FOPAM process data analytics workshop

Instructor: Leo Chiang
Guest speaker: Ivan Castillo
In collaboration with Richard Braatz and Joe Qin

August 6, 2019

Contributions from team Dow including Bea Braun, Swee-Teng Chin, Lloyd Colegrove, Mark Joswiak, Yoyo Peng, Ricardo Rendall, Alix Schmidt, Mary Beth Seasholtz, Monica Trevino, James Wade, Zhenyu Wang, and Mark Webb

Dow.com

3 - Industrial Experience and Tips, Interactive Discussions

3.1 Visualization

3.2 Outlier detection and data preprocessing

3.3 Method selection

3.4 How good is good enough? Industrial tips and tricks of the trade

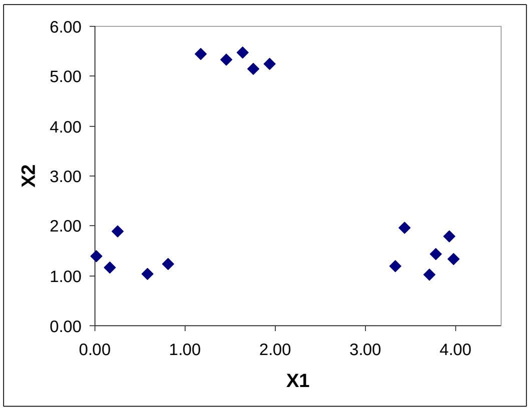
3.5 Industrial case studies

Data Driven Decision Making

Visualization: Data in context

X1	X2
1.76	5.14
1.46	5.34
1.63	5.47
0.81	1.23
1.17	5.45
0.25	1.89
1.93	5.25
0.02	1.40
3.33	1.19
0.17	1.16
…	…

vs.

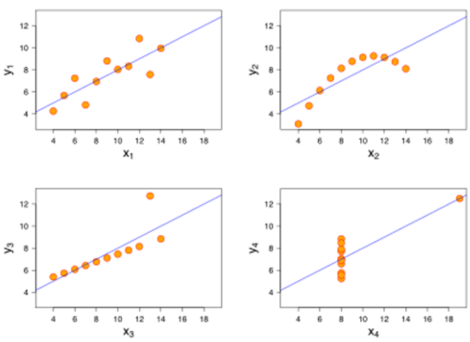


Dow | ANALYTICS 4.0
DATA DRIVEN DECISION MAKING

Visualization: Data in context

Correlation coefficient → caution

Example: Anscombe's quartet



Property	Value	Accuracy
Mean of x	9	Exact
Variance of x	11	Exact
Mean of y	7.50	To 2 decimal places
Variance of y	4.125	+/- 0.003
Correlation between x and y	0.816	To 3 decimal places
Linear regression	y = 3.00 + 0.500*x	To 2, 3 decimal places
Coefficient of determination of linear regression	0.67	To 2 decimal places

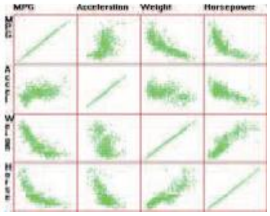
Dow | ANALYTICS 4.0
DATA DRIVEN DECISION MAKING

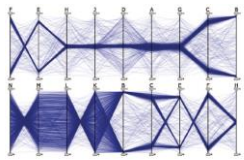
How to visualize a dataset: Direct method

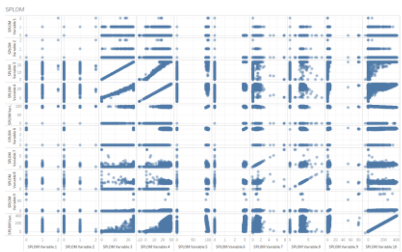
Direct

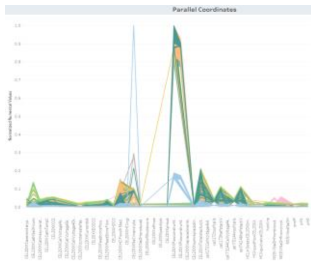
Correlation

Parallel coordinates





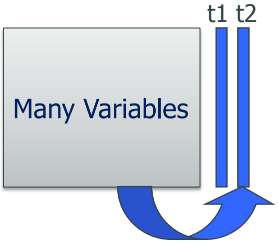




How to visualize a dataset: Indirect method

Indirect

Dimensionality reduction techniques (such as PCA) to preserve characteristics of original dataset

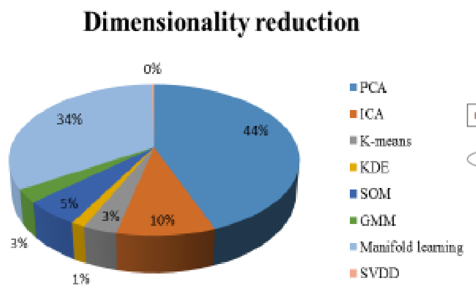


Principal component analysis is the industry workhorse

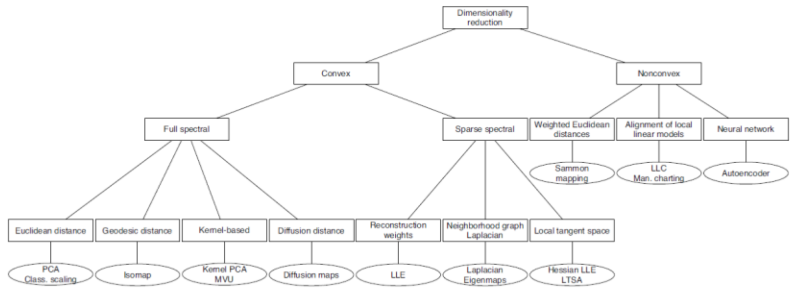
- Linear combinations of original variables
- Identifies correlations (structured variation)
- Leaves noise behind (unstructured variation)
- No parameters to tune
- Fast

Analogy
Interpreting the shadows of a complicated 3D-geometry after projection onto a 2D surface…

Besides PCA, what are the other options?



Ge et al. *IEEE Access* 2017



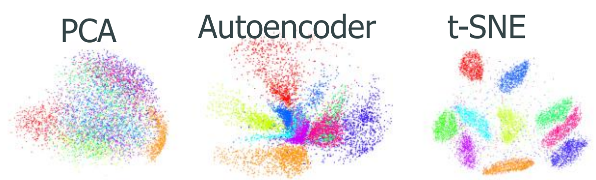
A variety of techniques exist
- Global/local
- Linear/non-linear
- Parametric/non-parametric
- Manifold learners
- Tunable parameters

Dow | Analytics 4.0
Data Driven Decision Making

7

t-SNE is the current gold standard

MNIST dataset (handwritten digits 0-9, colored by number) van der Maaten and Hinton *J. MLR* 2009



Identifies natural clusters
Solves crowding problem

How it works:
- Point-point similarity is a probability, which is sought to be preserved in reduction.
- Create latent graph most similar to original graph. Minimize reconstruction error of weights in graph edges

t-SNE Local focus

get clusters right

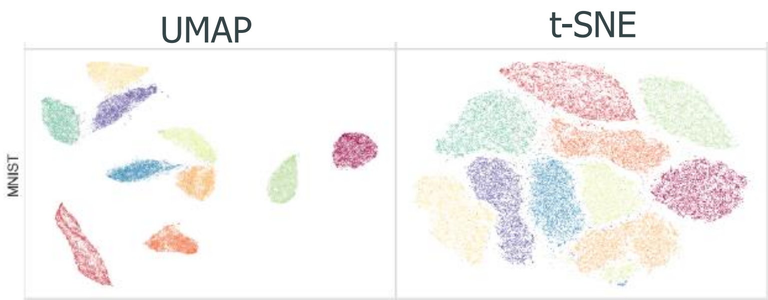
$$\min \sum_{e \in G} w_h \log \frac{w_h}{w_l}$$

Dow | Analytics 4.0
Data Driven Decision Making

8

t-SNE is the current gold standard…but here comes UMAP

Uniform Manifold Approximation and Projection



McInnes & Healy. ArXiv 1802.0342v2, 2018.

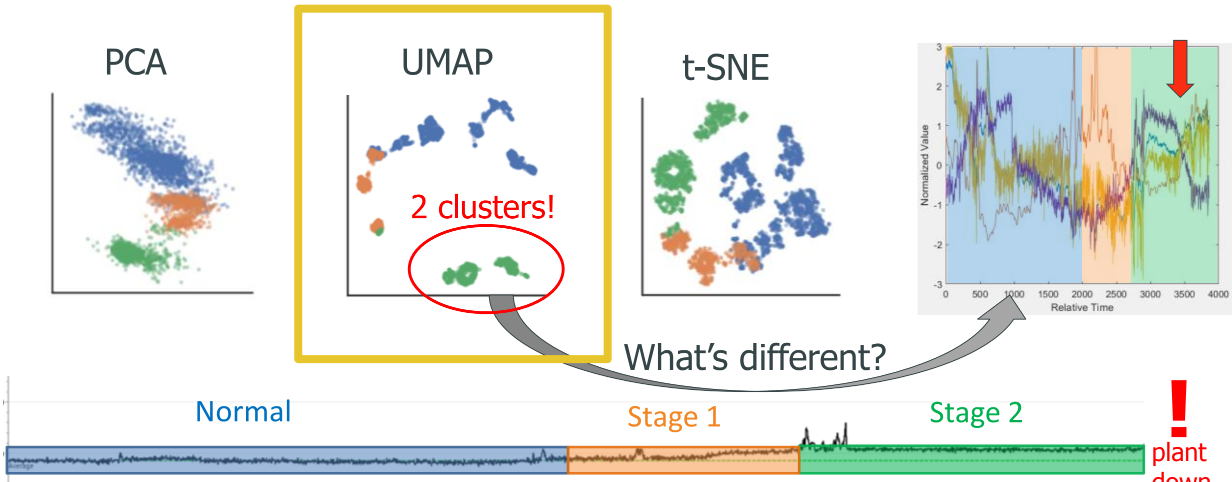
$$\min \sum_{e\in G} \overbrace{w_h \log\frac{w_h}{w_l}}^{\text{get clusters right}} + \overbrace{(1-w_h)\log\frac{1-w_h}{1-w_l}}^{\text{get gaps right}}$$

t-SNE Local focus

UMAP Balance local and global focus

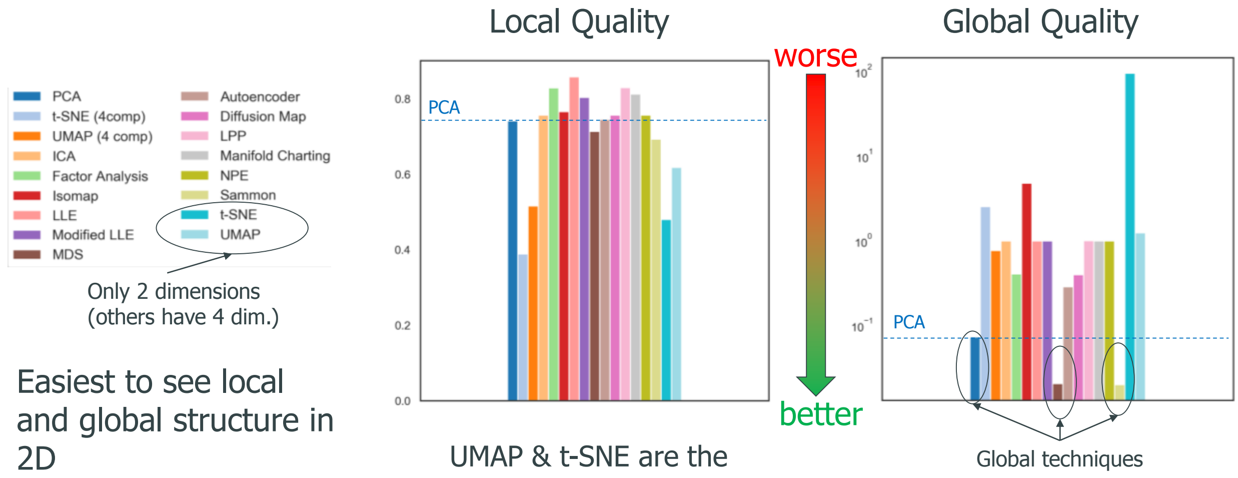
Case study 1: What occurs prior to an unplanned event?

UMAP and t-SNE immediately provide new insights



Labels from plant engineers Joswiak et al., *Control Engr Practice* 2019

Dimension reduction quality is a local/global tradeoff



Easiest to see local and global structure in 2D

UMAP & t-SNE are the best for local quality (even for only 2 dim.)

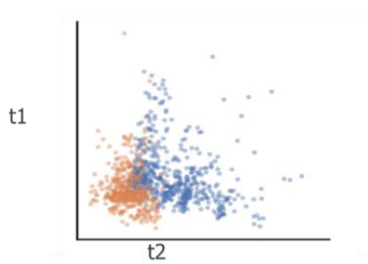
UMAP outperforms t-SNE, especially at 2 dimensions

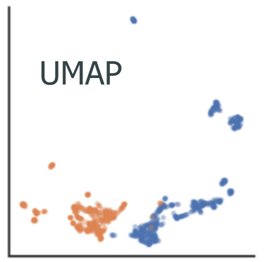
11

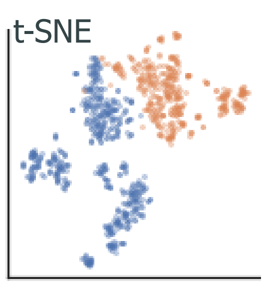
Data Driven Decision Making

Case Study 2: Different performance of two identical plants, A and B

PCA (and many other techniques) shows overlap of A and B







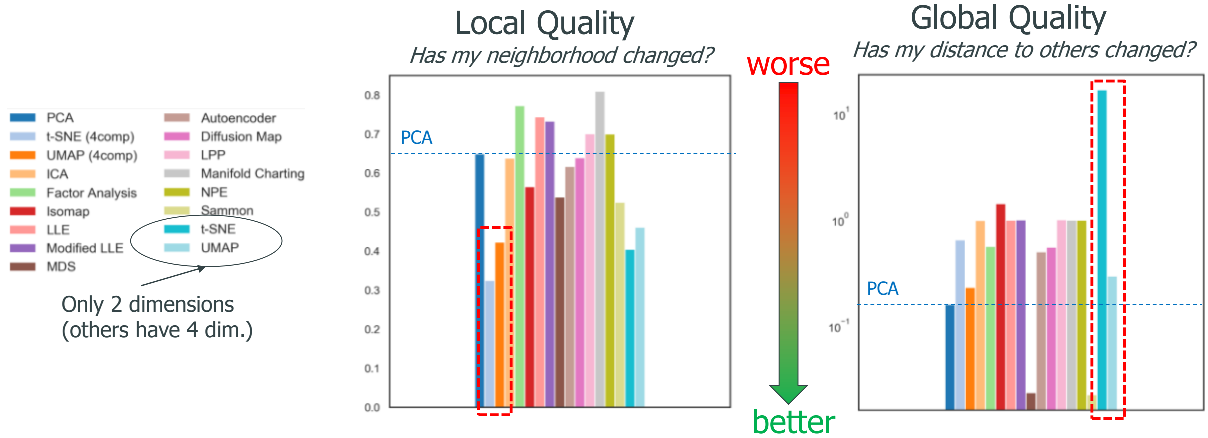
Advantages of UMAP

- Shows outlier cluster
- Quick visual analysis in just 2D
- Near t-SNE local quality with ~better than average global quality
- Clear (main) cluster separation even if data was not labeled
- Comparing UMAP clusters yields 7 more variables of interest over PCA-based analysis

12

Data Driven Decision Making

Dimension Reduction Quality Is a Local/Global Tradeoff



UMAP & t-SNE are the best for local quality (even for only 2 dim.)

UMAP outperforms t-SNE, especially at 2 dimensions

References

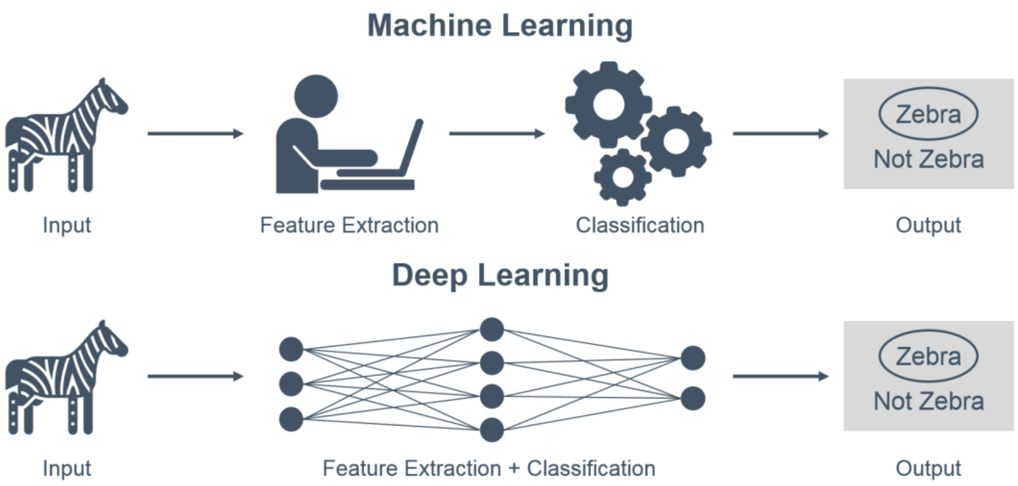
- Z. Ge, Z Song, SX. Ding, and B. Huang, Data mining and analytics in the process industry: The role of machine learning, *IEEE Access*, 5: 20590-20616, 2017.
- J. Zhang, H. Huang, J. Wang, Manifold Learning for Visualizing and Analyzing High-Dimensional, *IEEE Intelligent Systems*, 25(4): 54-61, 2010.
- LJP. van der Maaten, and GE Hinton, Visualizing data using t-SNE*, J. Mach Learn Res,* 9:2579–2605, 2008.
- L. McInnes, J. Healy, J. Melville, UMAP: Uniform manifold approximation and projection for dimensionality reduction, *arXiv*, 1802.03426v2, 2018.
- M. Joswiak, Y. Peng, I. Castillo, and L. Chiang, Visualizing Chemical Processes Utilizing Dimensionality Reduction Methods: Survey and Applications, *Control Engineering Practice*, 2019 (submitted).

3 - Industrial Experience and Tips, Interactive Discussions

Dow | ANALYTICS 4.0
DATA DRIVEN DECISION MAKING

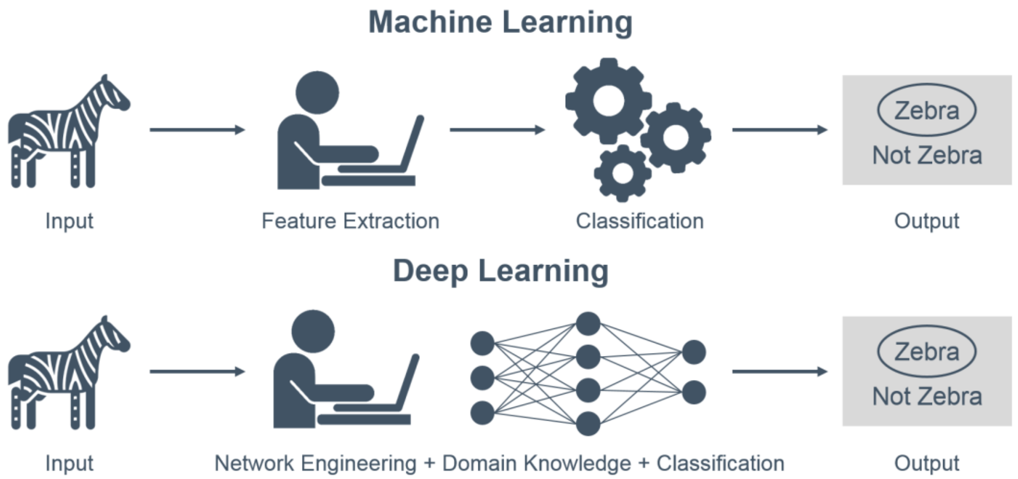
15

Some Opinions on Deep Learning vs Machine Learning

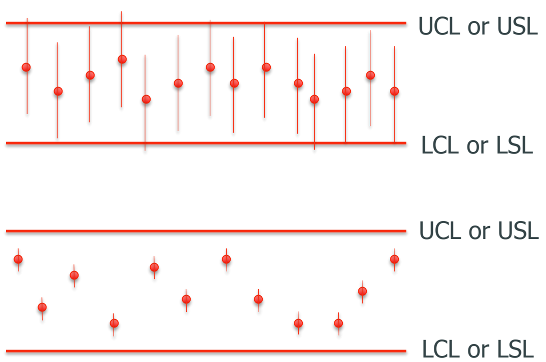


Dow | ANALYTICS 4.0
DATA DRIVEN DECISION MAKING

Some Opinions on Deep Learning vs Machine Learning



Data: Is it any good? How do you know?

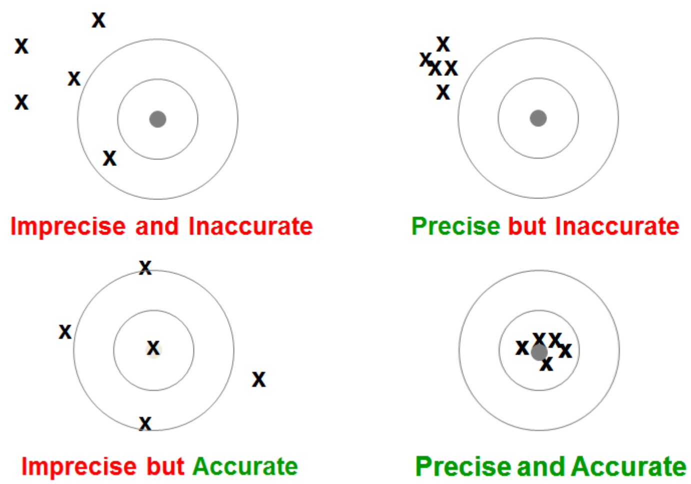


Data must be analyzed in context.

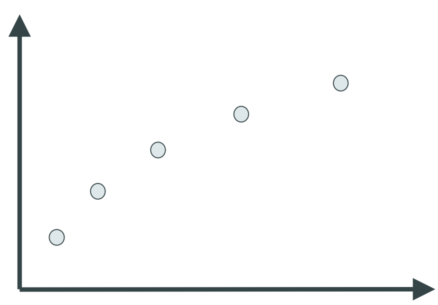
Do You Know Your Data?

What is accuracy?

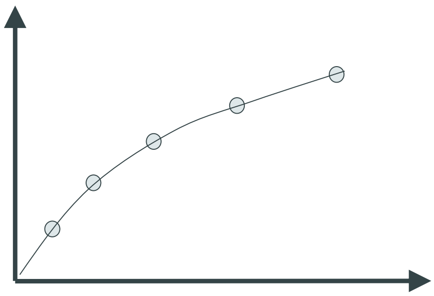
What is precision?



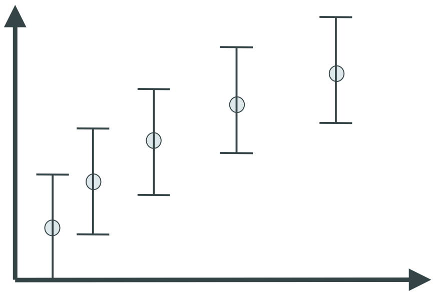
A Quick Scenario



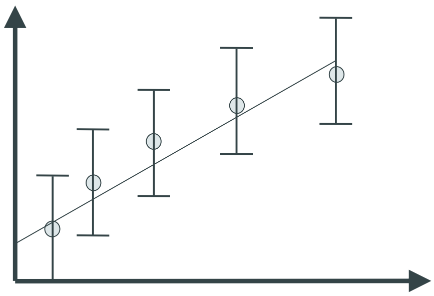
A Quick Scenario



A Quick Scenario



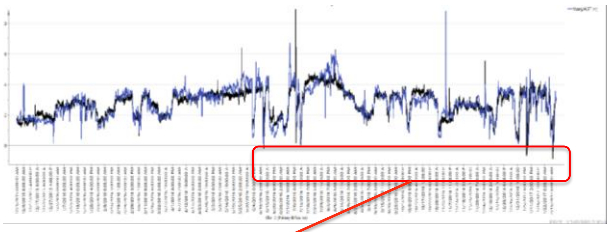
A Quick Scenario

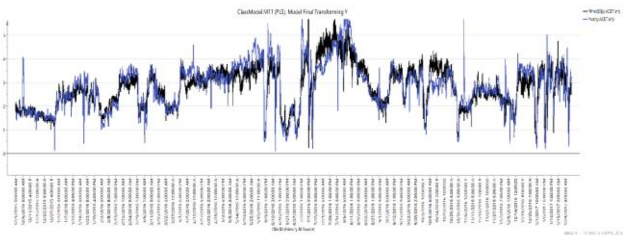


Variable transformation

Raw Y variable

Log(Y)

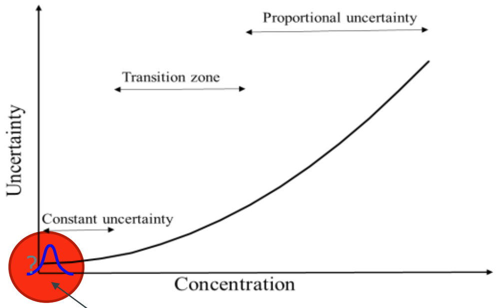




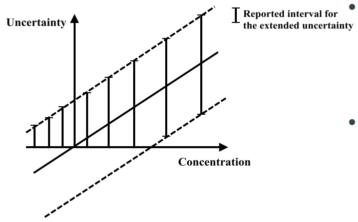
The model is predicting
negative concentration values

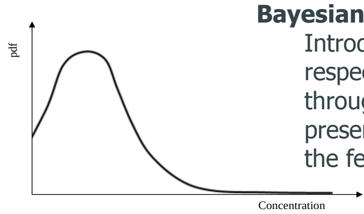
How to incorporate measurement uncertainty ?

Relative std. dev.



Measurement uncertainty near physical limits (absolute std. dev.)





Truncation of the interval

- first computing expanded uncertainty intervals using the current statistical methods (frequencist approach).
- Then, the interval is truncated in order to cover only the values in the feasible region.

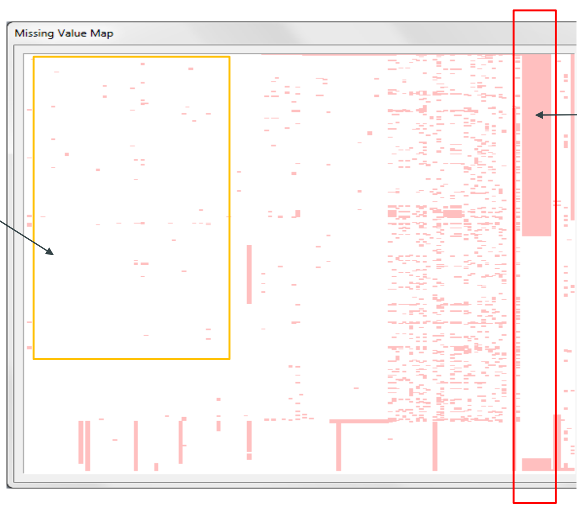
Bayesian intervals

Introduces the requirement of respecting the feasibility domain, through a prior distribution presenting finite density only in the feasible region.

M. Reis, R. Rendall, S. Chin, and L. Chiang, Challenges in the specification and integration of measurement uncertainty in the development of data-driven models for the chemical processing industry, *Industrial & Engineering Chemistry Research*, 54 (37):9159-9177, 2015.

Missing value estimation

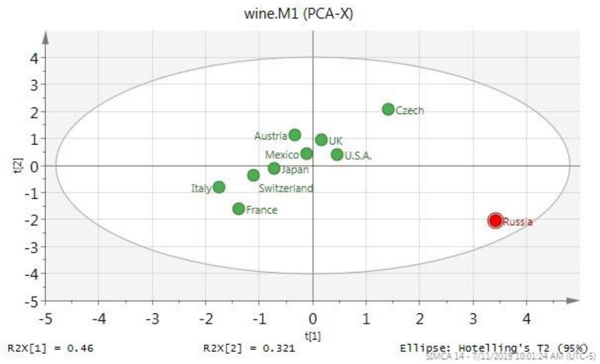
Random missing values are manageable: the NIPALS (PLS) algorithm works well



Systematic missing value are problematic and there is no one-size-fit-all solution: Start with:

1) Remove these variables
2) Use only data points with non-missing values

Outliers

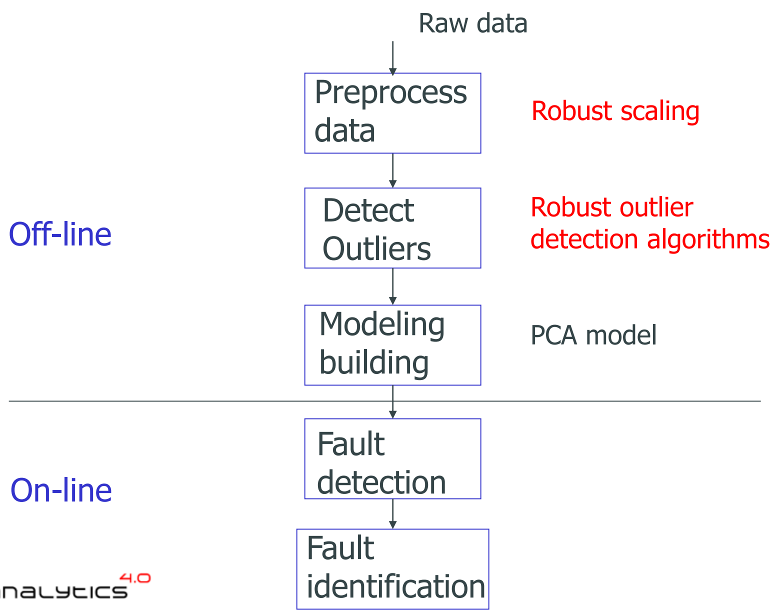


Pop quiz

What is the first thing you do when you see an outlier?

A. Eliminate the point
B. Investigation
C. Assume it is not an issue
D. Correct the number

Process monitoring workflow



Dow | Analytics 4.0
Data Driven Decision Making

29

Pre-processing methods that are less sensitive to outliers

Auto scaling (gold standard)

- Data are mean centered and then divided by the standard deviation
- With the presence of outliers, mean and standard deviation are biased

Robust scaling

- Replace mean with median
- Replace standard deviation with MAD (median absolute deviation from median)

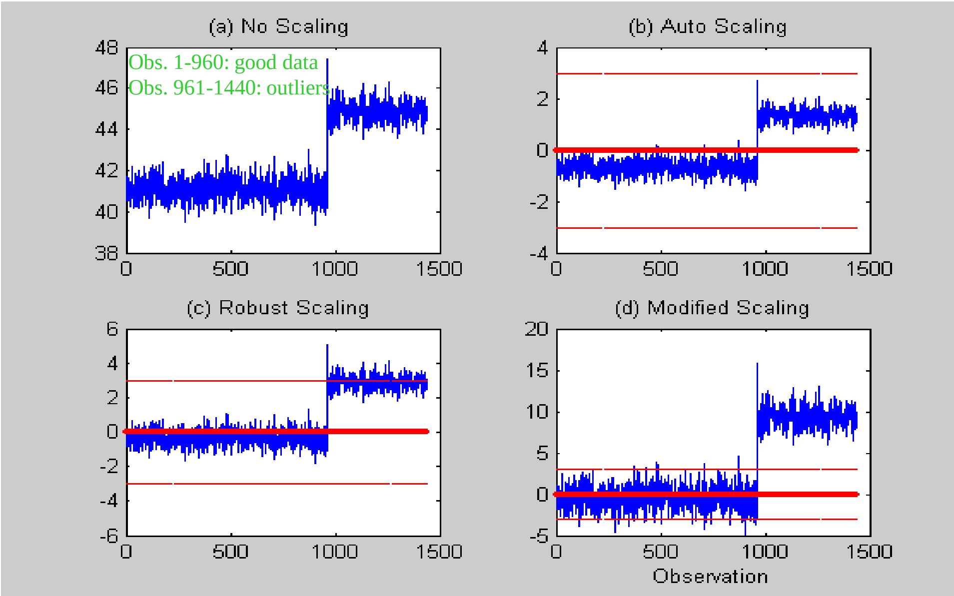
Dow modified scaling

- For each variable, find the *n/2* observations that are closest to the median
- Use these *n/2* observations to determine median and standard deviation

L. Chiang, R. Pell, and M.B. Seasholtz, Exploring process data with the use of robust outlier detection algorithms, *Journal of Process Control*, 13:437-449, 2003.

Dow | Analytics 4.0
Data Driven Decision Making

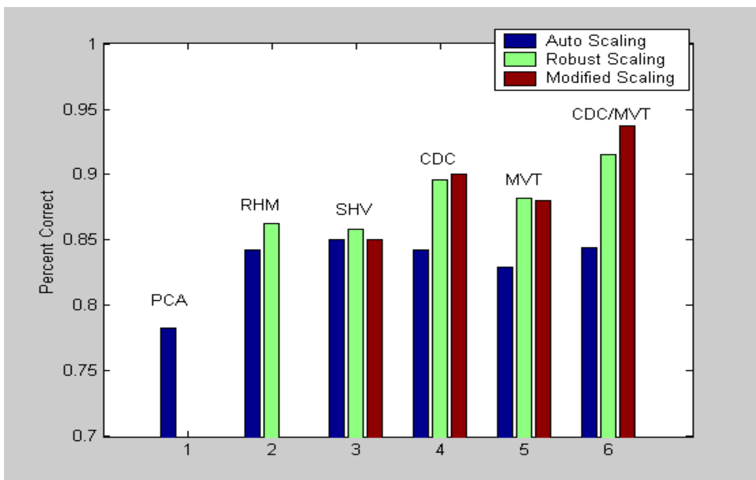
30



Dow | Analytics 4.0
Data Driven Decision Making

Robust outlier detection algorithm

- RHM (observations with small vector lengths after resampling)
- SHV (observations that are close together)
- CDC (observations that are close to the mean)
- MVT (observations with small T^2 stat. after iterations)
- CDC/MVT (use CDC in the initial step for MVT)



Dow | Analytics 4.0
Data Driven Decision Making

Outlier detection and data preprocessing summary

Outlier exclusion → invalid values, e.g. clamp transform for valid outliers

Missing data → NIPAS (PLS) algorithm works well for random missing values

Variable transformation → log

Normalization → mean centering, variance standardization

Robust statistics → less sensitive to outliers

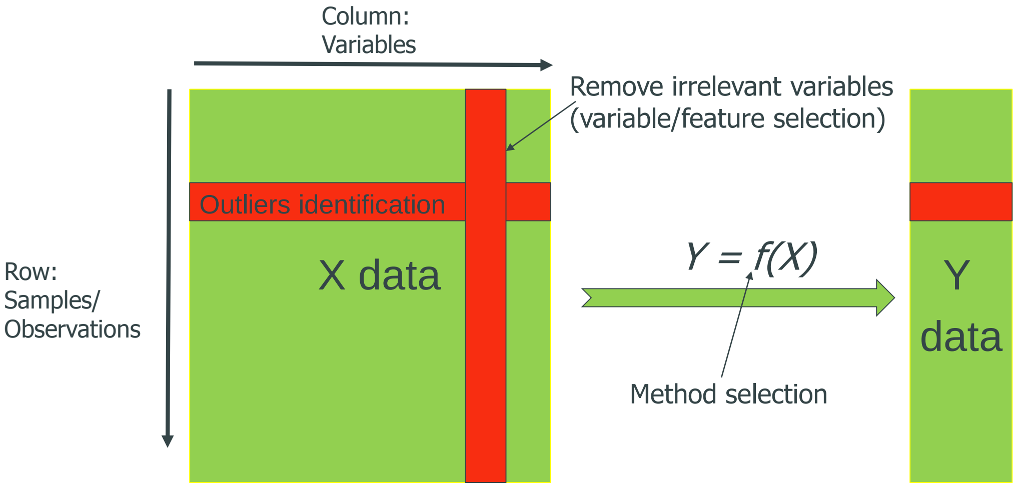
Use your domain knowledge (e.g., incorporate uncertainty Into decision making)

References

- M. Reis, R. Rendall, S. Chin, and L. Chiang, Challenges in the specification and integration of measurement uncertainty in the development of data-driven models for the chemical processing industry, *Industrial & Engineering Chemistry Research*, 54 (37):9159-9177, 2015.
- L. Chiang, R. Pell, and M.B. Seasholtz, Exploring process data with the use of robust outlier detection algorithms, *Journal of Process Control*, 13:437-449, 2003.

3 - Industrial Experience and Tips, Interactive Discussions

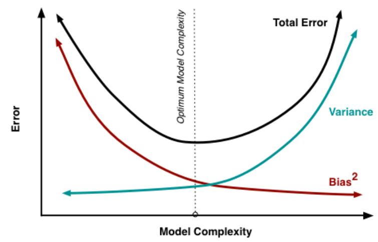
Challenge of method selection? That's why data scientist has a job!



Motivation: Why Utilize Feature Selection?

Feature Selection is the elimination of irrelevant variables from the model

1) Feature selection is relevant because it helps us address the “curse of dimensionality”:
   - Decreases the risk of overfitting
   - Improves prediction accuracy
   - Eliminates irrelevant and redundant features
   - Improves interpretability
   - Decreases computational time



2) Drastically simplify the complexity of implementation

Dow | ANALYTICS 4.0
Data Driven Decision Making

Image Source: <http://scott.fortmann-roe.com/docs/BiasVariance.html>

Current Feature Selection Approaches

Feature Selection Methods

Subset generation
Subset evaluation
Stopping criteria

Filter Methods: Use variable ranking methods. A threshold is used to remove variables

Wrapper Methods: Subset of variables is generated before and selection is based on the model performance

Embedded Methods: Subset of variables is proposed and evaluated during model development

Hybrid Methods: Combination of filter, wrapper and embedded approaches

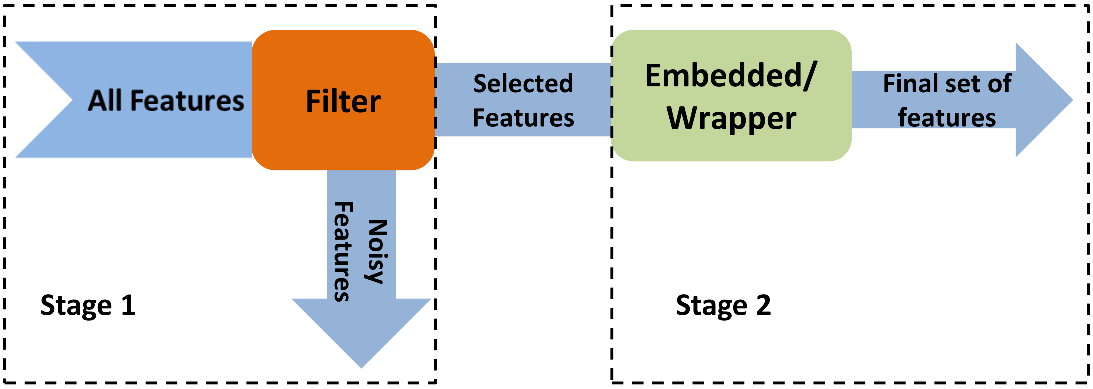
Other Methods: Unsupervised approaches

How can these methods be applied in Big Chemical Data?

Dow | ANALYTICS 4.0
Data Driven Decision Making

Wide spectrum feature selection (WiSe) Approach

1. Remove irrelevant or noisy features
2. Wrappers and/or embedded methods for further feature selection



R. Rendall, I. Castillo, A. Schmidt, S. Chin, L. Chiang, and M. Reis, Wide spectrum feature selection (WiSe) for regression model building, *Computers and Chemical Engineering*, 121:99-110, 2019.

Dow | analytics 4.0
Data driven decision making

Stage 1: Ranking Relevance to Response Variable

Ranking methods are helpful to understand the relevance of a feature

Most utilized filtering methods are entropy-based and statistical

- Information Gain (IG)
- Gain Ratio (GR)
- **Symmetrical Uncertainty (SU)**
- Mutual Information (MI)

- **Pearson's correlation**
- Chi-square test
- **Spearman's Correlation**

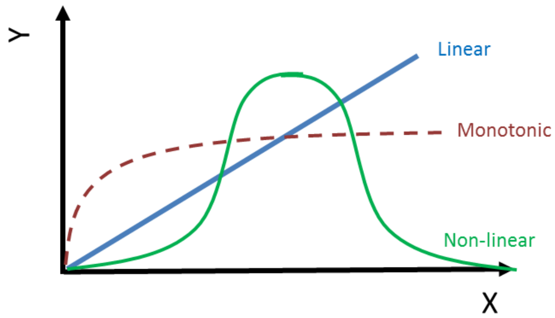
Filtering methods can be applied without significant computational burden (Thousands of variables)

Dow | analytics 4.0
Data driven decision making

Stage 1: Selected Filters

Univariate filters to efficiently remove noisy features

- Pearson correlation for linear relationships
- Spearman correlation for monotonic relationships
- Symmetrical Uncertainty (SU) for non-linear relationships
- Combinations of the aforementioned methods

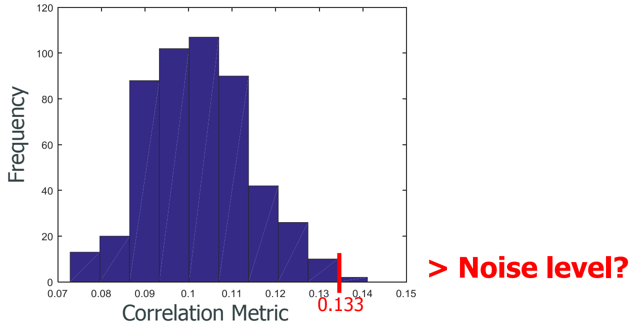


Dow | Analytics 4.0
Data Driven Decision Making

Stage 1: Determining the Threshold for Removal

Noise levels in the data are estimated utilizing random permutations

- Compute correlation metric for 100 random shuffles, using all features
- Estimate the p-value for each feature
- Select feature if p-value is below 0.2

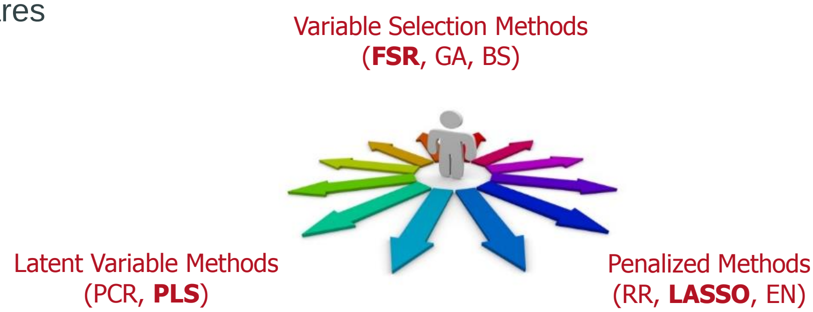


Dow | Analytics 4.0
Data Driven Decision Making

Stage 2: Model Building

The second step of feature selection is based on wrapper and embedded methods

- Forward Stepwise Regression
- LASSO
- Partial Least Squares



Dow | Analytics 4.0
Data Driven Decision Making

Dow data set

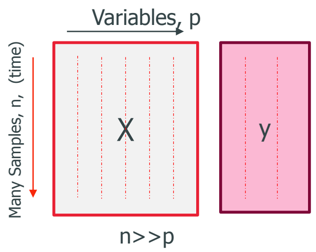
Industrial batch process of a functionalized silicone polymer

- Continuous quality parameter as a Y variable
- 29 batch conditions, 7 un-aligned batch trajectories of ~100 points
- >600 batches

Dow | Analytics 4.0
Data Driven Decision Making

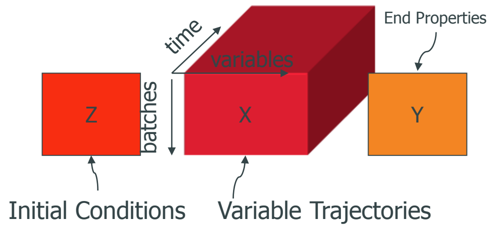
Our classic data-driven analysis problems

Continuous Process



- Steady-state continuous processes
- Dynamic behavior
- Correlated variables
- Usually for quality prediction

Batch Process



- Usually transient and dynamic
- Data variety / volume challenge, need to match context, multi-dimensional data
- High frequency, large volume datasets

R. Rendall, B. Lu, I. Castillo, S. Chin, L. Chiang, and M. Reis, A Unifying and Integrated Framework for Feature Oriented Analysis of Batch Processes, *Ind. Eng. Chem. Res.*, 56: 8590–8605, 2017.

Generated Features for industrial data set

Features: 29 batch conditions + SPA features for 7 trajectories

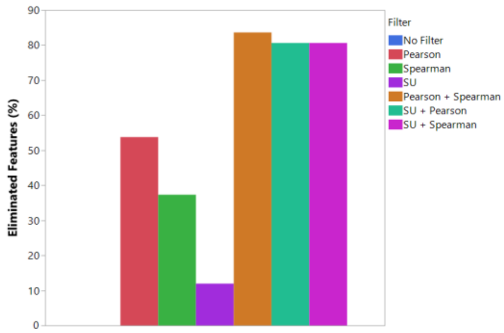
The following features are computed usng Statistical Pattern Analysis (SPA):

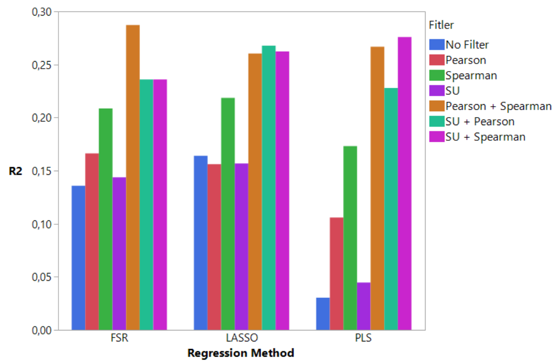
- Means of all process variables
- Variance of all process variables
- Skewness
- Kurtosis
- Covariance between variables

J. Wang and QP. He, Multivariate Statistical Process Monitoring Based on Statistics Pattern Analysis, *Ind. Eng. Chem. Res.*, 49: 7858-7869, 2010.

Results for Case 2 – Industrial Dataset

In this data, the filter combinations eliminated the most features and also resulted in the best model performance on test data





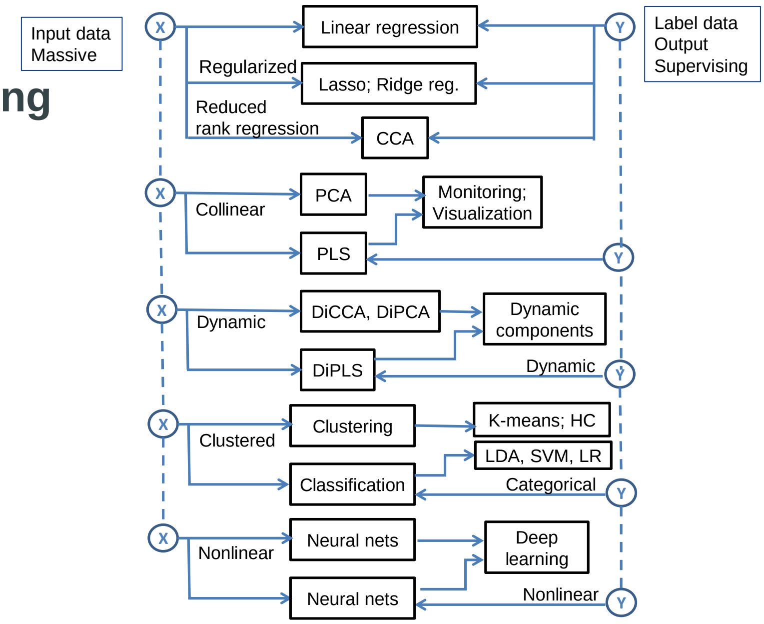
DOW | ANALYTICS 4.0
DATA DRIVEN DECISION MAKING

Conclusions

1. Filters efficiently reduced dimensionality in the first stage
    - Although performance is dependent on the dataset, most of the important predictors were selected
    - Many irrelevant variables were removed
2. Usefulness of eliminating features was demonstrated on an industrial dataset
    - The adoption of filters led to improved prediction performance across the three regression methods
    - Interpretation of which features are selected can bring insights

DOW | ANALYTICS 4.0
DATA DRIVEN DECISION MAKING

Machine Learning Algorithms



SJ. Qin and L. Chiang, Advances and opportunities in machine learning for process data analytics, *Computers and Chemical Engineering*, 126:465-473, 2019.

References

- R. Rendall, B. Lu, I. Castillo, S. Chin, L. Chiang, and M. Reis, A Unifying and Integrated Framework for Feature Oriented Analysis of Batch Processes, *Ind. Eng. Chem. Res*., 56: 8590−8605, 2017.
- R. Rendall, I. Castillo, A. Schmidt, S. Chin, L. Chiang, and M. Reis, Wide spectrum feature selection (WiSe) for regression model building, *Computers and Chemical Engineering*, 121:99-110, 2019.
- J. Wang and QP. He, Multivariate Statistical Process Monitoring Based on Statistics Pattern Analysis, *Ind. Eng. Chem. Res*., 49: 7858-7869, 2010.
- SJ. Qin and L. Chiang, Advances and opportunities in machine learning for process data analytics, *Computers and Chemical Engineering*, 126:465-473, 2019.

Data Driven Decision Making

3 - Industrial Experience and Tips, Interactive Discussions

3.1 Visualization

3.2 Outlier detection and data preprocessing

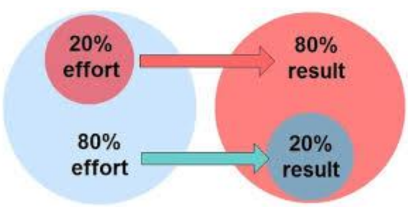
3.3 Method selection

3.4 How good is good enough? Industrial tips and tricks of the trade

3.5 Industrial case studies

"All Models are wrong, but some are useful"

George Box

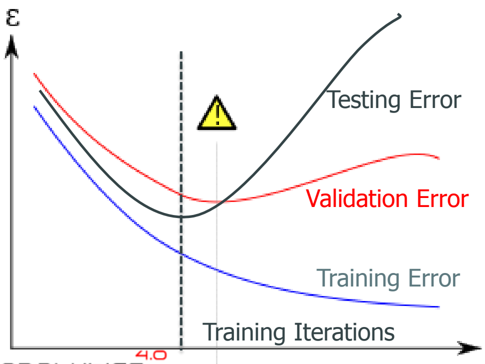


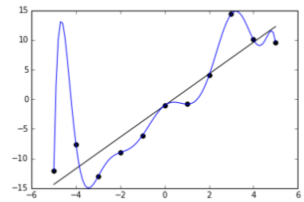
So, how good is good enough?

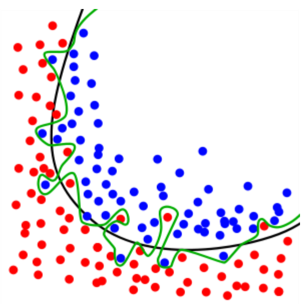
How to avoid overfitting

Tune model parameters	Tune hyper-parameters	True Evaluation
Training	Validation	Testing

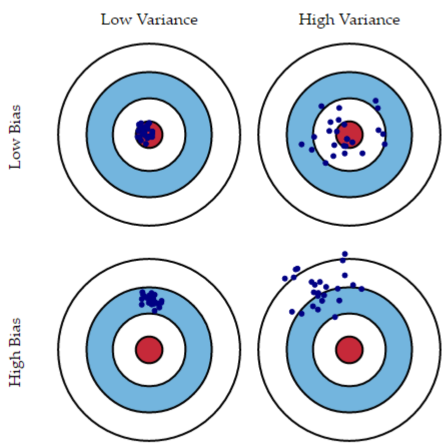
*Do not use the testing data **AT ALL** when you are developing the model



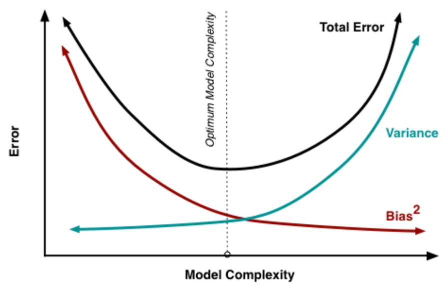




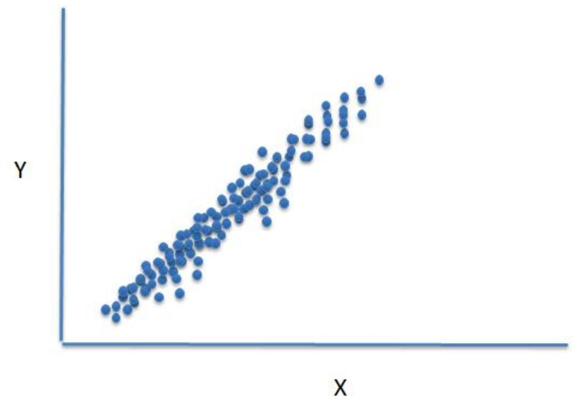
Bias / Variance Trade-Off



- Include more variables
- Transform/generate variables
- Add more components
- Remove outliers / ill-fitting data
- Use nonlinear techniques



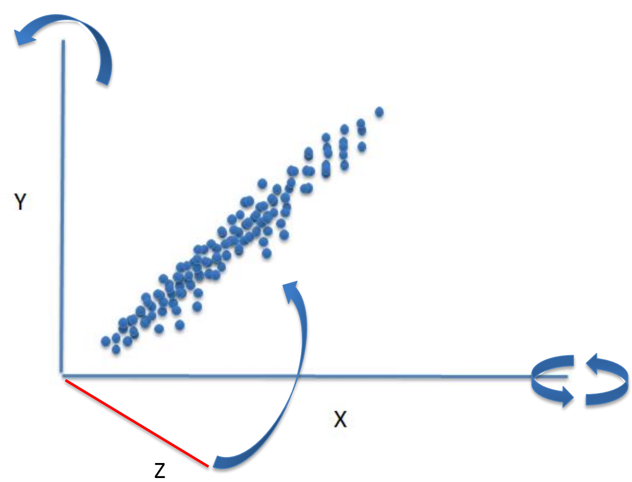
Another Dow example



DOW | ANALYTICS 4.0
DATA DRIVEN DECISION MAKING

55

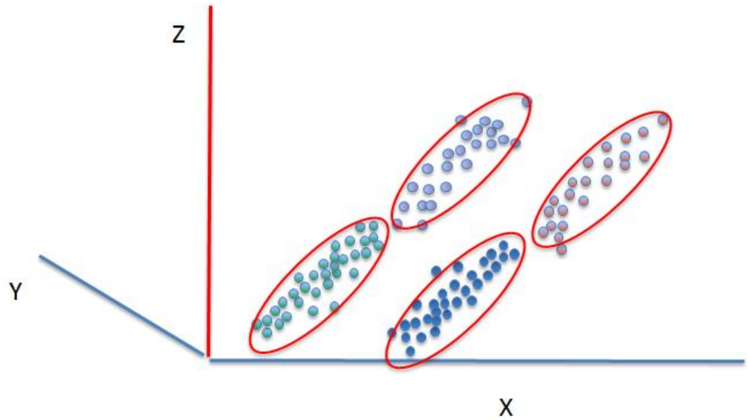
New data



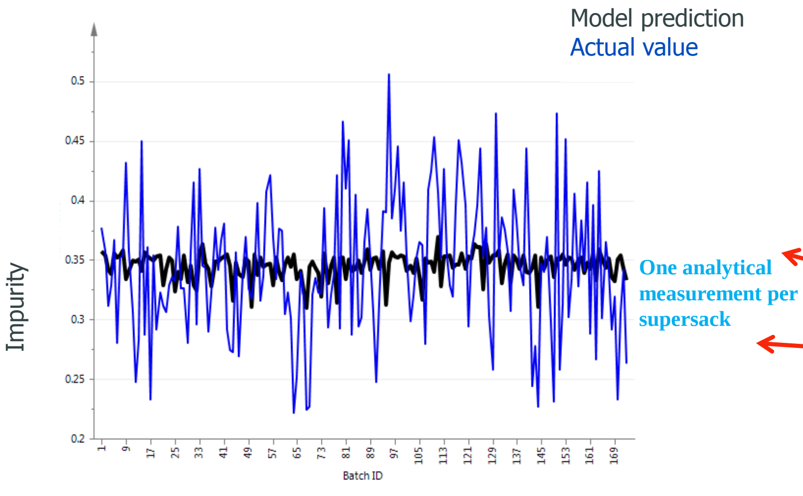
DOW | ANALYTICS 4.0
DATA DRIVEN DECISION MAKING

56

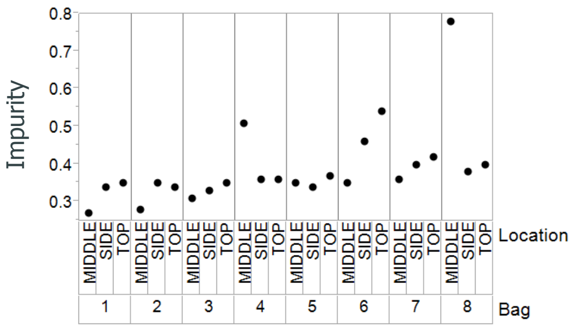
New data means new insight



One of the worst models I've seen at Dow



Traditional Approach

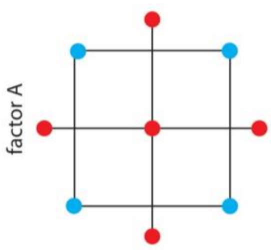


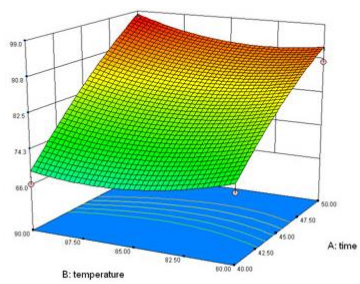
Who contributes more
to the variability?

Lab Process

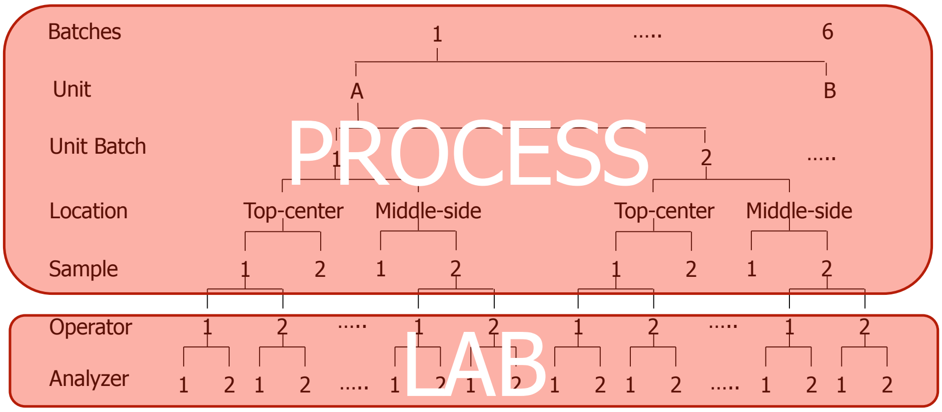
How to Improve Data Quality?

Design of Experiments





Design Of Experiments (DOE)

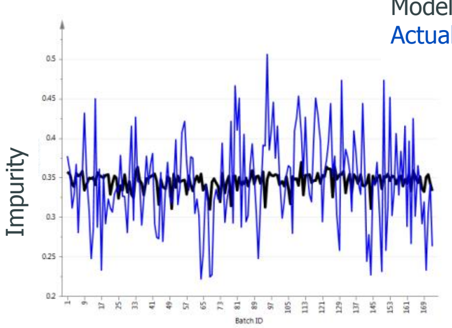


DOW | ANALYTICS 4.0
DATA DRIVEN DECISION MAKING

Systematic Statistical Approach

Historical data

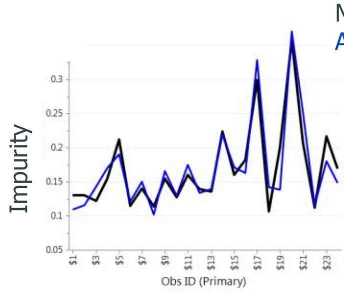
Model prediction
Actual value



Correlation Analysis

DOE plant data

Model prediction
Actual value



Correlation Analysis
+
DOE

DOW | ANALYTICS 4.0
DATA DRIVEN DECISION MAKING

Results: Process Variables that cause the difference performances in one major unit

Speed of propeller in Unit

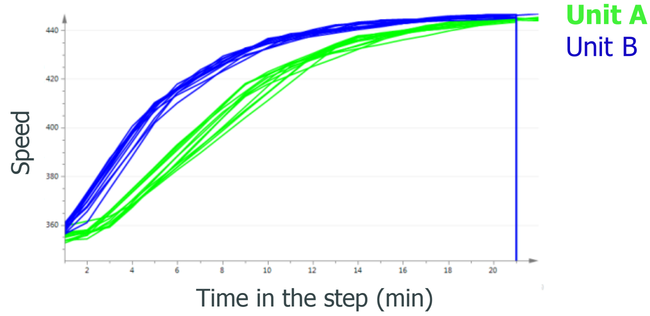
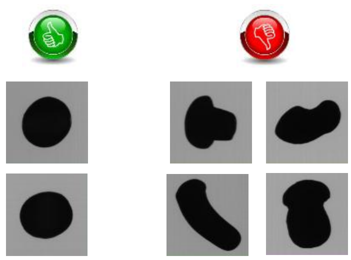


Image Classification at Dow

Good **vs** Bad



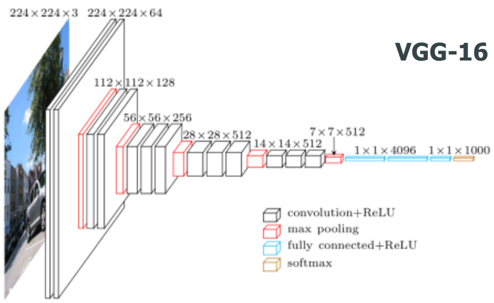
The shape of plastic pellets is a major quality factor

R. Rendall, M. Broadway B. Lu, I. Castillo, L. Chiang, B. Colegrove, and M. Reis, Image-based Manufacturing Analytics: Improving the Accuracy of an Industrial Pellet Classification System using Deep Neural Networks, *Chemometrics and Intelligent Laboratory Systems*, 180: 26-35, 2018.

Deep Neural Networks for Image Classification

A deep neural network contains many layers:

- Multiple layers allow the learning of high level features
- Each layer computes a specific function



VGG-16 [1]

Interesting Points of DNN:

- **Convolutional** layers and other type of layers
- Better optimization tools and other developments (ReLU activation, batch normalization, **transfer learning**, dropout, etc.)

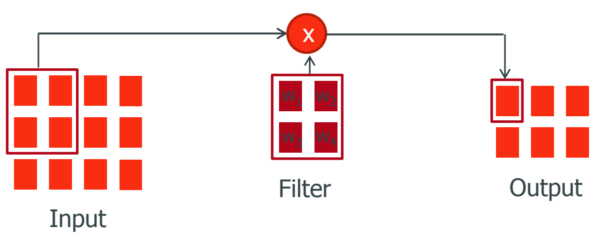
[1] - Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition (2015)

Deep Neural Networks for Image Classification

Convolutional layers:

- Contain filters that convolve with the input, outputting a matrix
- The parameters of the filter are optimized with training data

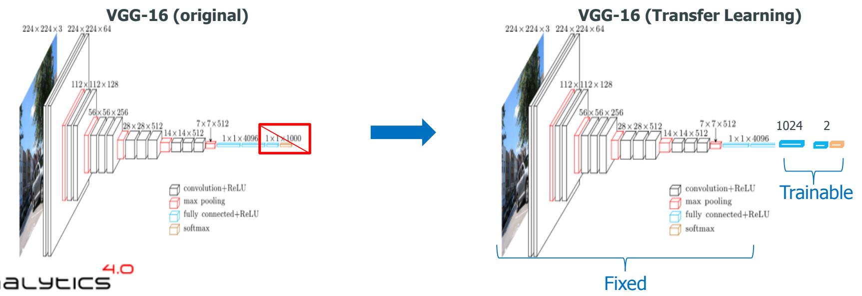
2D convolution



Deep Neural Networks for Image Classification

Transfer Learning:

- A pre-trained network is modified to a different classification task
- Relevant features in the original domain tend to be useful in the target domain



Dow | ANALYTICS 4.0
DATA DRIVEN DECISION MAKING

Deep Neural Networks for Image Classification

Two deep neural networks were tested

Simpler Deep Neural Network (SDNN)

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 46, 46, 32)	320
conv2d_2 (Conv2D)	(None, 44, 44, 64)	18496
max_pooling2d_1 (MaxPooling2	(None, 22, 22, 64)	0
dropout_1 (Dropout)	(None, 22, 22, 64)	0
flatten_1 (Flatten)	(None, 30976)	0
dense_1 (Dense)	(None, 128)	3965056
dropout_2 (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 2)	258

Total params: 3,984,130
Trainable params: 3,984,130
Non-trainable params: 0

VGG-16 with Transfer Learning

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 96, 96, 3)	0
block1_conv1 (Conv2D)	(None, 96, 96, 64)	1792
block1_conv2 (Conv2D)	(None, 96, 96, 64)	36928
block1_pool (MaxPooling2D)	(None, 48, 48, 64)	0
block2_conv1 (Conv2D)	(None, 48, 48, 128)	73856
block2_conv2 (Conv2D)	(None, 48, 48, 128)	147584
block2_pool (MaxPooling2D)	(None, 24, 24, 128)	0
⋮		
dense_1 (Dense)	(None, 1024)	525312
dense_2 (Dense)	(None, 2)	2050

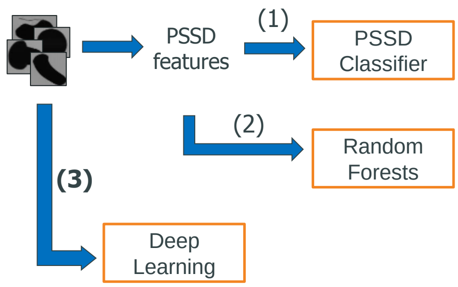
Total params: 15,242,050.0
Trainable params: 527,362.0
Non-trainable params: 14,714,688.0

Dow | ANALYTICS 4.0
DATA DRIVEN DECISION MAKING

Classification Methods Tested

Process experts manually labelled ~6000 images:

- These were split in training, validation and test sets
- Different classifiers were tested
  - PSSD
  - Random Forests
  - Deep Neural Networks

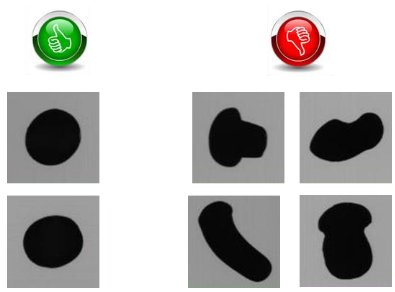


Dow | Analytics 4.0
Data Driven Decision Making

Classification Results

Good vs Bad Pellets

- Training: 2961 samples
- Validation: 1777 samples
- Test: 1185 samples



Accuracy

Set	PSSD[1]	Random Forests[1]	SDNN	SDNN[2]	Transfer Learning (VGG-16)[2]
Training	0.816	1	0.98	0.964	0.971
Validation	0.817	0.941	0.913	0.956	0.966
Test	0.805	0.937	0.917	0.957	0.967

[1] Approaches based on features
[2] Uses sample augmentation techniques

Dow | Analytics 4.0
Data Driven Decision Making

Complex interaction between analytics and expert knowledge

Current Model Residuals

Neither one alone is sufficient

Analytics Model

Expert Domain Knowledge

Both components are needed to improve model quality

Dow | Analytics 4.0
Data Driven Decision Making

71

Analytics culture change

Chemometric modelers
- **Innovation**
- **Advanced analytics and programming tools**

400+ Data scientists
- **Collaboration**
- **Data acumen**
  (Special analytics and programming tools)

35,000+ Dow people
- **Foundation**
- **Data literacy/acuity**
  (Practitioner analytics tools)

500,000+ US STEM graduates
- **Art of the possible**
- **Integrate data analytics into ChE curriculum**

72

Qin and Chiang, 2019
Chiang et al., 2017

References

- R. Rendall, M. Broadway B. Lu, I. Castillo, L. Chiang, B. Colegrove, and M. Reis, Image-based Manufacturing Analytics: Improving the Accuracy of an Industrial Pellet Classification System using Deep Neural Networks, *Chemometrics and Intelligent Laboratory Systems*, 180: 26-35, 2018.
- SJ. Qin and L. Chiang, Advances and opportunities in machine learning for process data analytics, *Computers and Chemical Engineering*, 126:465-473, 2019.
- L. Chiang, B. Lu, and I. Castillo, Big data analytics in chemical engineering, *Annual Review of Chemical and Biomolecular Engineering*, 8:4.1-4.23, 2017.

3 - Industrial Experience and Tips, Interactive Discussions

3.1 Visualization

3.2 Outlier detection and data preprocessing

3.3 Method selection

3.4 How good is good enough? Industrial tips and tricks of the trade

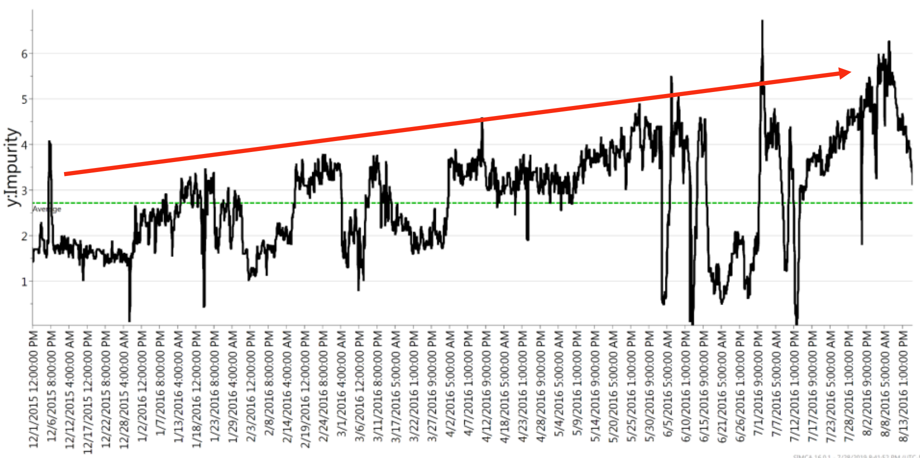
3.5 Industrial case studies

3.5 Industrial Case Studies

- Case Study 1: Impurity Estimation
- Case Study 2: Quality Classification

Case Study 1: Impurity Estimation

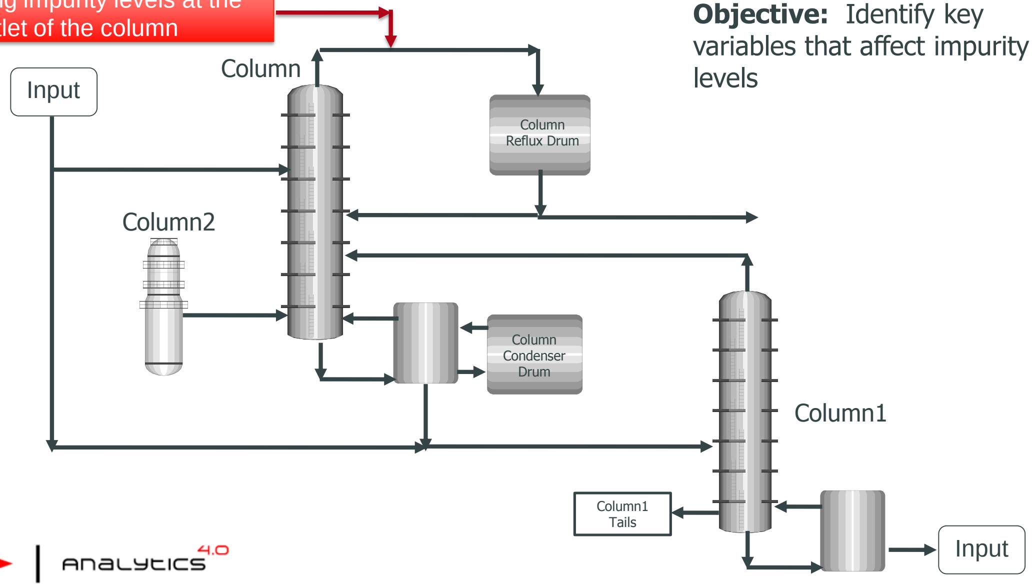
Impurity levels are constantly increasing, affecting production rate and the catalyst life of the reactor



Case Study 1: Block Diagram

Estimating impurity levels at the outlet of the column

Objective: Identify key variables that affect impurity levels



Case Study Variables

Column Variables	Column1 Variables	Column2 Variables
x1:Column Reflux Flow	x22: Column1 Base Concentration	x36: Column2 Recycle Flow
x2:Column Tails Flow	x23: Flow from Input to Column1	x37: Column2 Tails Flow to Column
x3:Input to Column Bed 3 Flow	x24: Column1 Tails Flow	x38: Column2 Calculated DP
x4:Input to Column Bed 2 Flow	x25: Column1 Tray DP	x39: Column2 Steam Flow
x5:Column Feed Flow from Column2	x26: Column1 Head Pressure	x40: Column2 Tails Flow
x6:Column Make Flow	x27: Column1 Base Pressure	
x7:Column Base Level	x28: Column1 Base Temperature	
x8:Column Reflux Drum Pressure	x29: Column1 Tray 3 Temperature	
x9:Column Condenser Reflux Drum Level	x30: Column1 Bed 1 Temperature	
x10:Column Bed1 DP	x31: Column1 Bed 2 Temperature	
x11:Column Bed2 DP	x32: Column1 Tray 2 Temperature	
x12:Column Bed3 DP	x33: Column1 Tray 1 Temperature	
x13:Column Bed4 DP	x34: Column1 Tails Temperature	
x14:Column Base Pressure	x35: Column1 Tails Concentration	
x15:Column Head Pressure		
x16:Column Tails Temperature		
x17:Column Tails Temperature 1		
x18:Column Bed 4 Temperature		
x19:Column Bed 3 Temperature		
x20:Column Bed 2 Temperature		
x21:Column Bed 1 Temperature		
Avg_Outlet_Impurity		
Avg_Delta_composition column		
y:Impurity		
Column reflux/feed		
Column make/reflux		

Data available:

ImpurityDataset_Validation.xlsx

ImpurityDataset_Traning.xlsx

Workflow

Outlier detection and data preprocessing

↓

Visualizing data

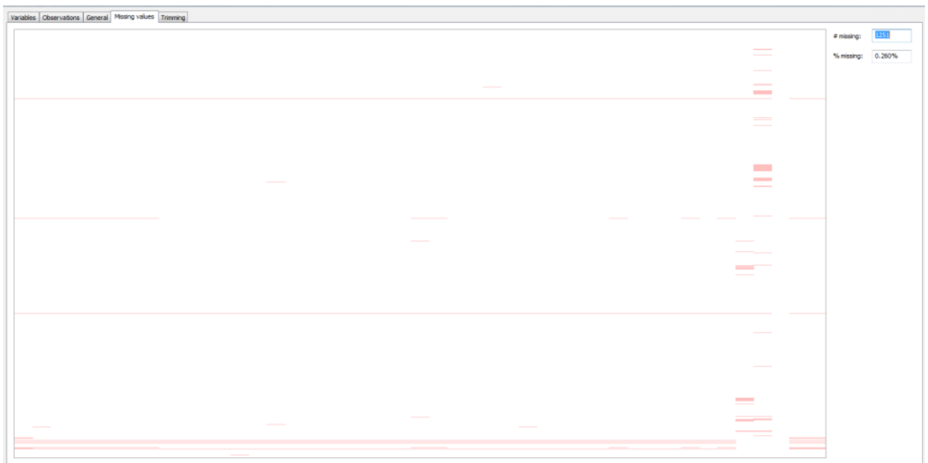
↓

Model development: Variable selection

↓

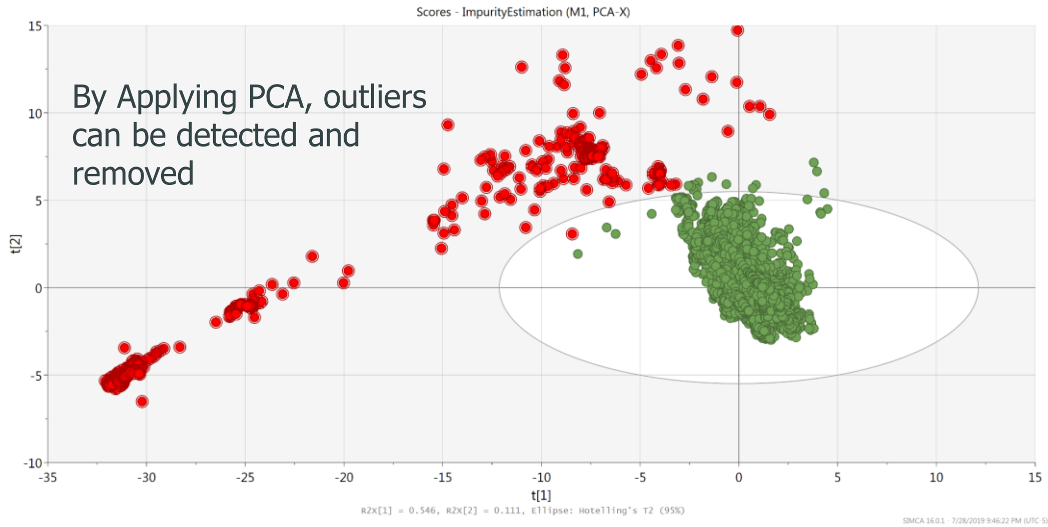
Model Validation: Is the model good enough?

Missing values

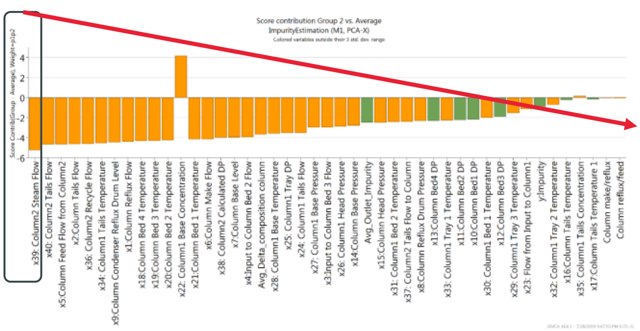


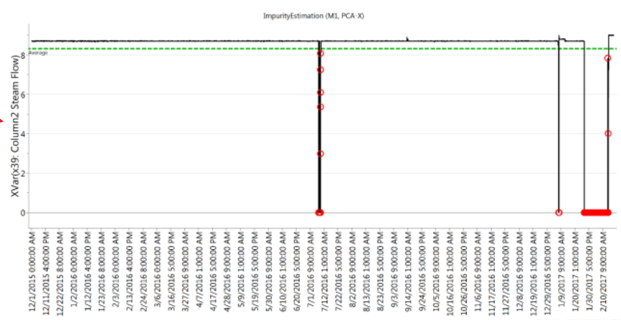
Variables	Missing (%)
Avg_Delta_composition column	2.719
Avg_Outlet_Impurity	0.906
x1:Column Reflux Flow	0.280
Column reflux/feed	0.280
Column make/reflux	0.280
x23: Flow from Input to Column1	0.215
x2:Column Tails Flow	0.206
x3:Input to Column Bed 3 Flow	0.196
x4:Input to Column Bed 2 Flow	0.196
x5:Column Feed Flow from Column2	0.196
x6:Column Make Flow	0.196
x7:Column Base Level	0.196
x8:Column Reflux Drum Pressure	0.196
x15:Column Head Pressure	0.196
x24: Column1 Tails Flow	0.196
x34: Column1 Tails Temperature	0.196
x38: Column2 Calculated DP	0.196
x40: Column2 Tails Flow	0.196
x13:Column Bed4 DP	0.187
x27: Column1 Base Pressure	0.187
x29: Column1 Tray 3 Temperature	0.187
x9:Column Condenser Reflux Drum Level	0.178
x10:Column Bed1 DP	0.178
x11:Column Bed2 DP	0.178
x12:Column Bed3 DP	0.178
x14:Column Base Pressure	0.178
x16:Column Tails Temperature	0.178
x17:Column Tails Temperature 1	0.178
x18:Column Bed 4 Temperature	0.178
x19:Column Bed 3 Temperature	0.178
x20:Column Bed 2 Temperature	0.178
x21:Column Bed 1 Temperature	0.178
x22: Column1 Base Concentration	0.178
x25: Column1 Tray DP	0.178
x26: Column1 Head Pressure	0.178
x28: Column1 Base Temperature	0.178
x30: Column1 Bed 1 Temperature	0.178
x31: Column1 Bed 2 Temperature	0.178
x32: Column1 Tray 2 Temperature	0.178
x33: Column1 Tray 1 Temperature	0.178
x35: Column1 Tails Concentration	0.178
x36: Column2 Recycle Flow	0.178
x37: Column2 Tails Flow to Column	0.178
x39: Column2 Steam Flow	0.178
y:Impurity	0.000

Outlier Detection (Visualizing Data Utilizing PCA)

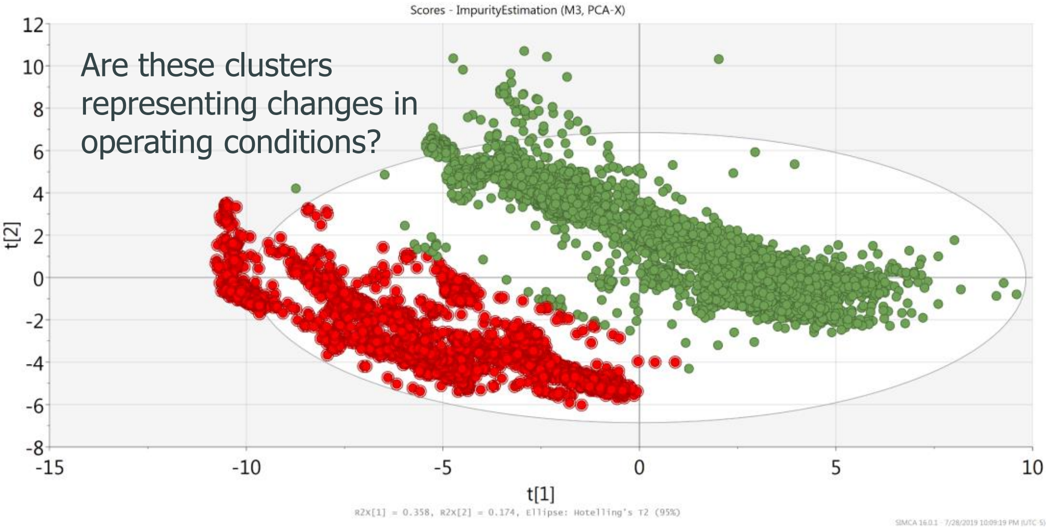


Are These Outliers?





Visualizing Data After Eliminating Outliers

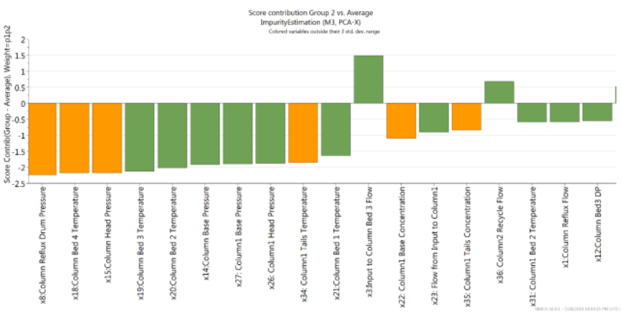


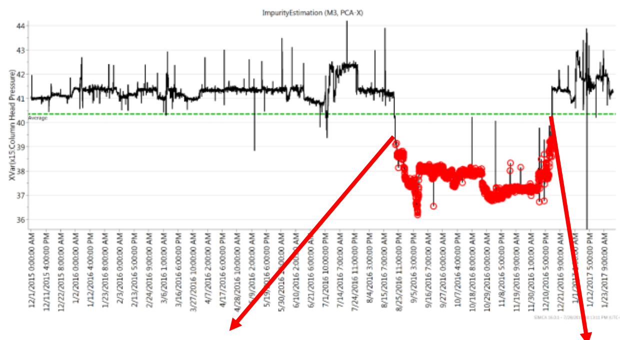
Dow | ANALYTICS 4.0
DATA DRIVEN DECISION MAKING

83

Visualizing Data After Eliminating Outliers

Significant changes in operating conditions from 8/22/2016 to 12/16/2016





Business reduced column head pressure

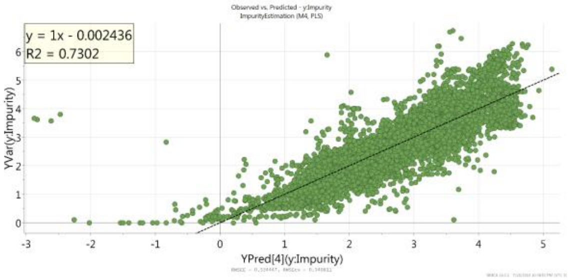
Dow | ANALYTICS 4.0
DATA DRIVEN DECISION MAKING

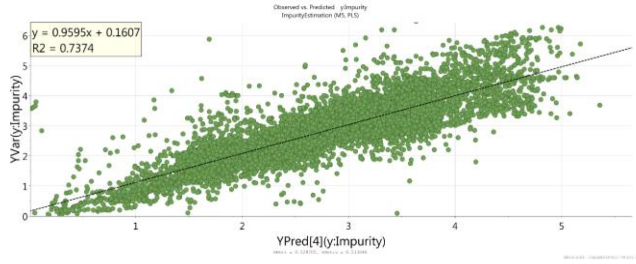
84

Model Development

The model is predicting negative values in the impurity value

The model prediction when applying a log transformation of the Y



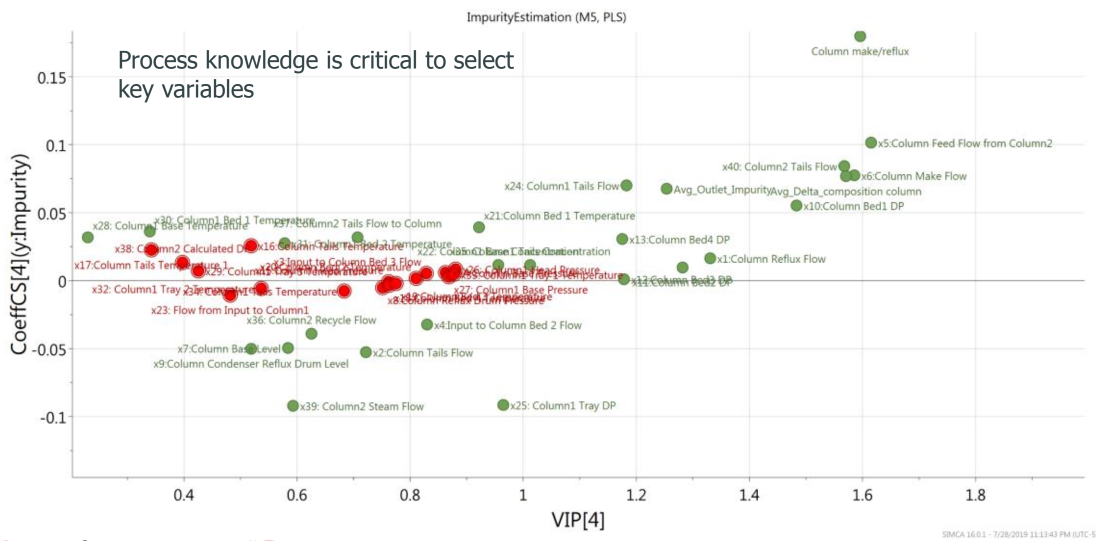


Accuracy can be improved by applying a nonlinear transformation to the output variable

ANALYTICS 4.0

Data Driven Decision Making

Variable Selection

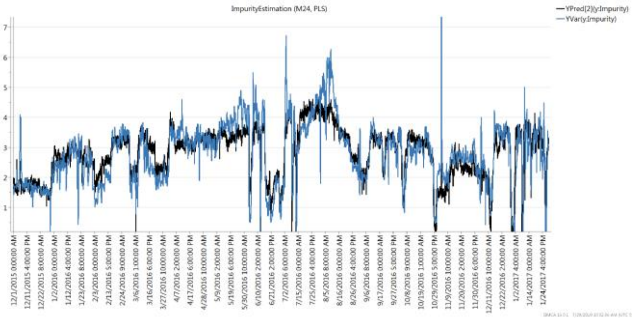


ANALYTICS 4.0

Data Driven Decision Making

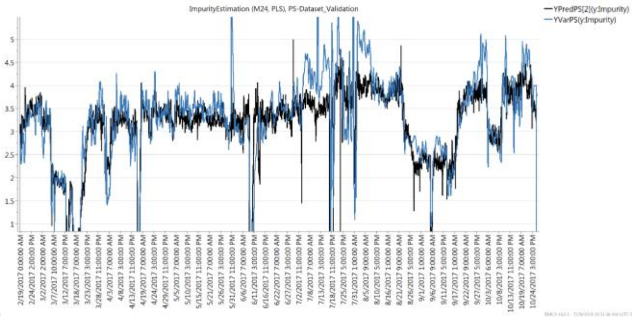
Model Results

Model Results Utilizing Training Dataset



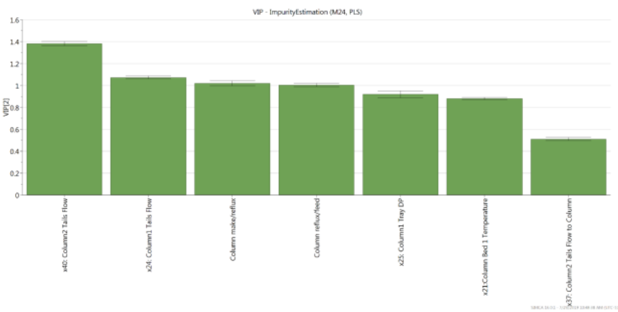
Model Results Utilizing Validation Dataset

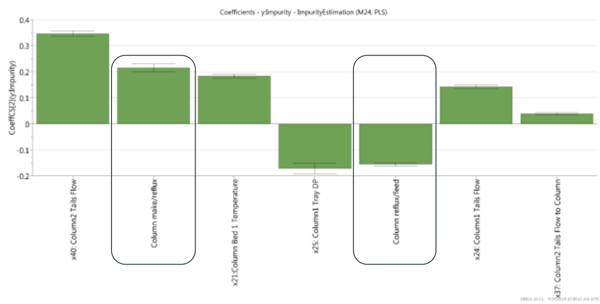
R2=0.6979



Is the model good enough? Best practice is to verify model accuracy by utilizing a validation dataset.

Model Results

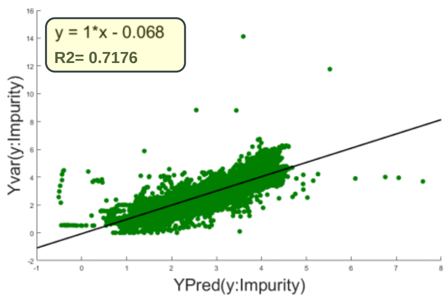


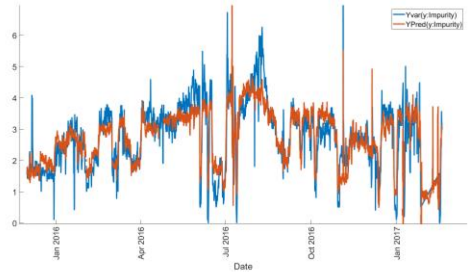


Is the model good enough? Are the selected variables in correspondence with first principles knowledge and plant operation?

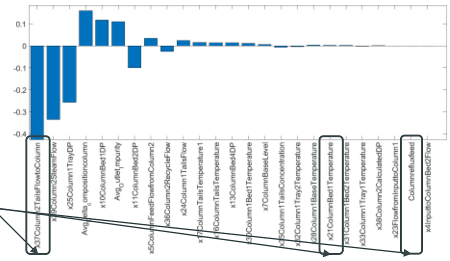
Comparing Performance Utilizing Lasso (Penalized Methods)

Model Development Results





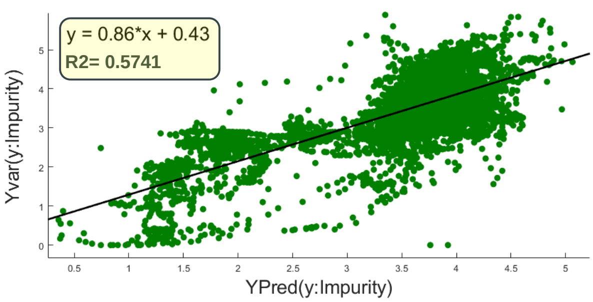
Coefficient Plot

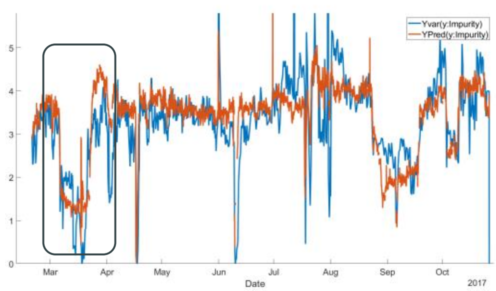


Similar performance than PLS. More variables selected by this method, requiring further elimination of variables.

Selected by PLS model

Comparing Performance with Lasso (Model Validation Results)



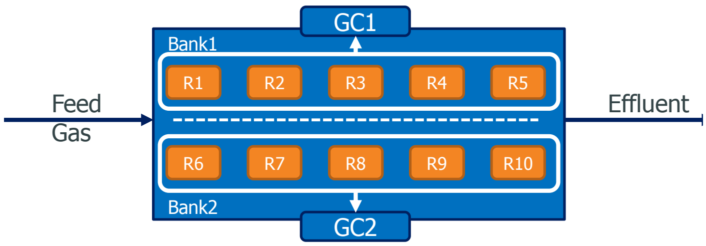


Between March and April 2017, the model is not capable to predict fast changes in the impurity. The PLS model has a better performance during the same time frame

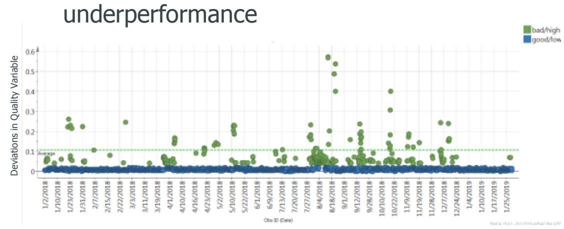
Case Study 2: Quality Classification-- Evaluating the Performance of Multiple Reactors

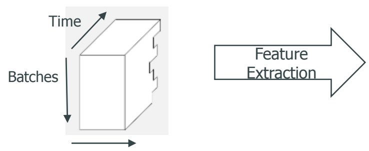
Batch operation (10 reactors, multiple catalyst, 38 variables per reactor)

Testing multiple catalyst families



Goal: Identify root cause of underperformance





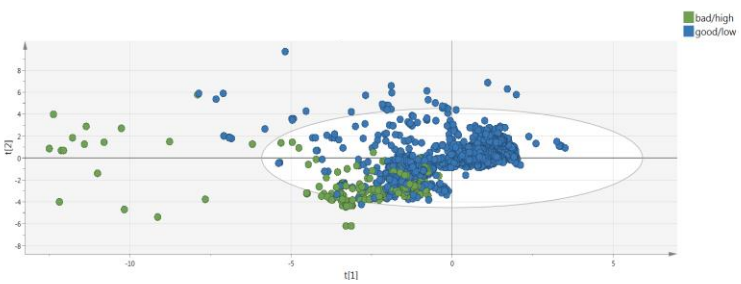
Features:

- Mean, Stdev, Skewness, Kurtosis, Pairwise Correlation
- Max, Min, Range, Medium, Slope, Area under the curve, Begin/End Delta
- Autocorrelation (lag=1)

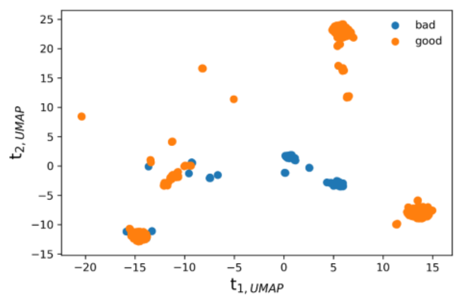
Dow | Analytics 4.0
Data Driven Decision Making

Supervised Case Study: Evaluating the Performance of Multiple Reactors

PLS-DA



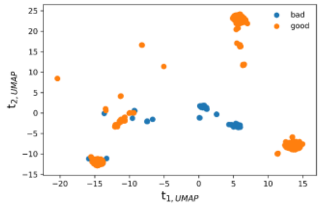
UMAP



When applying PLS-DA and UMAP, the separation of classes is unclear

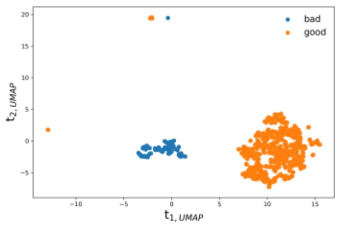
Dow | Analytics 4.0
Data Driven Decision Making

Separation Between Classes Can Be Improved by Performing Feature Selection

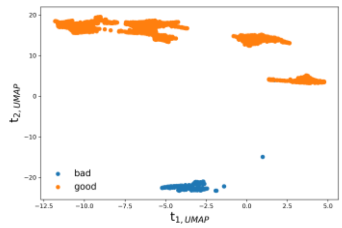


X: (2828 (observations) , 1832 (features))

Mutual Information:
Top 10 variables



PLS-DA:
Top 10 variables



Objective: Find variables that separate classes and minimize number of clusters based on the following metric:

Metric: $J = \min_i \max_j \frac{\sum_k I(y_k \in (n_j \cup c_i))}{|c_i|}$

ci refers to cluster i (|ci| is the cardinality)
nj refers to given class/label j
I(x) is the indicator function where I(x)=1 if x is true else I(x) = 0

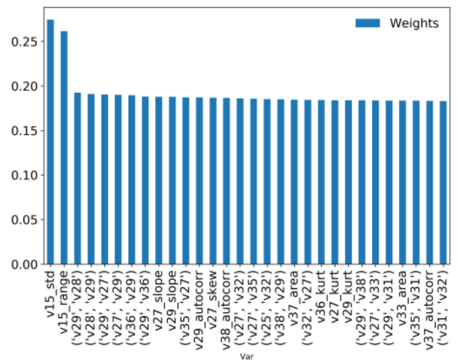
Objective: $\max_{k,n,i \leq \theta} J$

k refers to the number of features
n refers to the number of neighbors (hyper-parameter used for UMAP)
Θ is a constraint on the number of clusters

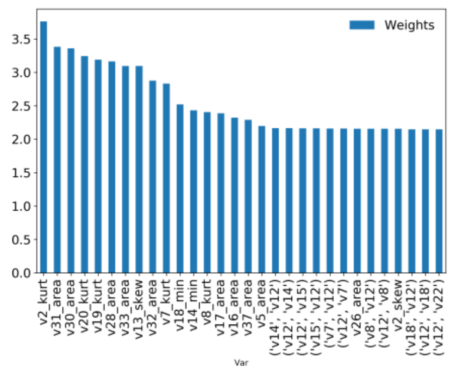
DOW | ANALYTICS 4.0
DATA DRIVEN DECISION MAKING

Selecting Relevant Features

Feature selection utilizing Mutual Information (MI)



Feature selection utilizing PLS-DA VIP



The selected features for each method are not consistent and are located in different order. Which classifier model is better?

DOW | ANALYTICS 4.0
DATA DRIVEN DECISION MAKING

Classifier Performance (Validating with new data)

Performance classifiers utilizing full feature space

Accuracy	Random Forest	PLS-DA	NN
Good	99.4%	98.0%	93.7%
Bad	54.2%	32.2%	49.2%

Bad class performance is not ideal

Performance classifiers built upon selected features

Accuracy	Top 10 Features (MI)	Top 10 Features (VIP)	Top 10 Features (RF)	Top 5 Features (RF)
Good	98.8%	96.1%	98.8%	99.2%
Bad	89.8%	39.0%	79.8%	86.4%

RF=Random Forest; MI=Mutual Information and VIP=Variable Influence of Projection based on PLS-DA

Model validation is very helpful to identify best classifier model

Dow | Analytics 4.0
Data Driven Decision Making

95

Summary – Case Studies

- Two supervised case studies were illustrated by utilizing industrial cases studies. Feature selection and model validation are key steps to evaluate model performance.
- Process knowledge is key for generating best models. Dimensionality reduction techniques are helpful to visualize high dimensional data and bring process understanding

Dow | Analytics 4.0
Data Driven Decision Making

Package Resources

Visualization: Tableau, PowerBi, Python-plotly, seaborn and matplotlib

Design of Experiment: JMP

Random Forest: Python- sklearn

Dimensionality Reduction: Python-sklearn, Matlab-toolbox by Laurens van der Maaten

PLS/PLS-DA: Sartorius-Stedim/Umetrics SIMCA, Python-pychemometrics

Deep Neural Networks: Python-keras (tensorflow)