Making the Most of Data: Feature Engineering for Applied Supervised Machine Learning

Advanced Infrastructure Integrity Modeling

Presented to DOE Data Days
June 2, 2022 - LLNL
Disclaimer

This project was funded by the United States Department of Energy, National Energy Technology Laboratory, in part, through a site support contract. Neither the United States Government nor any agency thereof, nor any of their employees, nor the support contractor, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.
Authors and Key Personnel

PIs
Lucy Romeo\textsuperscript{1,2} and Jennifer Bauer\textsuperscript{1}

Key Personnel
Rodrigo Duran\textsuperscript{3}, Alec Dyer\textsuperscript{1,2}, Isabelle Pfander\textsuperscript{1,2}, Thomas Martin\textsuperscript{1,2}, Chukwuemeka Okoli\textsuperscript{1,2}, Kelly Rose\textsuperscript{1}, Michael Sabbatino\textsuperscript{1,2}, Madison Wenzlick\textsuperscript{1,2}, Patrick Wingo\textsuperscript{1,2}, Dakota Zaengle\textsuperscript{1,2}

\textsuperscript{1) National Energy Technology Laboratory, 1450 Queen Avenue SW, Albany, OR 97321, USA  
\textsuperscript{2) NETL Support Contractor, 1450 Queen Avenue SW, Albany, OR 97321, USA  
\textsuperscript{3) Theiss Research, 7411 Eads Avenue, La Jolla, CA 92037, USA

Thomas.Martine@netl.doe.gov  
Lucy.Romeo@netl.doe.gov  
Jennifer.Bauer@netl.doe.gov
Setting the Stage

Machine Learning (ML)

- **Supervised ML** – machine is trained, taught with labeled examples
- **Unsupervised ML** – machine creates its own labels (i.e. clustering)

**Big Data & Big Data Computing** –
Large volumes, variety, variability, velocity of data and the computing engineering & systems to handle them

**Features** – Variables or attributes (ex. continuous or categorical)

**Feature Engineering** – Select, transform, process, and visualize input features of a given dataset

Offshore Infrastructure Hazards

- Aging infrastructure
- Operational wear-and-tear
- Offshore environment:
  - Extreme weather
  - Climate change
  - Corrosion hazards
  - Geohazards
- Need:
  - Identify & prevent hazards
  - Inform safe lifespan extension strategies
  - Environmentally prudent planning in low-carbon economy

>60% of platforms >30 years old

How AIIM Operates

Utilizes **big data, big data computing, and multiple ML models** to forecast infrastructure lifespan and risk.

Key points:
- Data **analysis** and **visualization** at every step
- Subject Matter Expert **QAQC**
- A focus on the final product being explainable, logical, and defendable

**Critical Insights!**

**Do the results make sense?**

**Yes**

**Visualize Results**

**No**

**Explore Data**

**Apply ML**

**Feature Engineering**

**Built a dataset of >11k platforms with >2k features** representing natural-engineered offshore system

- Structures
- Metocean
- Incidents
- Biochemical
- Geohazards
- Production

**Singular Value Decomposition (SVD)**

**Self-Organizing Maps (SOM)**

**Other Methods!**

**Nelson et al., 2021**

**Dyer et al., 2022**

Getting to Know the **Integrated** Data

**Challenges & Opportunities**

> 11,000 platform records * > 2,000 features  
  = >22,000,000 data values

~50% of the dataset has ~90% coverage on a per feature basis

**Feature Breakdown**

- Structure: 40%
- Incident: 14%
- Metocean Stats: 11%
- Annual Production: 4%
- Well and Entity Stats: 3%
- Geohazards: 5%

Integration has increased data complexity

**Feature Breakdown**

- **Metocean Variables**
- **Production Records**
- **Reported Incidents**

- **Structural Information**
- **Geohazards**

Singular Value Decomposition (SVD)

SVD efficiently identifies and summarizes important information in a correlation or covariance matrix.

Pros

- Interpretable
- Appropriate for time series and continuous spatial data
- Most efficient way to summarize data in a matrix (Eckart-Young Theorem)

Cons

- Does not work with categorical features
- Incomplete data requires pre-processing
- Expert opinion needed to select features

First three right singular vectors of a data correlation matrix, showing relations between input variables and the target variable “Age of Removal”. 76% percent of features explained by 6 features.

Self Organizing Maps (SOM)

SOM is an *unsupervised* ML technique that is a specific type of neural network. SOMs identify non-linear feature relationships.

**Pros**
- Can be used with nonlinear features
- Relatively fast
- Threshold for different features is user-selected
- Can be used to create composite features

**Cons**
- Can’t be used with categorical data
- Expert opinion needed to select features
- Like all neural networks, complete and pre-processed data helps with convergence and speed

Letting the Model Decide – Feature Importance

Gradient Boosted Decision Trees (GBDTs) are a common and well-used ML algorithm. This is one method to assess every feature’s importance.

Pros
- Handles all data types
- Easily interpretable
- Great ML model to be used for prediction as well

Cons
- Shared feature importance (potentially collinearity w/other input variables).
- Scores are presented quantitively, easy to overinterpret.
- Model accuracy has limited impact on feature importance.

Feature Importance from a GBDT (using CatBoost). This specific model removed many features while still retaining similar accuracy. Low importance features could be further removed.

### Comparison of Methods

#### Overcoming incomplete, complex, multivariate data

<table>
<thead>
<tr>
<th>Method</th>
<th>Best for</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVD</td>
<td>numeric data (time series, spatial)</td>
<td>Identified variables containing duplicate information. Highlighted storm-related features as important.</td>
</tr>
<tr>
<td>SOM</td>
<td>deciding between closely related non-linear features</td>
<td>Confirmed age variables are key. Confirmed findings from SVD testing.</td>
</tr>
<tr>
<td>GBDT</td>
<td>categorical, incomplete data</td>
<td>Using top 5–10 variables does not degrade model performance. Continued interpretation of environmental loadings and age variables is key.</td>
</tr>
</tbody>
</table>

Using any method alone will give an incomplete picture.

Key Findings

Feature Engineering Matters!
- Reduced input features from >2,000 to >20
- *Minimizing* complexity & error, *maintaining* accuracy
- Insights to inform safe infrastructure reuse & removal
- Identify hazards to support environmental and operational risk prevention

Next Steps:
- Finalize feature engineering
- Expand to evaluate pipelines & wellbores
- Develop & compare additional models
- Build an interactive AIIM modeling and visualization tool


Thomas Martin
thomas.martin@netl.doe.gov
References


