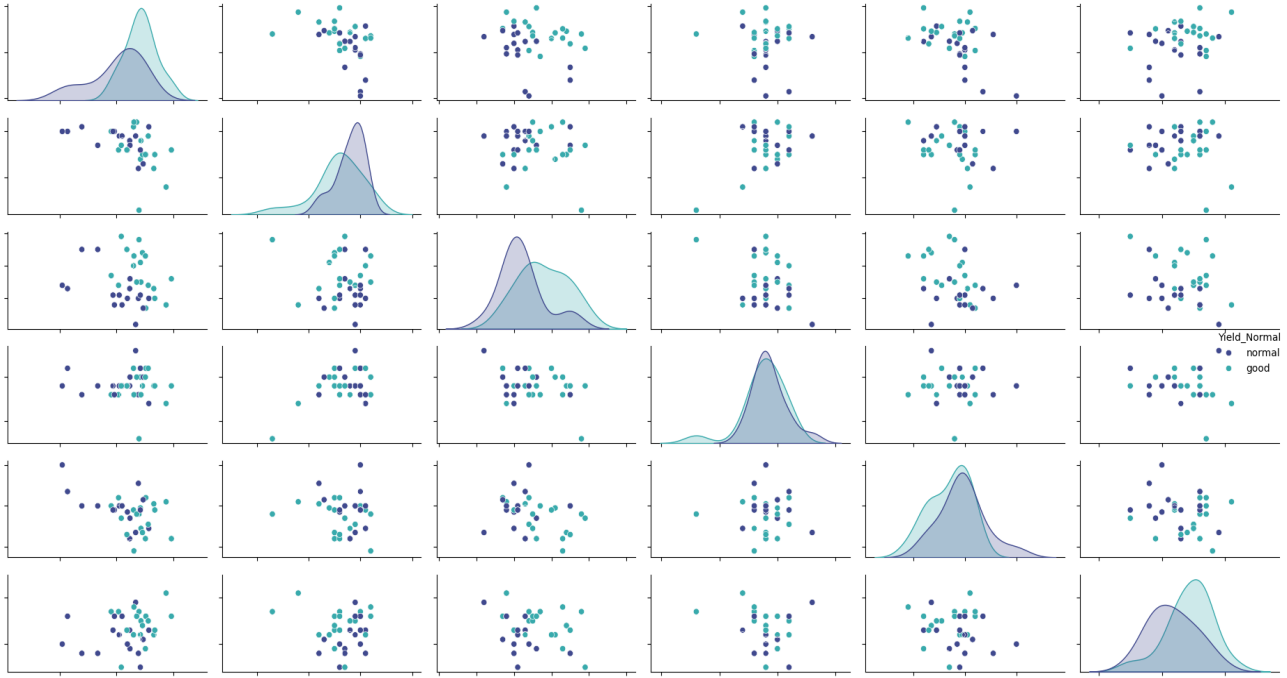
Steps Taken

1. Select unit operation
    - Univariate analysis
2. Select subset of variables in unit operation
    - Colinear analysis
    - Manual feedback
3. Quantify impact of variables on target
    - IML
4. Deep dive by Biostats+SMEs



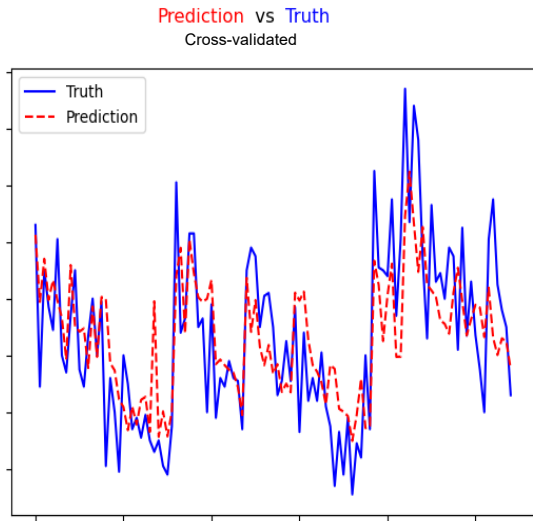
Select Unit Operations

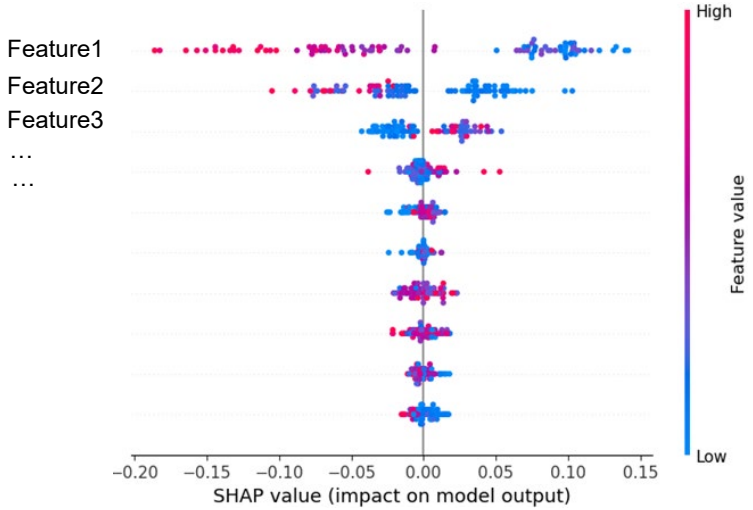
- Different between good vs bad
- Correlated between steps?



Feature Selection and IML

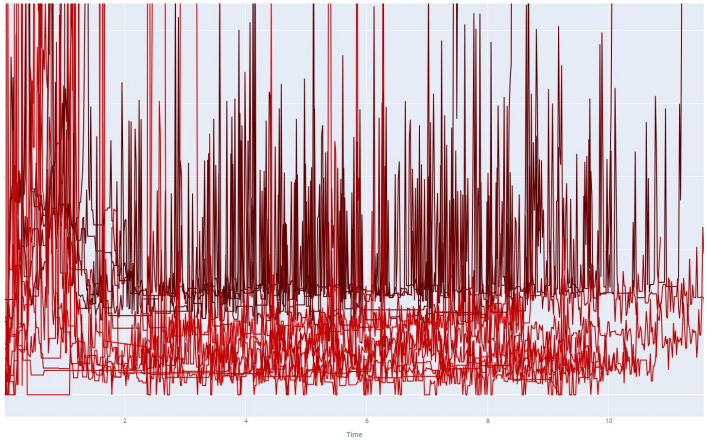
- Colinear removal
  - Univariate: Correlation analysis
  - Multivariate: VIF
- Conduct IML
  - Proper development of ML models
    - K-fold cross validation
  - Calculate Shapley score for each historical batch

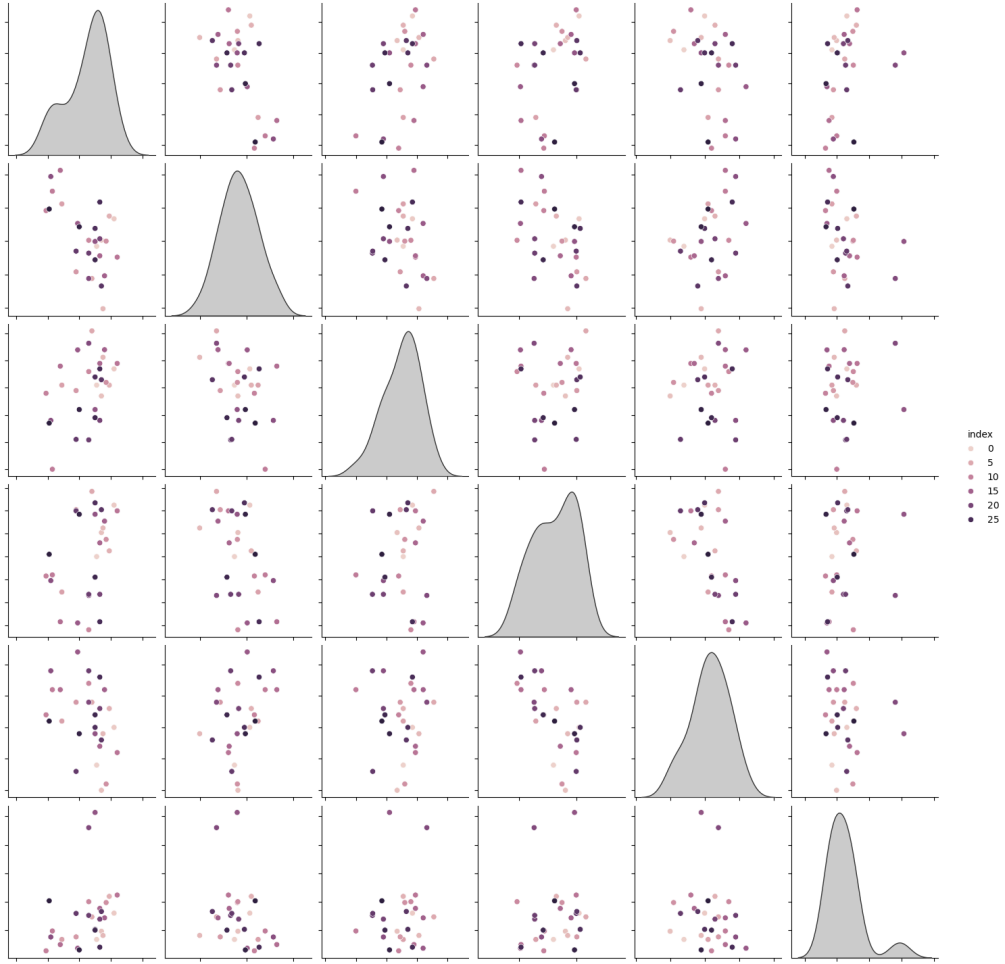




Post IML Check

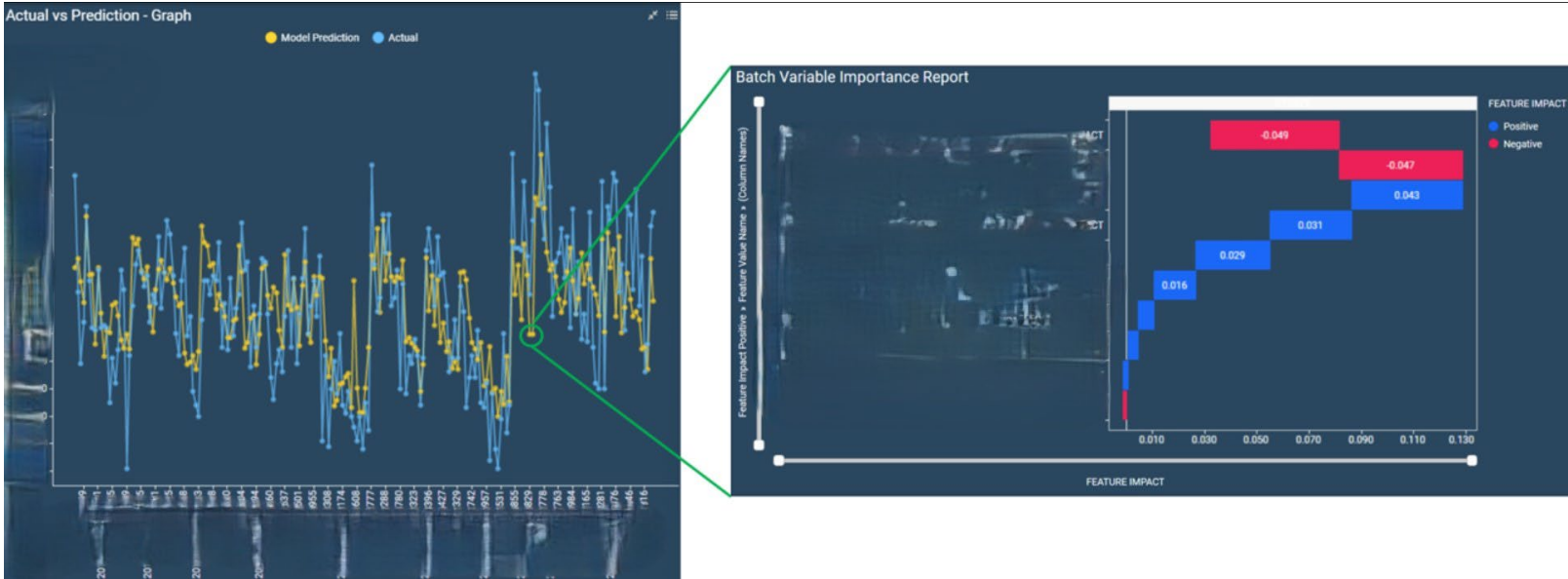
- Top correlated variable of picked ones
  - Possible confounder?
- Look into related batch evolution data
  - Colored by y





Results

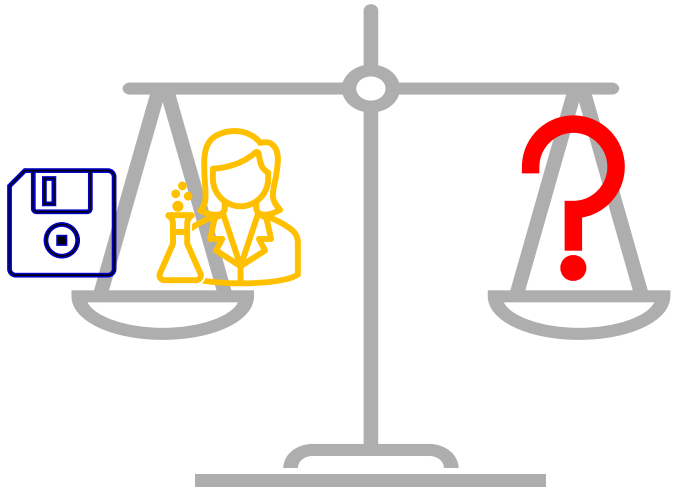
- “Surprising” factors found: 6
- Accelerated investigation: <2 month in analysis
- Yield improvement :**$3.5M**
- Dashboard deployed supporting continuous improvement



Discussion

Insufficient, But Useful

- Unmeasured Variables
- Large Hypothesis Set
  - Which Unit operation
  - Which tag
  - Which timepoint
- Nonlinear
- Multivariate
- Causality?



Acknowledgment

Tech Service :
- Samantha Paniagua
- Kathleen Long
- Solomon Gobaw
- Steven Pereira
- Ethan Wright
- Sri Marri
- Paul Haupt-Renaud
- Todd Miller
- Wilson French
- Lee Mcoy
- Scott Orlando

Operations:
- Scott Wolfrom
- Dennis Woodby
- Christina Cokely
- Sadie MacLean
- Kayla Sica
- Laura Moore

MSAT:
- Jonathan Kinross
- Allan Maxwell
- Krishana Gulla
- Aparajita Dasgupta

Digital:
- Ankur Nagpal
- Rashid Mijumbi

Manufacturing Intelligence
- William Fahey
- Simon Glaude
- Debadri Dutta
- Reza Kamyar

- Chi-Shi Chen
- Jeffery Doyle
- Meghan Griffin

Backup

Recap

- Univariate:
  - t-Test/ANOVA/Chi-Square
    - Which variable is “different” between “good vs bad”
  - Correlation Analysis
    - Pearson: linear correlation
    - Spearman: monotonicity

Recap

- Multivariate Linear
  - OLS/Logistic regression: p-value
  - Variable selection
    - LASSO (Elastic Net)
    - Anything with a Laplacian prior/L1 regularization
- Nonlinear:
  - Stepwise regression
    - Unstable with correlated variables