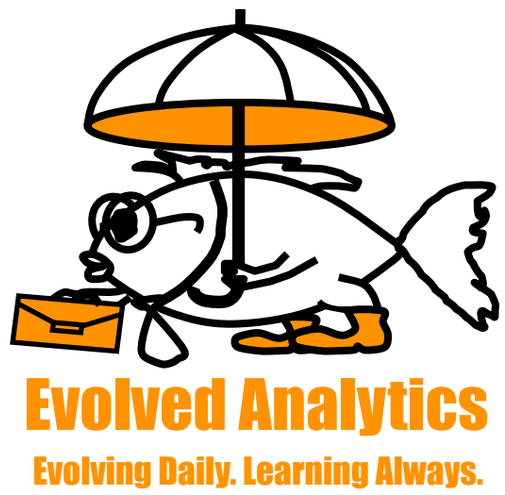
Life
(Data Science)
Doesn't Have to be Hard

Mark Kotanchek
Evolved Analytics LLC
www.evolved-analytics.com



I never guess. It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.

– Sir Arthur Conan Doyle (1859 - 1930), The Sign of Four, A Scandal in Bohemia

AI ⇒ Augmented Intelligence

- Recognized ~20 years ago that we needed better tools to extract actionable insight from real-world multivariate data

- The human needs to be in the loop with the machine serving not the converse

- If an analysis component can be automated, it should be automated!

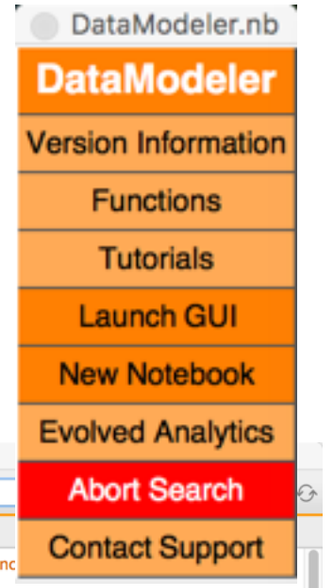
Real-World Data

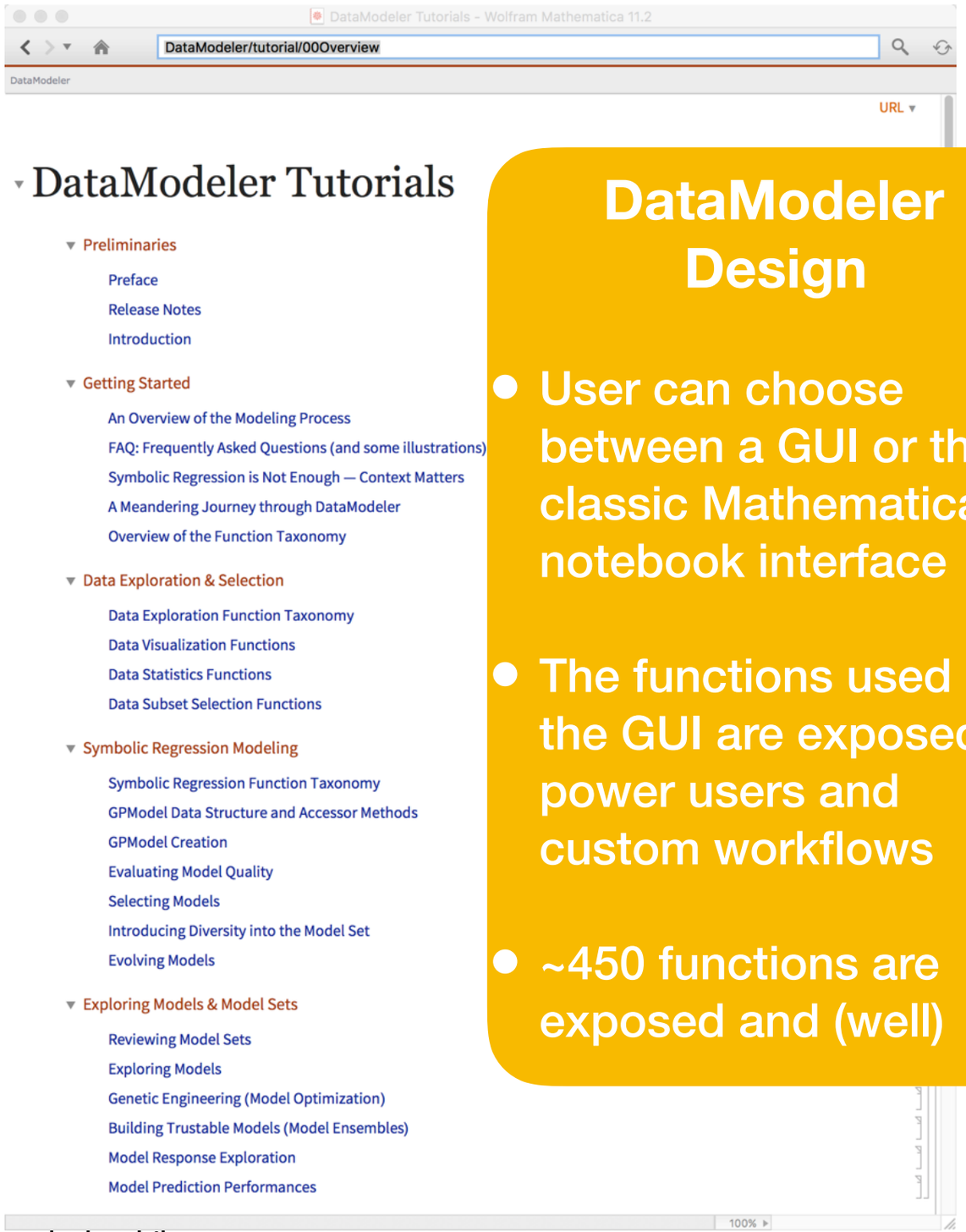
- Multivariate
- Correlated variables
- Unbalanced (closed-loop)
- Nonlinear
- Missing Data
- Wrong Data

Questions & Needs

- Exploratory Data Analysis
- Which variables matter?
- Are there important metavariables?
- What variable combinations are useful?
- How to handle extrapolation and system changes
- How to deploy models?

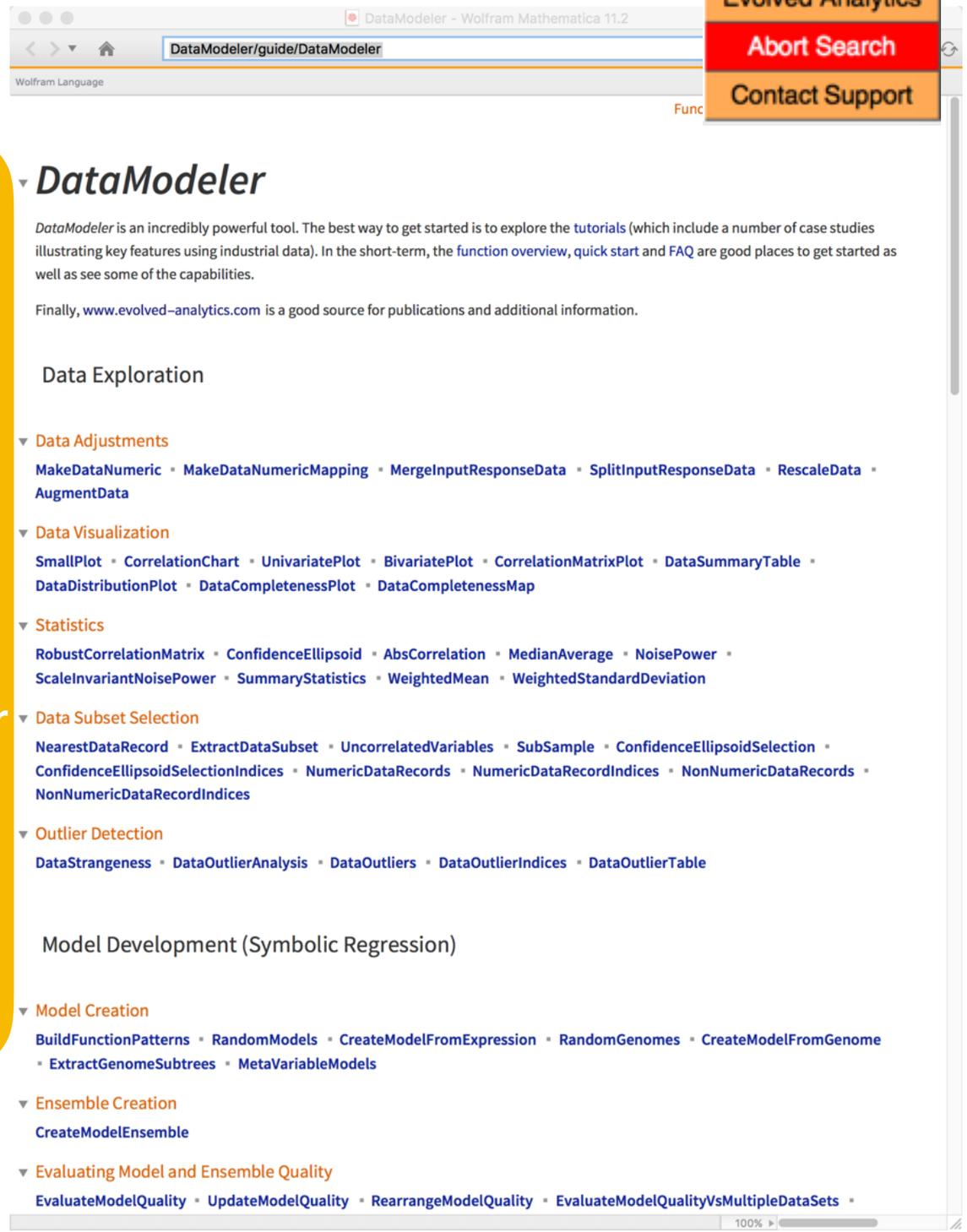
GUI and/or Notebook



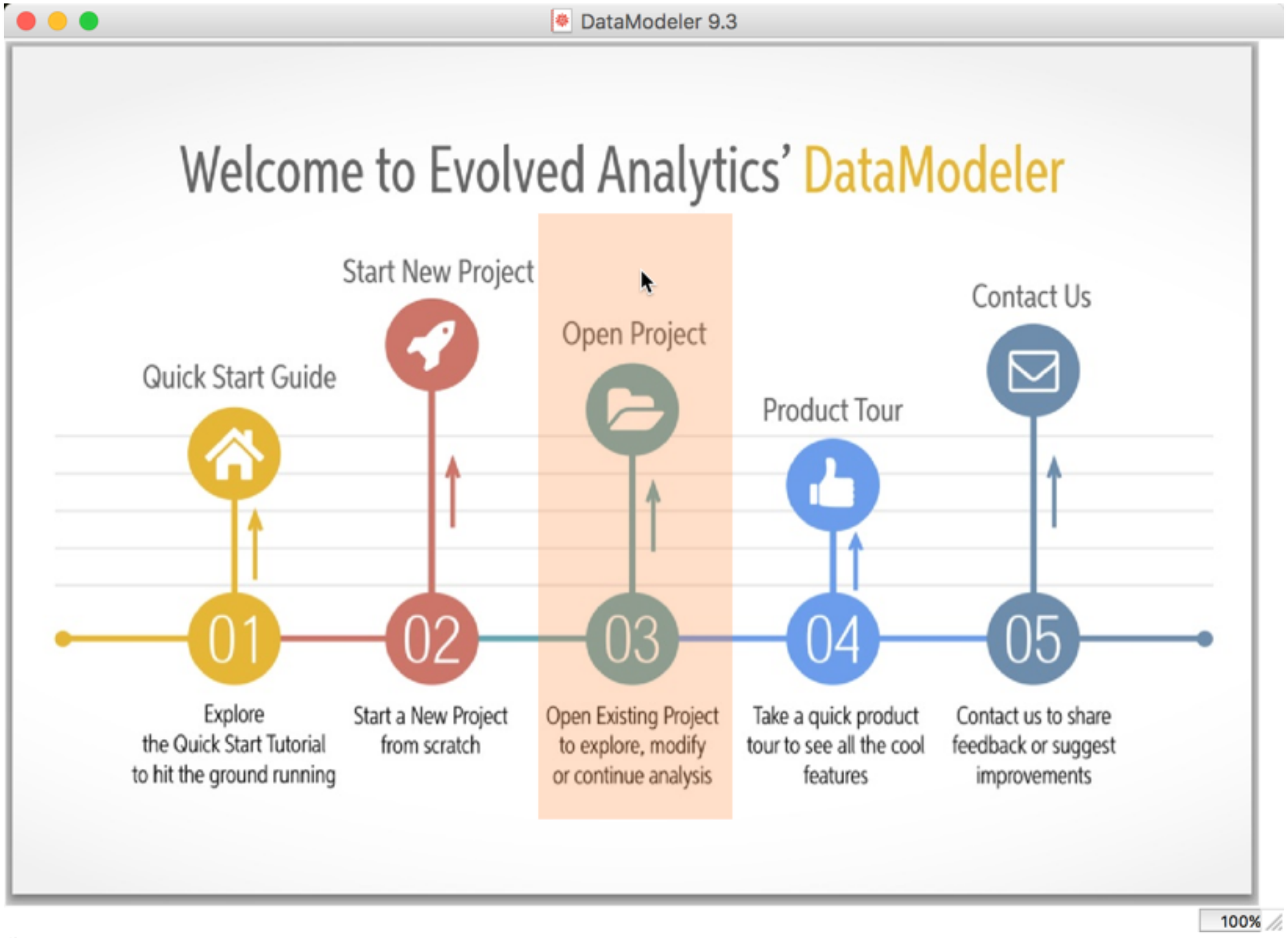


DataModeler Design

- User can choose between a GUI or the classic Mathematica notebook interface
- The functions used by the GUI are exposed for power users and custom workflows
- ~450 functions are exposed and (well)

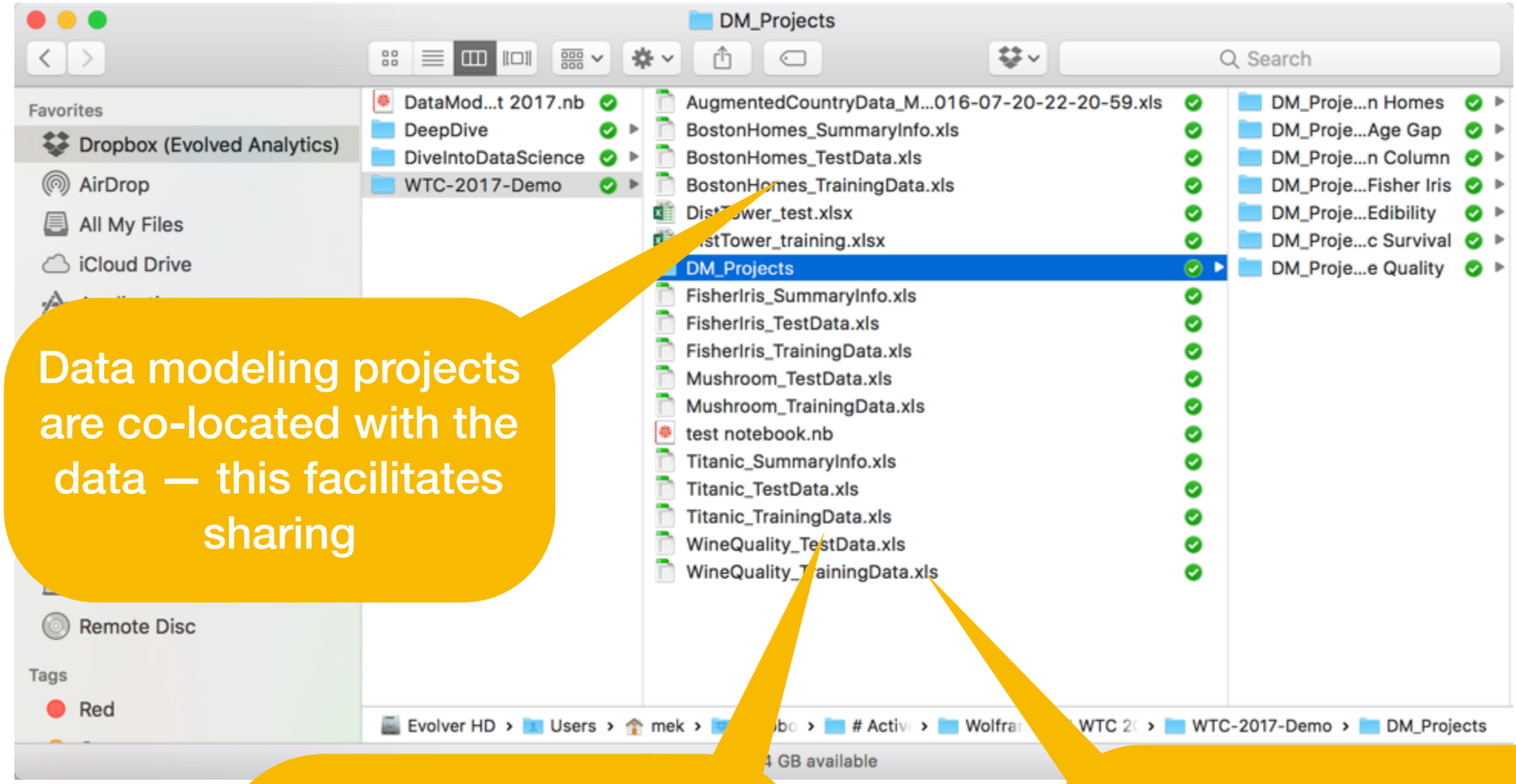


GUI used for Demo

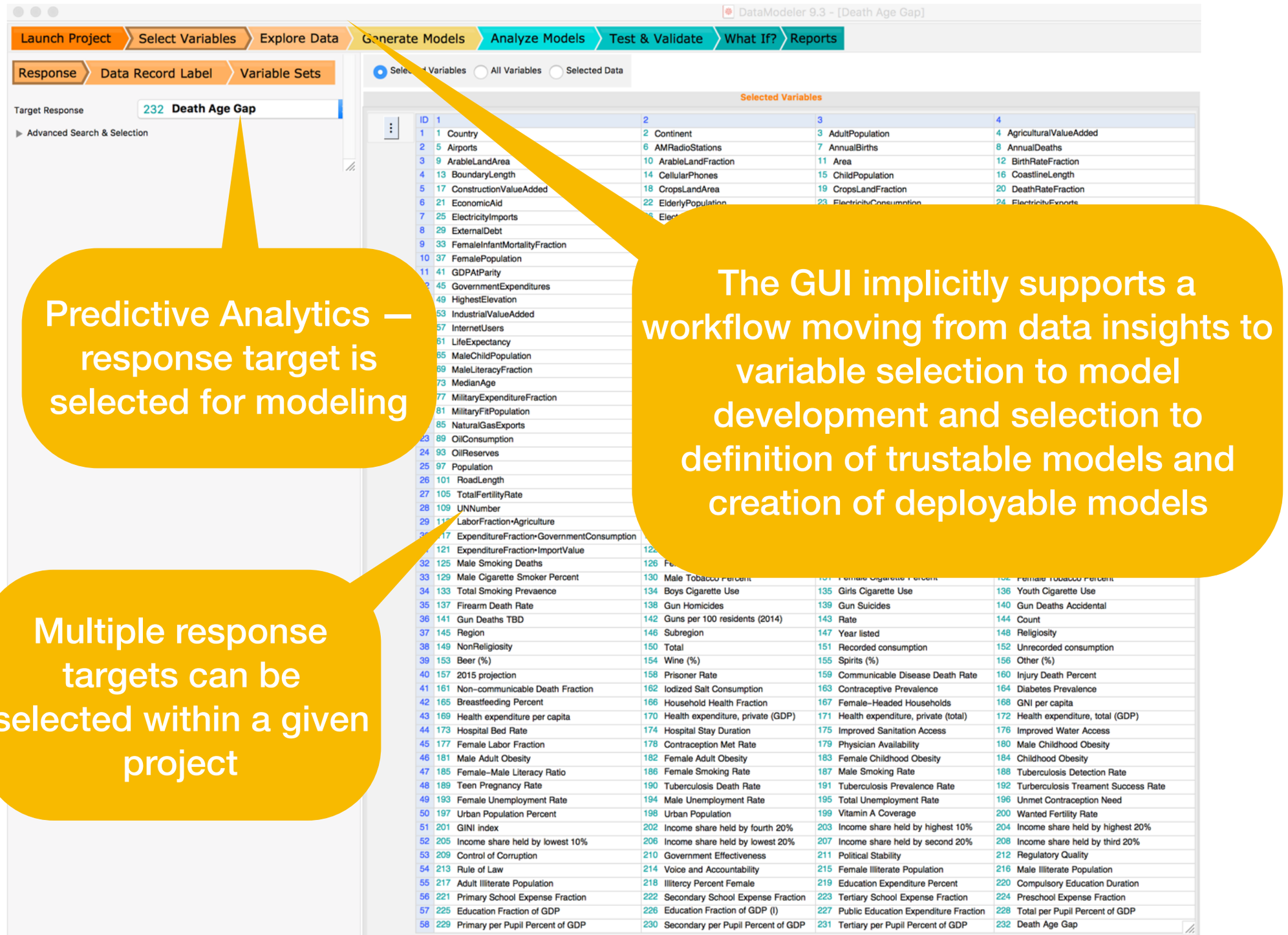


www.evolved-analytics.com

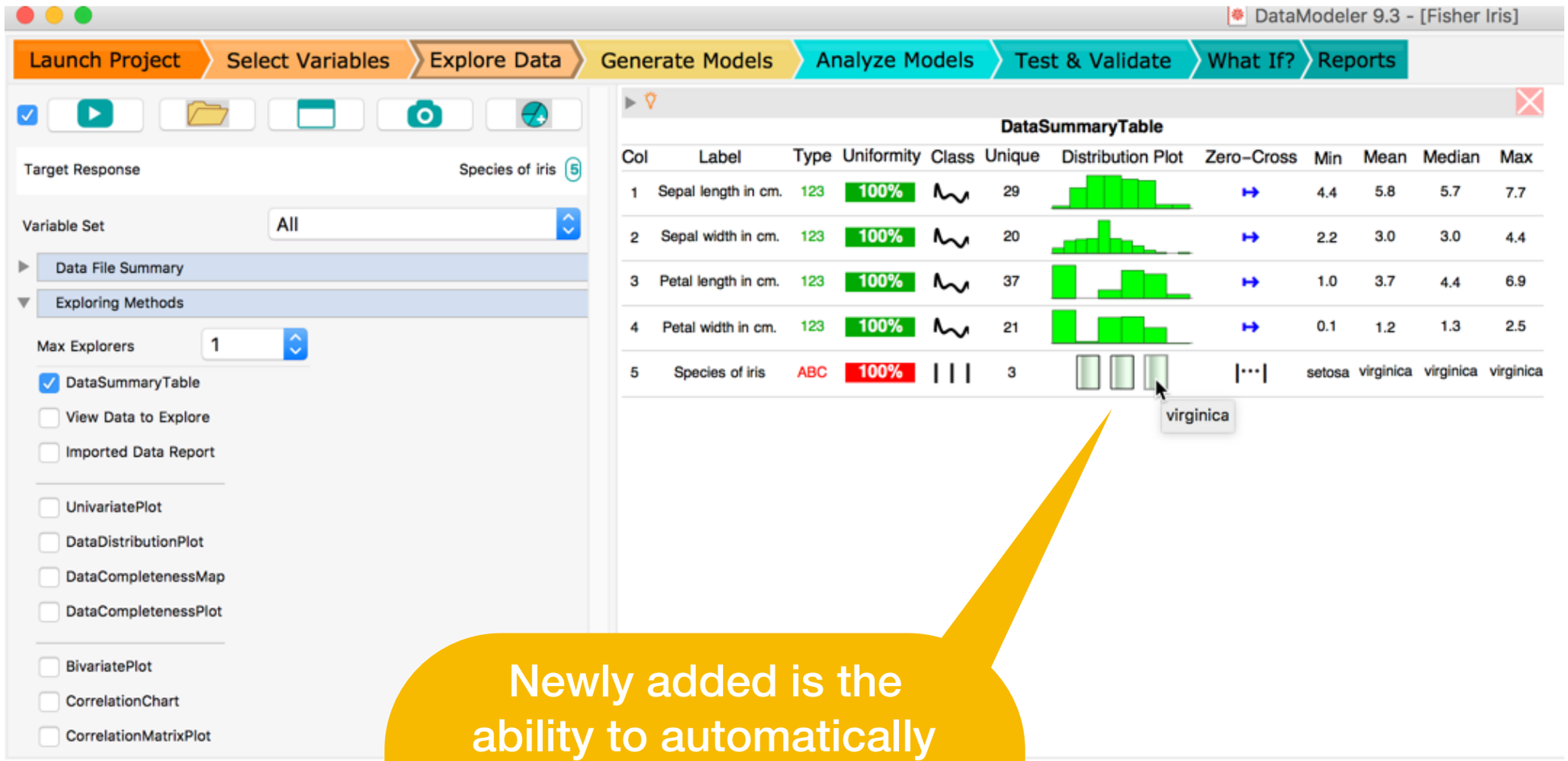
Project co-located with data



Choose a Target



Analysis now supports low-order categoricals

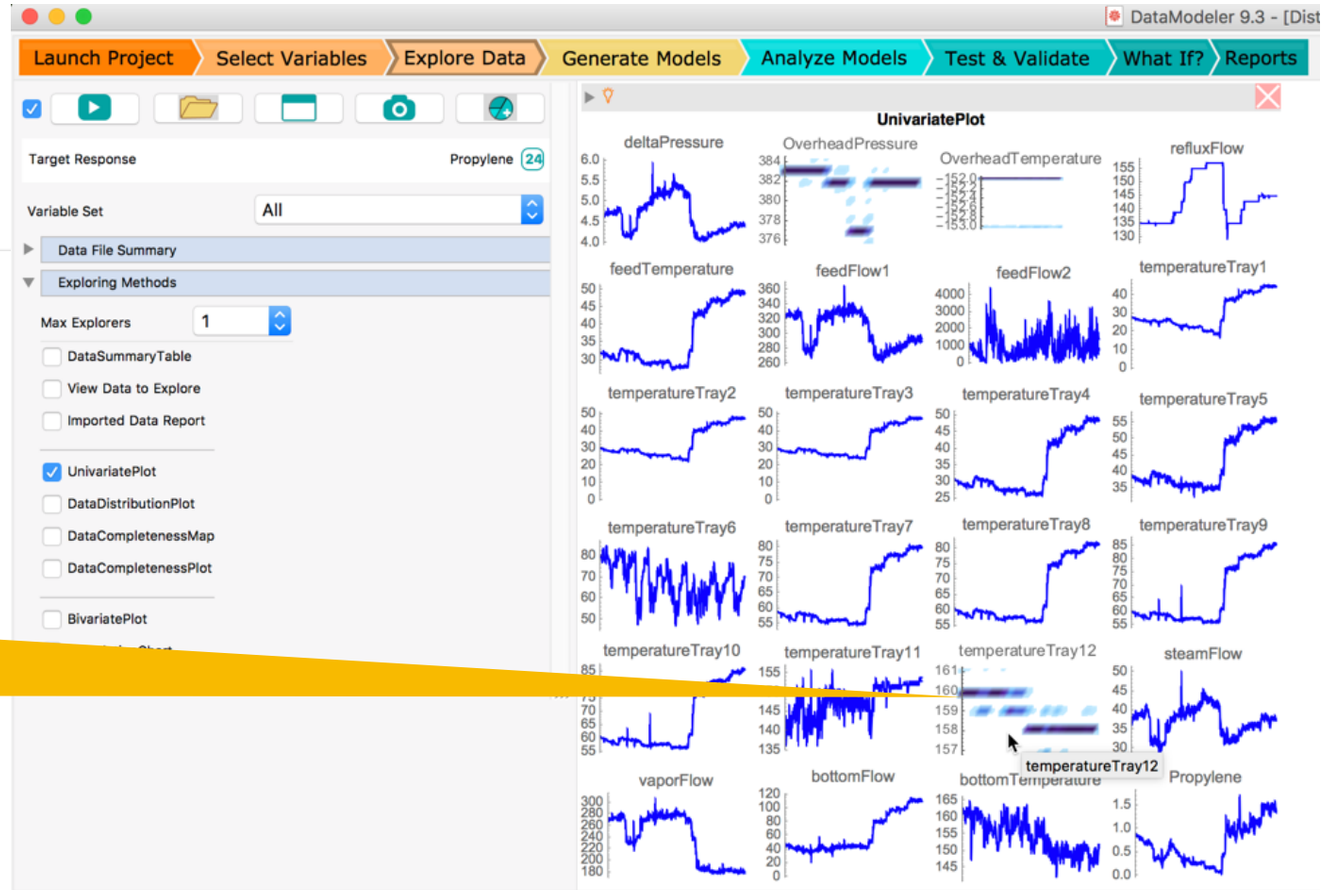


Newly added is the ability to automatically convert low-order categoricals to numerics

Getting the Zen of the Data

Plots adapt if a low-order categorical is detected and tooltips generated to show the mapping being used.

If a variable is detected as numeric but quantized, the displays shift appropriately.



DataModeler 9.3 - [Death Age Gap]

Launch Project | Select Variables | Explore Data | Generate Models | Analyze Models | Test & Validate | What If? | Reports

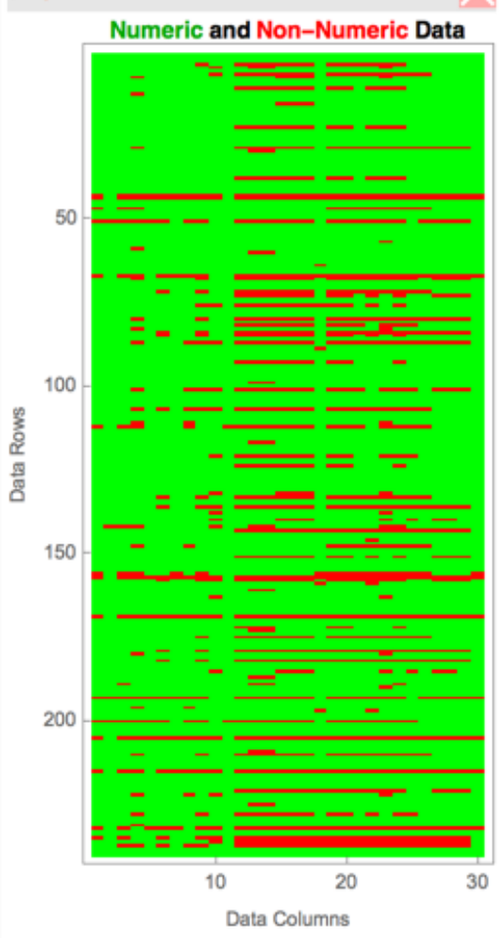
Target Response: Death Age Gap 232

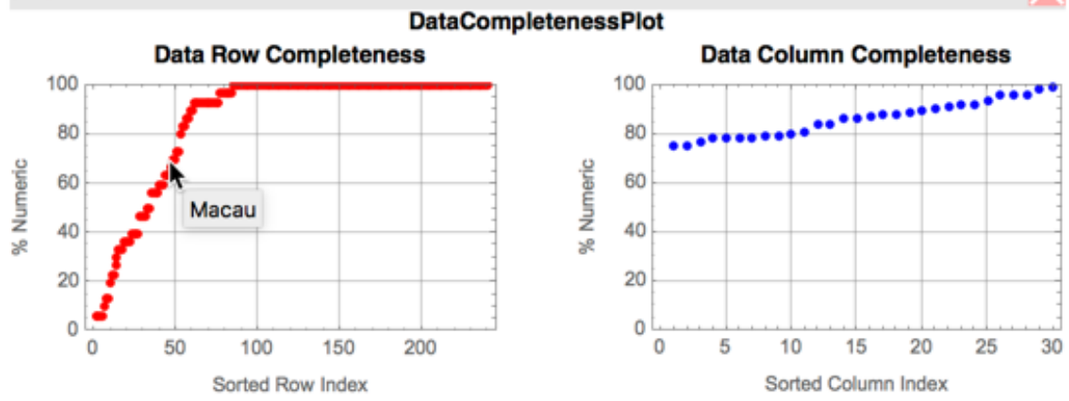
Variable Set: Round 2 Drivers (1%)

- Data File Summary
- Exploring Methods

Max Explorers: 3

- [x] DataSummaryTable
- [ ] View Data to Explore
- [ ] Imported Data Report
- [ ] UnivariatePlot
- [ ] DataDistributionPlot
- [x] DataCompletenessMap
- [x] DataCompletenessPlot
- [ ] BivariatePlot
- [ ] CorrelationChart
- [ ] CorrelationMatrixPlot





The default rule for model development is that a model must be supported by having at least 75% of the data records be completely numeric in the variables used by the model. Incomplete records are not considered in the quality assessment.

Missing data is handled seamlessly for both modeling and data exploration and model development

Page 1 | Page 2

DataSummaryTable

Col	Label	Type	Uniformity	Class	Unique	Distribution Plot	Zero-Cross	Min	Mean	Median	Max
20	DeathRateFraction	123	94%		202		↦	0.0	0.0	0.0	0.0
21	EconomicAid	123	98%		213		↔	-2.4×10^{10}	3.6×10^{7}	8.7×10^{7}	2.2×10^{10}
33	FemaleInfantMortalityFraction	123	92%		214		↦	0.0	0.0	0.0	0.2
35	FemaleLiteracyFraction	123	86%		142		↦	0.1	0.8	0.9	1.0
42	GDPPerCapita	123	96%		231		↦	137.6	15952.0	5514.7	211500.0
43	GDPRealGrowth	123	91%		218		↔	-0.1	0.0	0.0	0.2
61	LifeExpectancy	123	96%		228		↦	45.6	71.5	74.0	84.4
62	LiteracyFraction	123	92%		144		↦	0.2	0.9	0.9	1.0

DataSummaryTable

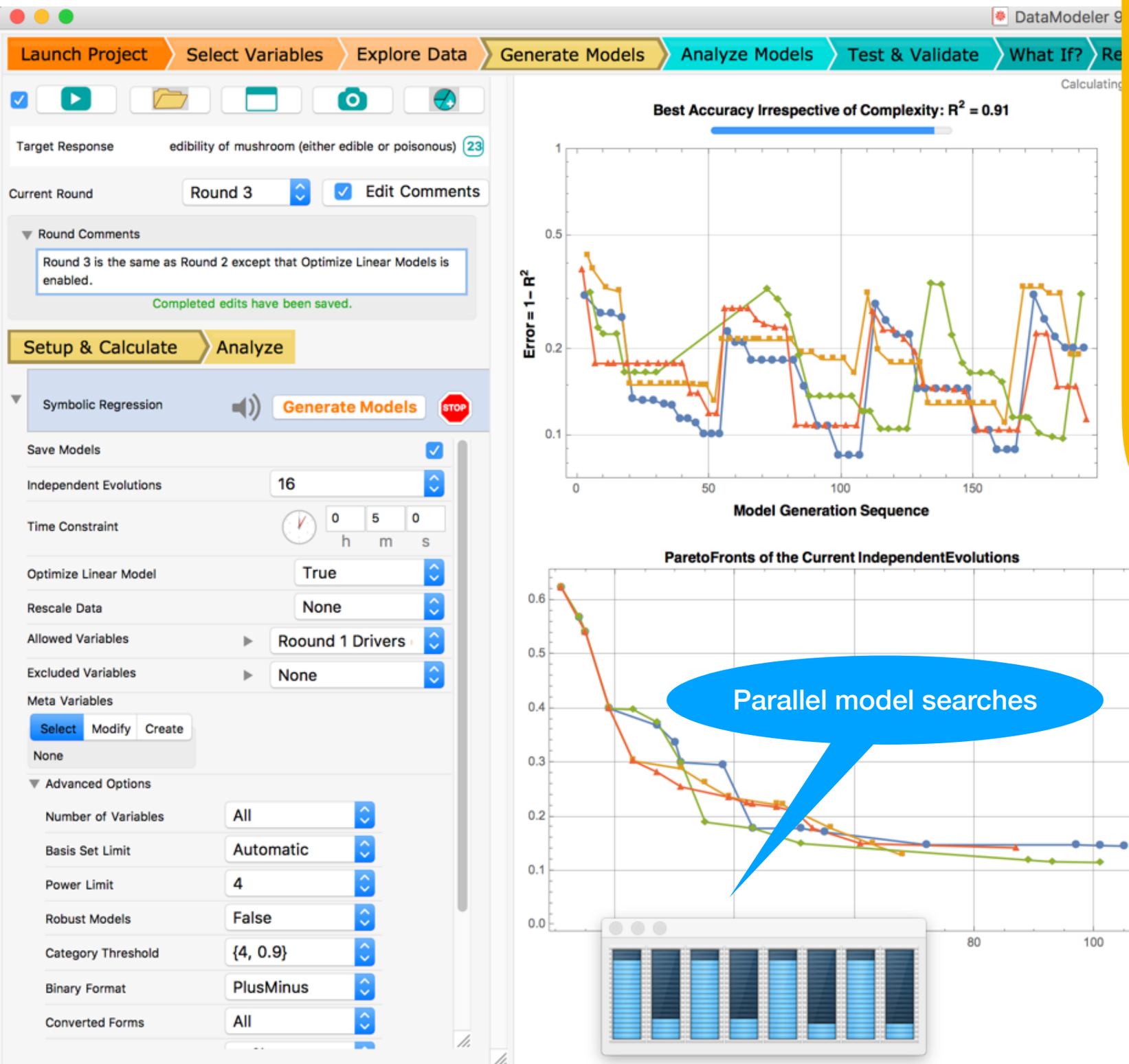
Col	Label	Type	Uniformity	Class	Unique	Distribution Plot	Zero-Cross	Min	Mean	Median	Max
1	cap-shape	ABC	100%	| | |	6		|···|	bell	convex	convex	sunken
2	cap-surface	ABC	100%	| | |	4		|···|	fibrous	scaly		
3	cap-color	ABC	100%	⎍	10		|···|	brown	brow		
4	bruises?	ABC	100%	=	2		| |	False	Fals		
5	odor	ABC	100%	⎍	9		|···|	almond	non		
6	gill-attachment	ABC	100%	—	2		| |	attached	free		
7	gill-spacing	ABC	100%	=	2		| |	close	clos		
8	gill-size	ABC	100%	=	2		| |	broad	broa		
9	gill-color	ABC	100%	⎍	12		|···|	black	buf		
10	stalk-shape	ABC	100%	=	2		| |	enlarging	taper		
11	stalk-root	ABC	100%	| | |	5		|···|		bulbo		
12	stalk-surface-above-ring	ABC	100%	| | |	4		|···|	fibrous	smoc		
13	stalk-surface-below-ring	ABC	100%	| | |	4		|···|	fibrous	smoc		
14	stalk-color-above-ring	ABC	100%	⎍	9	fibrous	|···|	brown	whit		
15	stalk-color-below-ring	ABC	100%	⎍	9		|···|	brown	whit		
16	veil-type	ABC	100%	—	1		•	partial	parti		
17	veil-color	ABC	100%	—	4		|···|	brown	whit		
18	ring-number	ABC	100%	=	3		|···|	none	one		
19	ring-type	ABC	100%	| | |	5		|···|	evanescent	penda		
20	spore-print-color	ABC	100%	| | |	9		|···|	black	white		
21	population	ABC	100%	⎍	6		|···|	abundant	several	several	solitary
22	habitat	ABC	100%	⎍	7		|···|	grasses	woods	woods	woods
23	edibility of mushroom (either edible or poisonous)	ABC	100%	=	2		| |	edible	edible	edible	poisonous

Categorical Analysis

- A purely textual data set – is the mushroom edible or poisonous?
- Goal is to identify key attributes & develop predictive models
- We cap the number of categories converted with those not converted ignored
- Still doing symbolic regression to produce a predictive model

Model Search

- The assumption is that not all variables are equal and that we can consolidate to a relative handful of inputs
- Models are evolved rewarding those which are simple and accurate with breeding rights
- Developed models are simple algebraic expressions — human interpretable!
- We can view the search as an automated hypothesis generation & refinement
- Lots of diverse accurate-but-simple models are developed from which we can extract insight.



DataModeler 9.3 - [Mushroom Edibility]

Launch Project | Select Variables | Explore Data | Generate Models | Analyze Models | Test & Validate | What If? | Reports

Target Response: edibility of mushroom (either edible or poisonous) 23

Models | Analyze

Select Models 307/2813 — Reset Filters

Candidates | Focus | Ensembles

Quality Box

All

Complexity 100

$1-R^2$ 0.2

Selection Fraction 50%

Number of Variables: All (Max)

Required Variables: None

Allowed Variables: All

Excluded Variables: None

Power Limit: 4

Robust Models: False

ANOVA Trim: Apply

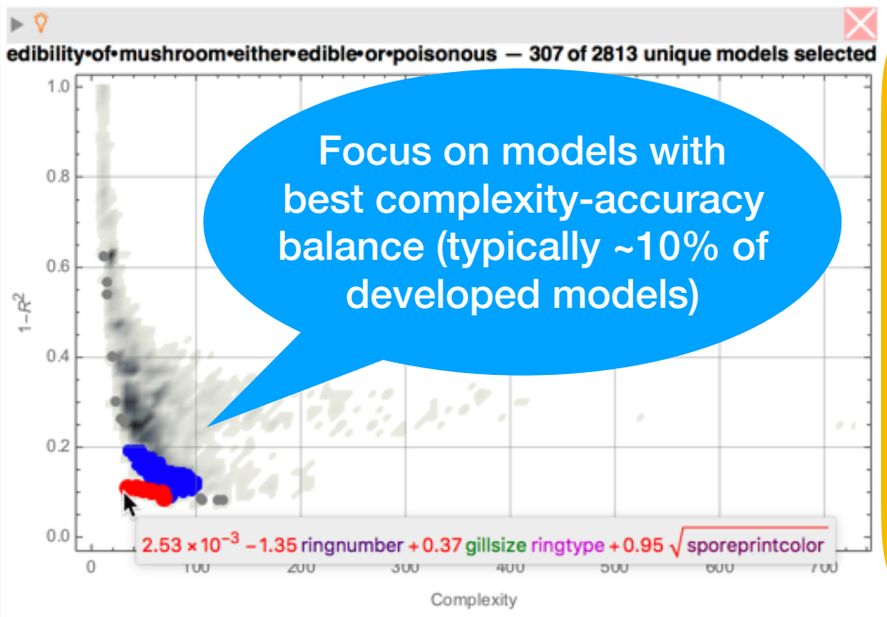
Selected Model(s): Save Model Set...

Model Ensemble: Create...

Explore Models

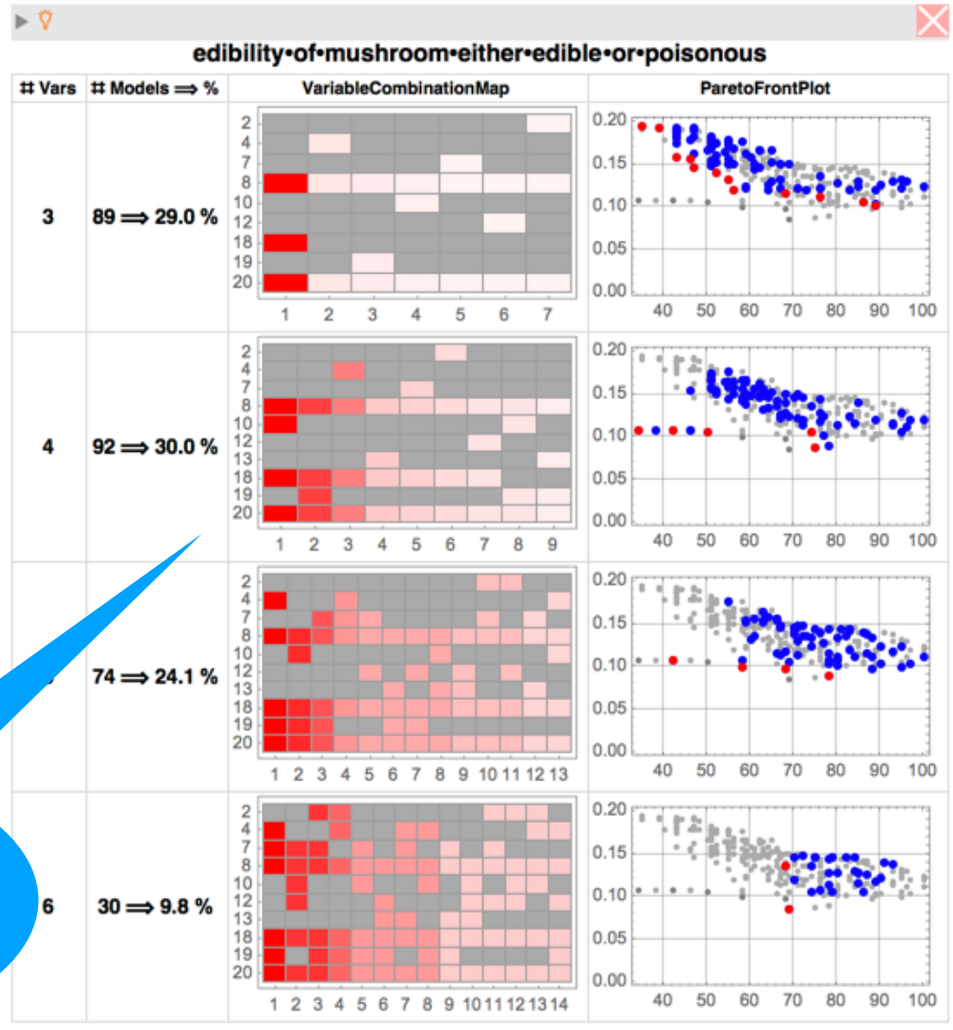
Max Explorers: 3

- Model Quality Plots (1)
  - ParetoFrontPlot
  - ParetoFrontContextPlot
  - ParetoFrontLogPlot
  - ParetoFrontContextLogPlot
- Model Dimensionality (1)
  - ModelDimensionalityTable
- Variable Presence (1)
  - VariablePresenceChart*
  - VariablePresenceTable
  - VariablePresenceDist
  - VariablePresence
- Meta Variables
- Variable Combinations
- Model Selection Tables
- Variable Filters Table



Focus on models with best complexity-accuracy balance (typically ~10% of developed models)

- Model search converting categoricals to numerics (max 4 categories covering at least 90% of records)
- 16 model searches of 5 minutes on quad-core laptop — 20 minutes of clock time



There are multiple variable combinations giving comparable performance

Variables have different model inclusion — an indicator of importance

DataModeler 9.3 - [Mushroom Edibility]

Launch Project | Select Variables | Explore Data | Generate Models | Analyze Models | Test & Validate | What If? | Reports

Target Response: edibility of mushroom (either edible or poisonous) 23

Models | Analyze

Select Models 307/2813 — Reset Filters

Candidates | Focus | Ensembles

Quality Box

All

Complexity 100

$1-R^2$ 0.2

Selection Fraction 50%

Number of Variables: All (Max)

Required Variables: None

Allowed Variables: All

Excluded Variables: None

Power Limit: 4

Robust Models: False

ANOVA Trim: Apply

Selected Model(s): Save Model Set...

Model Ensemble: Create...

Explore Mode

Max Explor

Model C

Model Di

Variable Pres

VariablePresence

VariablePresenceTable

VariablePresenceDistributionChart

VariablePresenceMap

Meta Variables (1)

ModelBasisSetTable*

ModelBasisSetDistributionTable*

ModelBasisSetDistributionChart

MetaVariableTable*

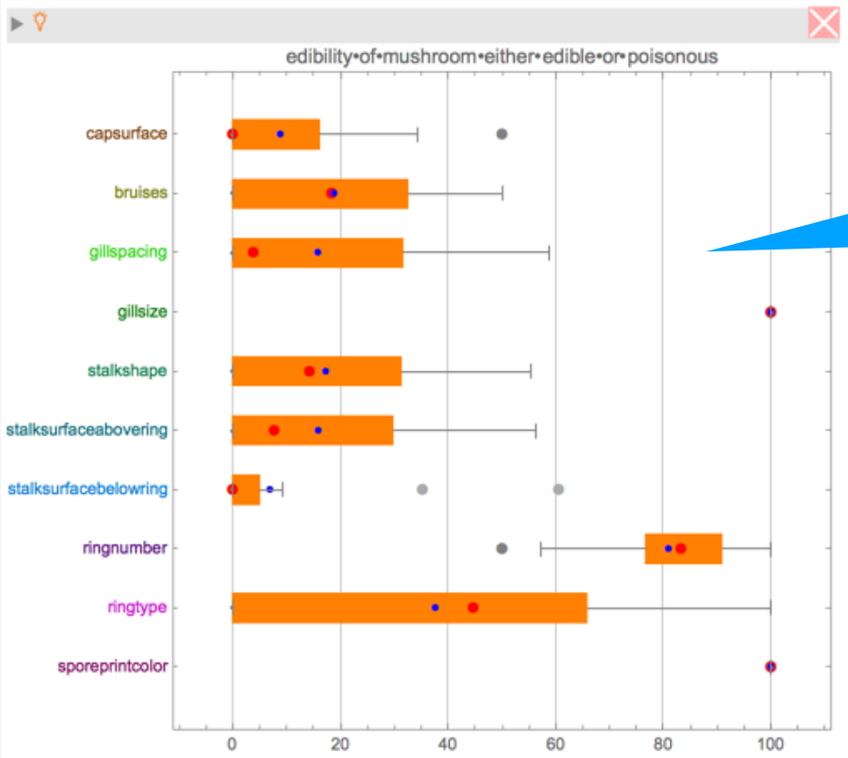
MetaVariableDistributionTable*

MetaVariableDistributionChart

Variable Combinations (1)

VariableCombinationTable*

VariableCombinationMap

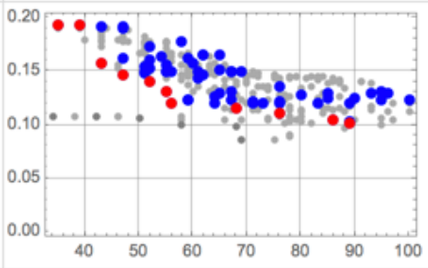
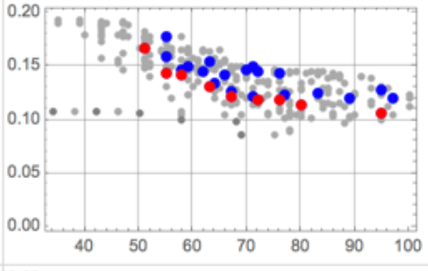
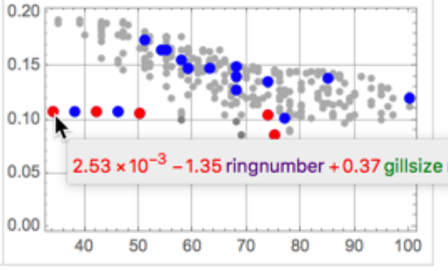


Here we look at the variable presence in each of the independent model searches. Each pursues their own path to success

We can also look for model substructures which have been exploited. These metavariables can provide insight as well as be exploited in subsequent rounds of modeling

Here we look at the most prevalent variable combinations and their performance

edibility•of•mushroom•either•edible•or•poisonous

	num ⟹ %	Variables Used	ParetoFrontPlot
1	64 ⟹ 20.8 %	gillsize ringnumber sporeprintcolor	
2	28 ⟹ 9.1 %	gillsize stalkshape ringnumber sporeprintcolor	
3	21 ⟹ 6.8 %	gillsize ringnumber ringtype sporeprintcolor	

$2.53 \times 10^{-3} - 1.35$ ringnumber $+ 0.37$ gillsize ringtype $+ 0.95 \sqrt{\text{sporeprintcolor}}$

edibility•of•mushroom•either•edible•or•poisonous

	Rank	Count	MetaVariable	% of models	% of MetaVariables
++	1	53	gillsize + $\sqrt{\text{sporeprintcolor}}$	17.3	35.1
++	2	49	$\sqrt{\text{gillsize} + \sqrt{\text{sporeprintcolor}}}$	16.0	32.5
++	3	28	gillsize + ringnumber + $\sqrt{\text{sporeprintcolor}}$	9.1	18.5
++	4	22	gillsize^2 stalksurfacebelowring	7.2	14.6
++	5	22	$(\text{gillsize} + \text{ringnumber} + \sqrt{\text{sporeprintcolor}})^{1/3}$	7.2	14.6
−−	6	21	gillsize ringtype	6.8	13.9
++	7	14	$\frac{1}{-1.03928+\text{sporeprintcolor}}$	4.6	9.3
++	8	14	$\text{gillsize}^2 + \sqrt{\text{sporeprintcolor}}$	4.6	9.3
++	9	13	$(\text{gillsize}^2 + \sqrt{\text{sporeprintcolor}})^{1/3}$	4.2	8.6
++	10	12	$\sqrt{\text{sporeprintcolor}^3}$	3.9	7.9

DataModeler 9.3 - [Mushroom Edibility]

Launch Project | Select Variables | Explore Data | Generate Models | Analyze Models | Test & Validate | What If? | Reports

Target Response: edibility of mushroom (either edible or poisonous) 23

Models | Analyze

Select Models 301/1937 — Reset Filters

Candidates | Focus | Ensembles

Quality Box

All

Complexity 80

$1-R^2$ 0.15

Selection Fraction 50%

Number of Variables: All (Max)

Required Variables: None

Allowed Variables: All

Excluded Variables: None

Power Limit: 4

Robust Models: False

ANOVA Trim: Apply

Selected Model(s): Save Model Set...

Model Ensemble: Create...

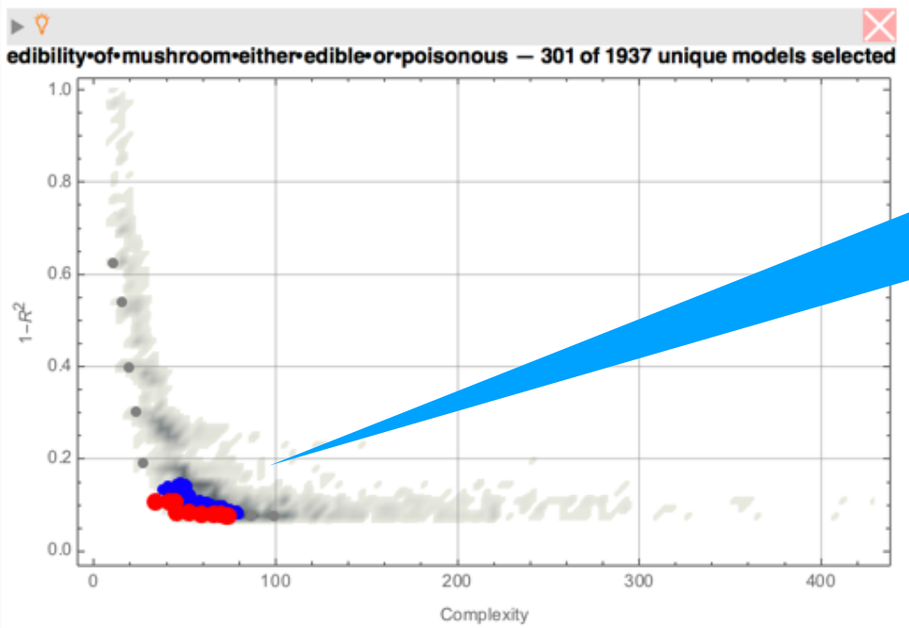
Explore Models

Max Explorers: 3

Model Quality Plots (2)
- ParetoFrontPlot
- ParetoFrontContextPlot
- ParetoFrontLogPlot
- ParetoFrontContextLogPlot

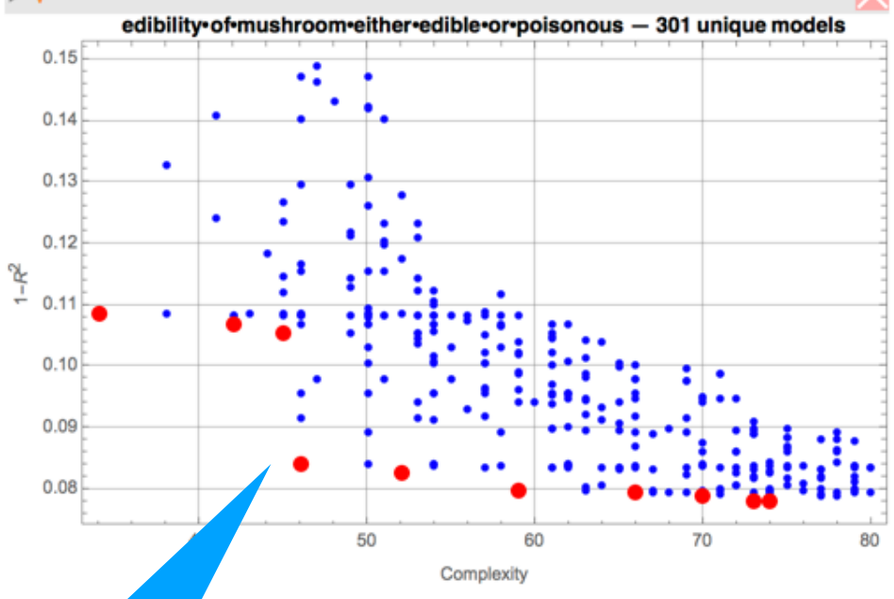
Model Dimensionality
- ModelDimensionalityTable

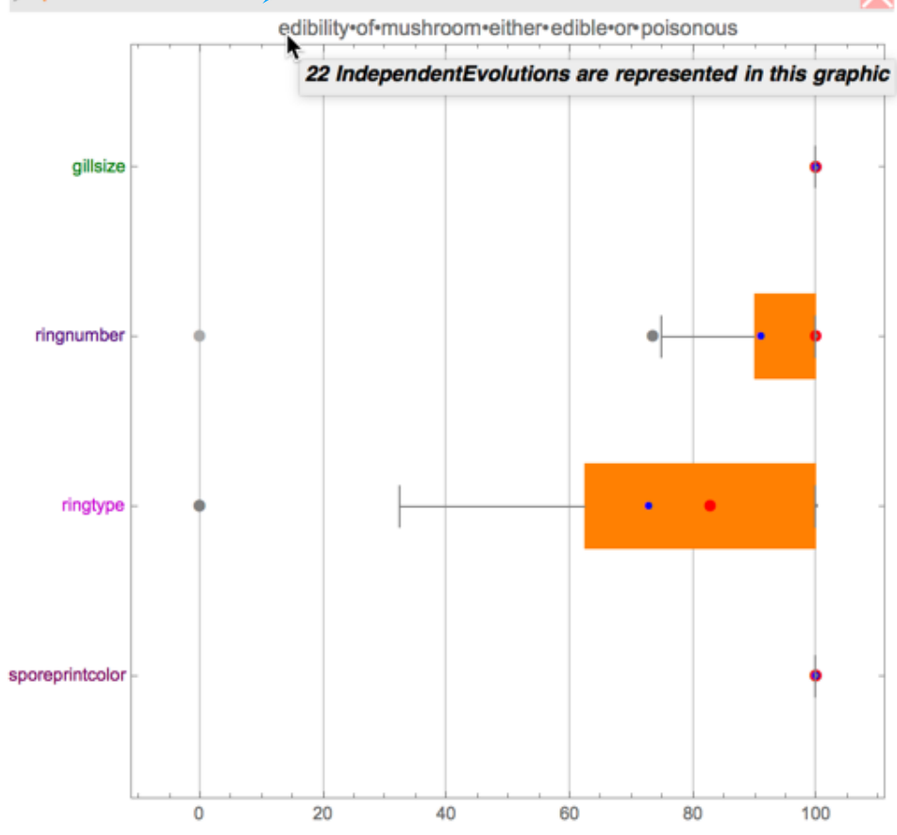
Variable Presence (1)
- VariablePresenceChart*
- VariablePresenceTable
- VariablePresenceDistributionChart



We did additional model searches only allowing four variables and selected 301 models as being both simple and accurate

110 minutes of model search across four rounds of analysis created this model set





The data defines the models and there are many models which are of comparable quality

DataModeler 9.3 - [Mushroom Edibility]

Target Response: edibility of mushroom (either edible or poisonous) 23

Models | Analyze

Select Models 1/301

Candidates | Focus | Ensembles

Model Picker

Approx. Max Number of Models: 25

Approx. Min Number of Models: 1

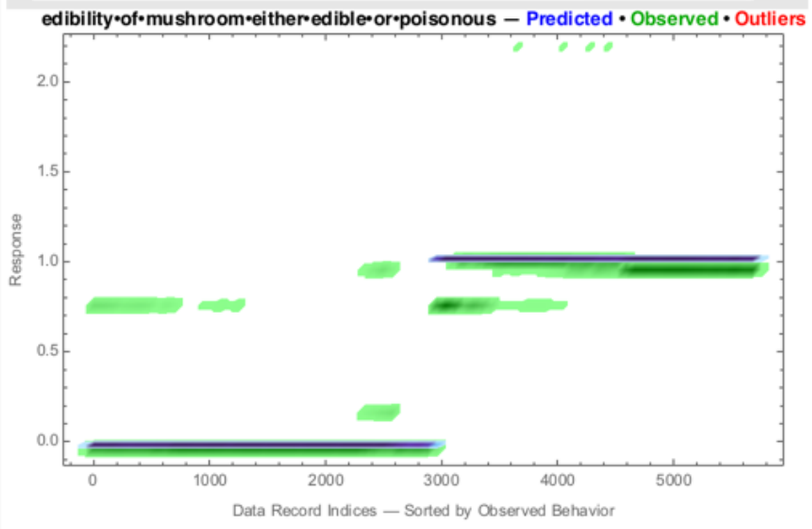
Selected Model(s) — Save Model Set...

Explore Models

Max Explorers: 3

- Model Statistics
- Prediction Plots (3)
  - ModelPredictionComparisonPlot
  - ModelPredictionPlot
  - ModelResidualPlot
  - ResponsePlotExplorer
  - ResponseSurfacePlotExplorer
  - ResponseContourPlotExplorer
- Data Outliers
- Meta Variables
  - ModelBasisSetTable*
  - MetaVariableTable*
- Model Selection Tables
- Variable Filters Table

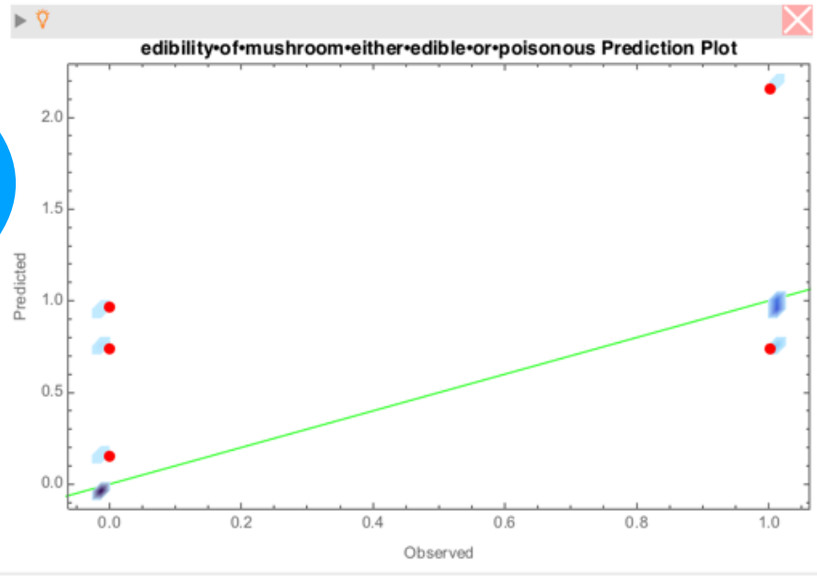
Plot Markers	None
Sort By	Observed
Connect Plot Points	False
Data Record Labels	Automatic
Data Outliers	Outliers
Front	Automatic
Tooltip Limit	100
Outlier Tooltip Limit	100
Grid Lines	False
Image Size	500
Aspect Ratio	3/5 (Landscape)
Plot Range	All
Plot Label	Model Prediction Comparison
Reset	<<<

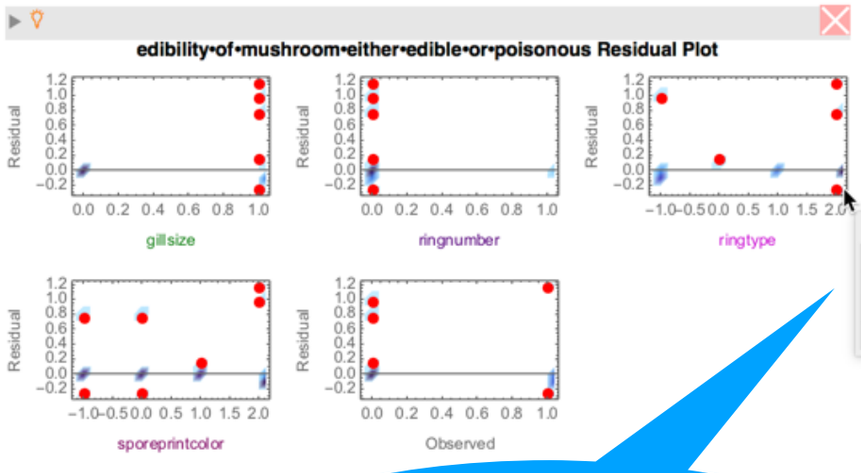


Each graphic has controls which can be exposed and used to tweak the display

- Here we are looking at one of the candidate models
- Unfortunately, THE model does not typically exist in data-derived models
- We can use ensembles of models to create a trustable model to detect extrapolation or underlying system changes

Automatic shifts to a DataDensityPlot reveals that the model is doing a good job with a few erroneous predictions





Tooltips are used to show the variable mapping being used

Test & Validate

DataModeler 9.3 - [Distillation Colum

Launch Project | Select Variables | Explore Data | Generate Models | Analyze Models | Test & Validate | What If? | Reports

Target Response: Propylene 24

Data Sets | Models | Analyze

Pick Models

Model Sets

propylene (190)

Individual Models

No single-model sets retrieved.

Model Ensembles

No ensembles retrieved.

Select Target Data Set

Data Sets for Model Analysis: DistTower_training, DistTower_test

Model Quality

Quality Metrics: Updated

Updated Modeling Objective

Select Models 190/190

Candidates

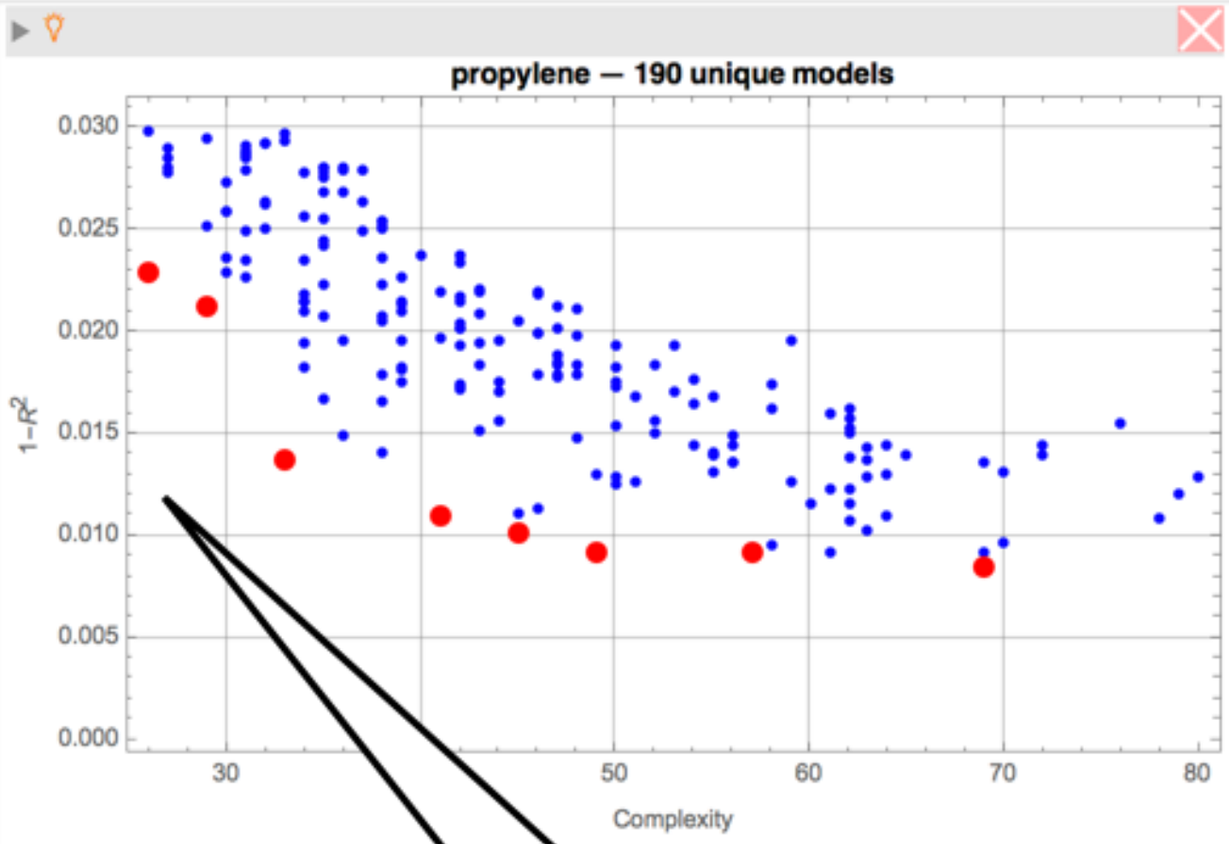
Quality Box

Complexity 81

$1-R^2$ 0.03

Selection Fraction 100%

Number of All



Original Metrics

Complexity & NoisePower

Complexity & 2-Norm

NoisePower

2-Norm

R2 is not a good metric for evaluating against multiple data sets since it assesses shape match rather than true prediction error. NoisePower is the norm of the prediction error normalized by the norm of the observed values. Here we reassess the models quality against the specified data sets.

DataModeler 9.3 - [Distillation Column]

Launch Project | Select Variables | Explore Data | Generate Models | Analyze Models | Test & Validate | What If? | Reports

Target Response: Propylene 24

Data Sets | Models | Analyze

Pick Models

Model Sets
- propylene (190)

Individual Models
No single-model sets retrieved.

Model Ensembles
No ensembles retrieved.

Select Target Data Set

Data Sets for Model Analysis: DistTower_training, DistTower_test

Model Quality

Quality Metrics: Updated

Updated Modeling Objective: Complexity & NoisePower

Select Models 137/190 — Reset Filters

Candidates | Focus | Ensembles

Quality Box: All / Complexity 81 / DistTower_training:Noise Power 0.0082 / DistTower_test:Noise Power 0.05

Selection Fraction: 100%

Number of Variables: All (Max)

Required Variables: None

Allowed Variables: All

Excluded Variables: None

Power Limit: 6

Robust Models: False

Selected Model(s): Save Model Set...

Model Ensemble: Create...

Explore Models

Max Explorers: 1

- Model Quality Plots (1)
- Model Dimensionality
- Variable Presence

Density Plot: Automatic

Displayed Objectives: Complexity; DistTower_training:Noise Power; DistTower_test:Noise Power

Graphics Array Columns: Automatic

Image Size: 925.

Aspect Ratio: 3/5 (Landscape)

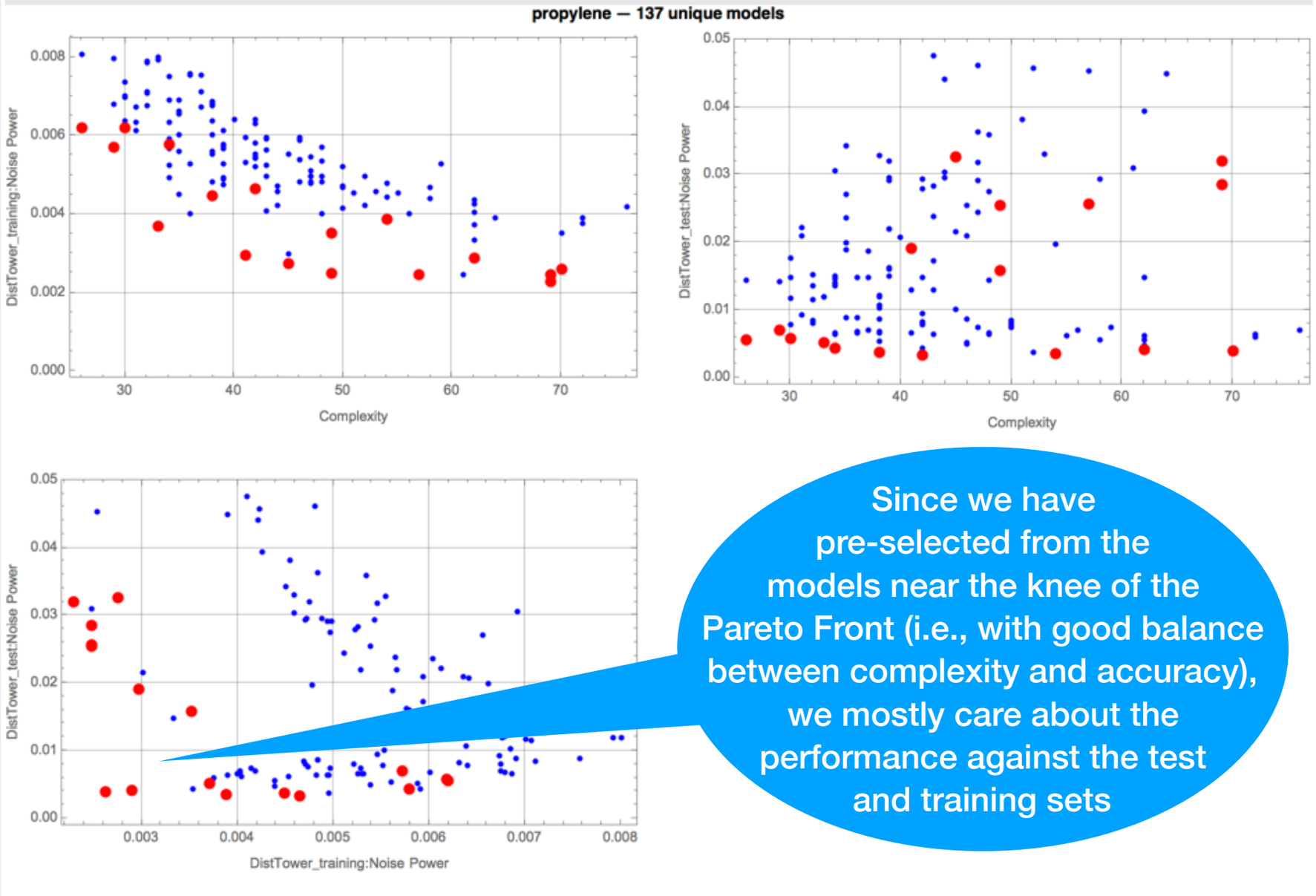
Plot Label: Pareto Front Plot

Plot Range: All

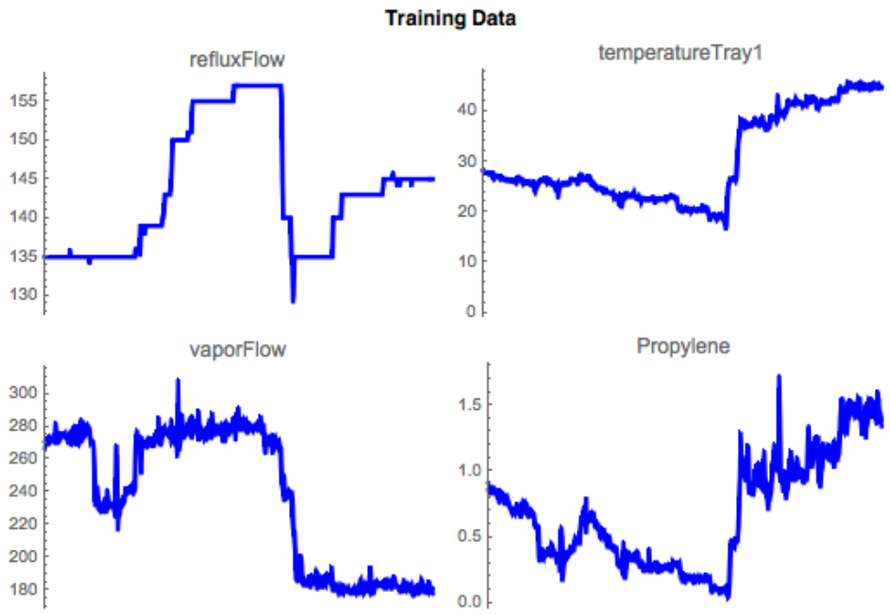
Tooltip Limit: 750

Reset: <<<

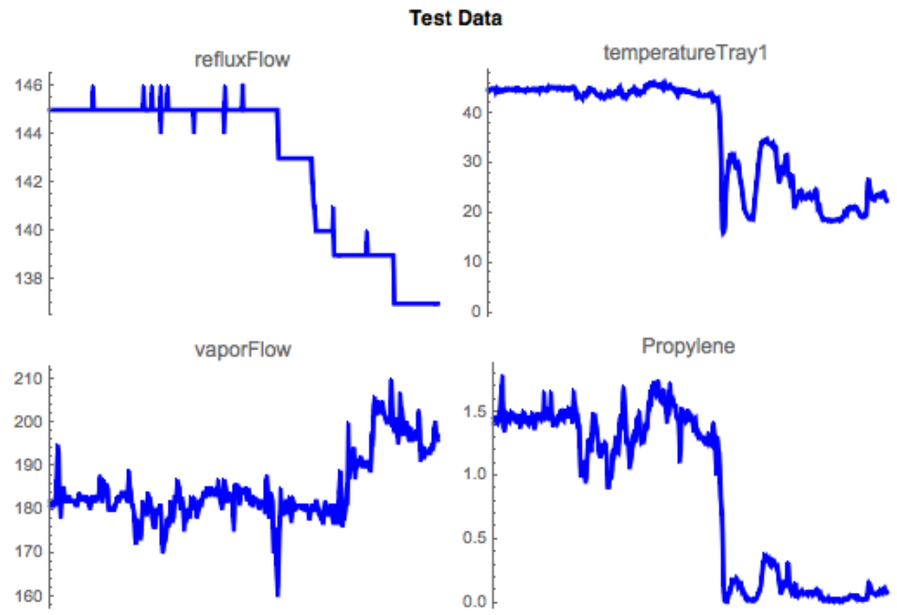
We can choose which quality metrics we want to display via the (normally hidden) controls



Since we have pre-selected from the models near the knee of the Pareto Front (i.e., with good balance between complexity and accuracy), we mostly care about the performance against the test and training sets

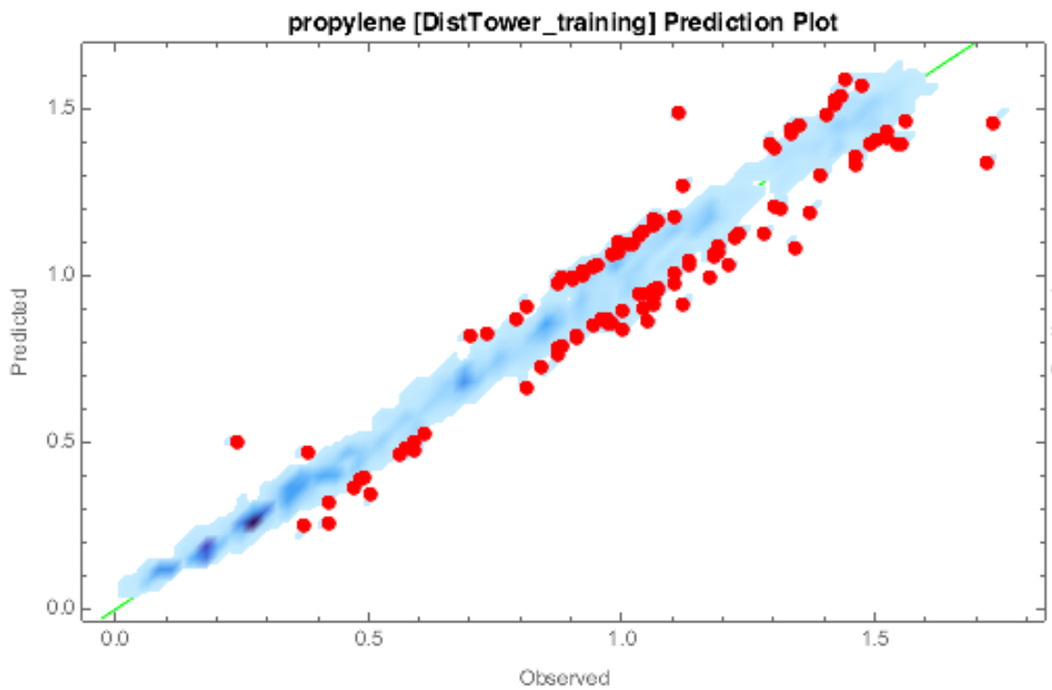


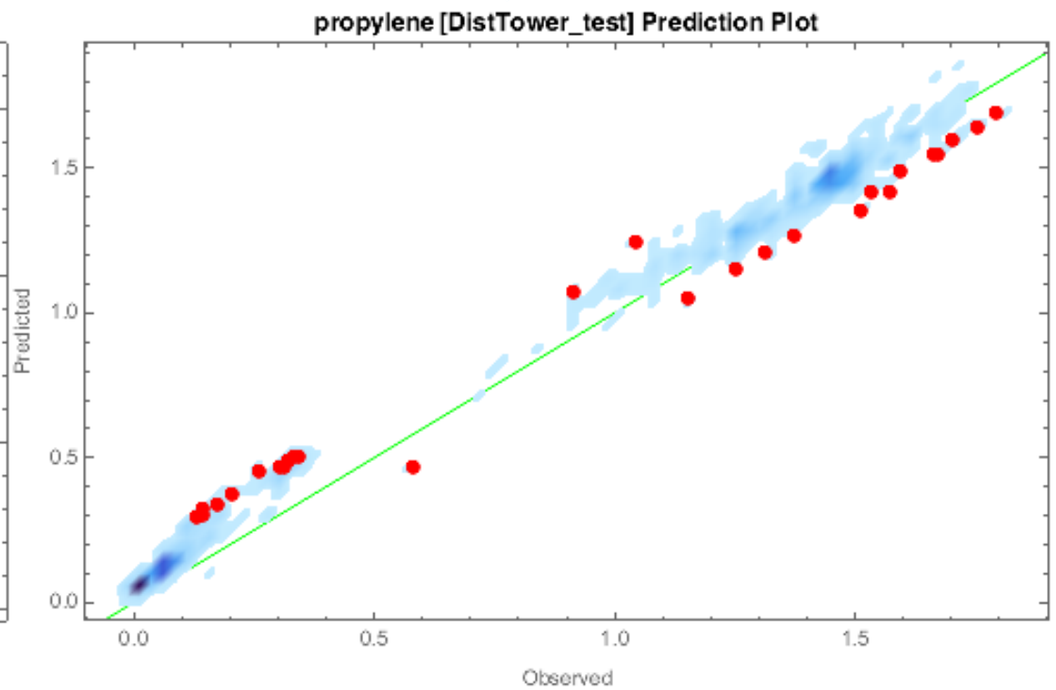
The normal purpose of a test set is to avoid over-fitting. With DataModeler, we avoid that by explicitly considering the complexity-accuracy trade-off — so we can develop fully-informed models.



This data set was intentionally partitioned to test model extrapolation — such is a reality of deployed predictive analytics models.

Here we found a model that extrapolates well; however, it is better to build a fully-informed model from the start.





Round 3 Ensemble (3 Vars) Summary Table

MedianAverage[models, 0.7]

Expression

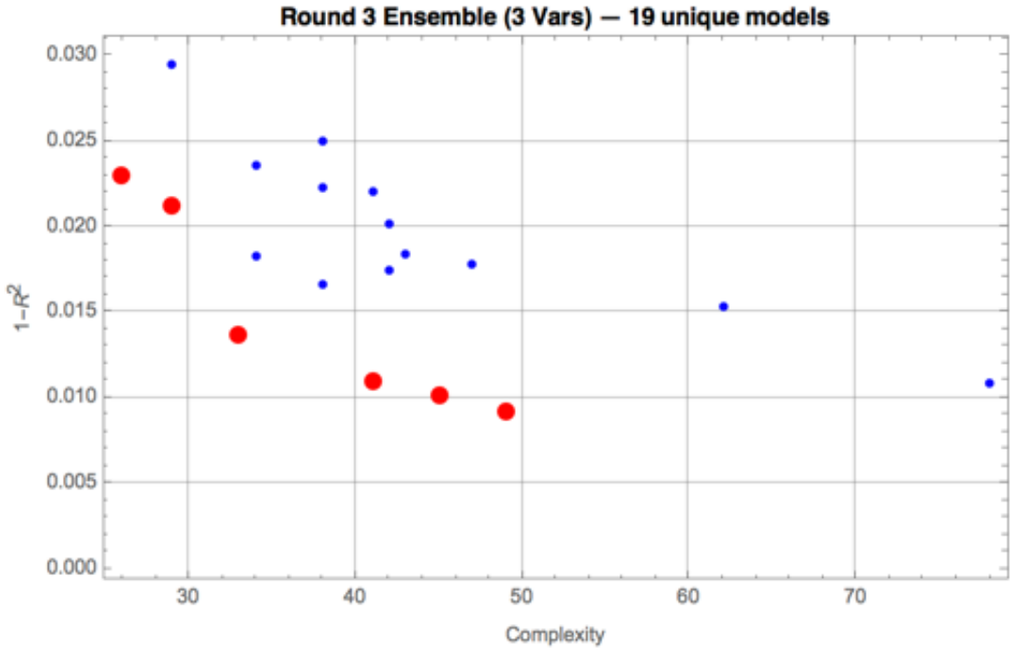
$-0.28-\left(3.55\times10^{-7}\right)(127.67-\text{vaporFlow})\,\text{temperatureTray}_1{}^3$
$-0.13+\dfrac{\left(7.59\times10^{-8}\right)\text{vaporFlow}^2\,\text{temperatureTray}_1{}^3}{\text{refluxFlow}}$
$-0.48+\left(3.58\times10^{-8}\right)\text{vaporFlow}^2\,\text{temperatureTray}_1\,(-9.37+\text{temperatureTray}_1)$
$-0.27-\dfrac{\left(5.00\times10^{-5}\right)(127.67-\text{vaporFlow})\,\text{temperatureTray}_1{}^3}{\text{refluxFlow}}$
$1.07-\left(6.86\times10^{-3}\right)\text{refluxFlow}-\left(4.49\times10^{-2}\right)\text{temperatureTray}_1+\left(9.18\times10^{-6}\right)\text{vaporFlow}\,\text{temperatureTray}_1{}^2$
$-0.41+\left(7.69\times10^{-3}\right)\text{vaporFlow}-\left(8.69\times10^{-4}\right)\text{refluxFlow}\,\text{temperatureTray}_1+\left(3.04\times10^{-3}\right)\text{temperatureTray}_1{}^2$
$4.42-\left(9.75\times10^{-3}\right)\text{refluxFlow}-1.40\,\text{temperatureTray}_1{}^{1/3}+\left(9.26\times10^{-6}\right)\text{vaporFlow}\,\text{temperatureTray}_1{}^2$
$-0.25-\dfrac{\left(6.92\times10^{-3}\right)(127.67-\text{vaporFlow})\,\text{temperatureTray}_1{}^3}{\text{refluxFlow}^2}$
$-0.65+\left(2.22\times10^{-3}\right)\text{temperatureTray}_1{}^2-\dfrac{7223.34\,\text{temperatureTray}_1{}^2}{\text{vaporFlow}^3}$
$0.59-\left(4.50\times10^{-4}\right)\text{vaporFlow}-\dfrac{0.12\,\text{refluxFlow}\,\text{temperatureTray}_1}{\text{vaporFlow}}+\left(2.63\times10^{-3}\right)\text{temperatureTray}_1{}^2$
$-0.50+\dfrac{\left(3.60\times10^{-5}\right)\left(7+\sqrt{\text{vaporFlow}}\right)\text{vaporFlow}\,\text{temperatureTray}_1{}^2}{\text{refluxFlow}}$
$-4.61-\left(2.11\times10^{-2}\right)\text{vaporFlow}+\dfrac{96.47}{\text{temperatureTray}_1}+\left(1.05\times10^{-3}\right)\text{vaporFlow}\,\text{temperatureTray}_1-\left(9.81\times10^{-6}\right)\text{temperatureTray}_1{}^3$
$-0.75-\left(9.49\times10^{-3}\right)\text{refluxFlow}+\left(8.19\times10^{-3}\right)\text{vaporFlow}+\dfrac{68.78\,\text{temperatureTray}_1{}^3}{\text{refluxFlow}^3}$
$-0.28+\dfrac{\left(6.12\times10^{-9}\right)\text{vaporFlow}^{5/2}\,\text{temperatureTray}_1{}^3}{\text{refluxFlow}}$
$-7.19+\left(5.44\times10^{-4}\right)\text{vaporFlow}+\dfrac{70.58}{\text{temperatureTray}_1}+0.25\,\text{temperatureTray}_1-\dfrac{0.13\,\text{refluxFlow}\,\text{temperatureTray}_1}{\text{vaporFlow}}$
$6.19-\left(1.11\times10^{-2}\right)\text{refluxFlow}-\dfrac{7459363.20}{\text{vaporFlow}^3}-2.10\sqrt{\text{temperatureTray}_1}+0.27\,\text{temperatureTray}_1$
$8.85-\left(4.51\times10^{-4}\right)\text{vaporFlow}-3.98\sqrt{\text{temperatureTray}_1}+0.54\,\text{temperatureTray}_1-\dfrac{0.13\,\text{refluxFlow}\,\text{temperatureTray}_1}{\text{vaporFlow}}$
$-24.15+\dfrac{2865.36}{\text{vaporFlow}}+\left(1.96\times10^{-2}\right)\text{vaporFlow}+\dfrac{78.56}{\text{temperatureTray}_1}+0.50\,\text{temperatureTray}_1-\dfrac{76.91\,\text{temperatureTray}_1}{\text{vaporFlow}}+\left(1.02\times10^{-5}\right)\text{temperatureTray}_1{}^3$
$1.08-1.81\,\text{refluxFlow}^{1/3}+\left(6.55\times10^{-5}\right)\text{refluxFlow}^2+\left(5.14\times10^{-2}\right)\text{vaporFlow}-\left(8.83\times10^{-5}\right)\text{vaporFlow}^2-\dfrac{7.85}{\text{temperatureTray}_1}+\dfrac{75.70\,\text{temperatureTray}_1{}^3}{\text{refluxFlow}^3}$

Ensemble Quality

R^2	0.988397
⌗ Variables	3
⌗ Models	19
Avg Model Complexity	41.5263

Model Quality

Model	Complexity	$1-R^2$
1	26	0.0229559
2	29	0.0212007
3	29	0.029529
4	33	0.0137302
5	34	0.0182884
6	34	0.0235753
7	38	0.0166514
8	38	0.02229
9	38	0.025075
10	41	0.0109854
11	41	0.0220358



Instead of trying to find THE model, we can exploit the abundance of explored model forms and select a diverse set from the good-and-simple category to form a TRUSTABLE model

Ensemble Performance

The (algorithmically generated) ensemble gives pretty good composite results when asked to extrapolate

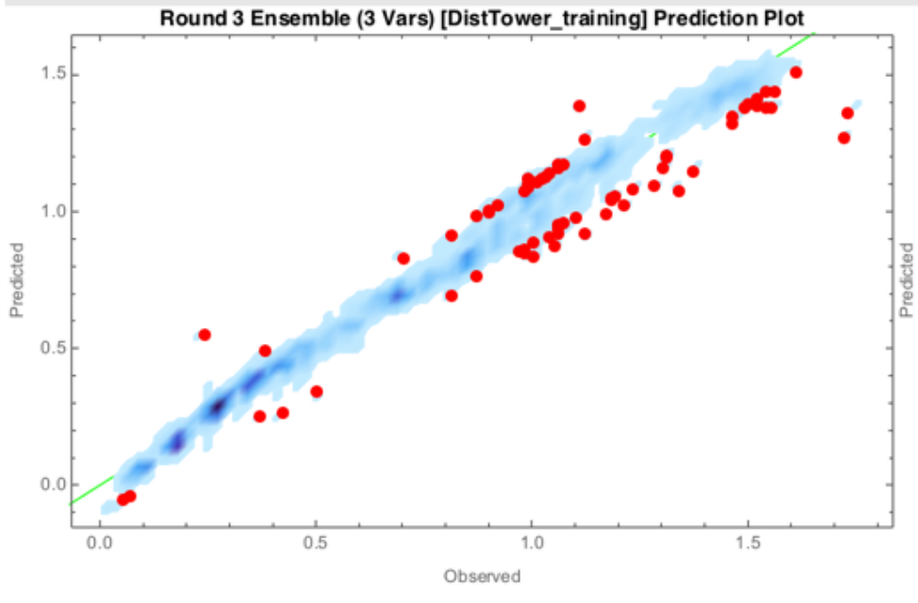
DataModeler 9.3 -

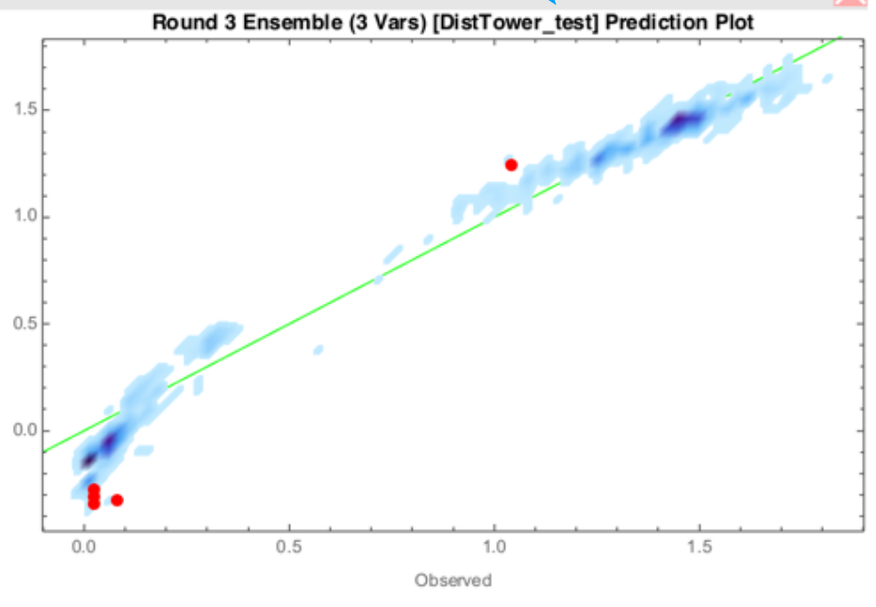
Launch Project | Select Variables | Explore Data | Generate Models | Analyze Models | Test & Validate | What If? | Reports

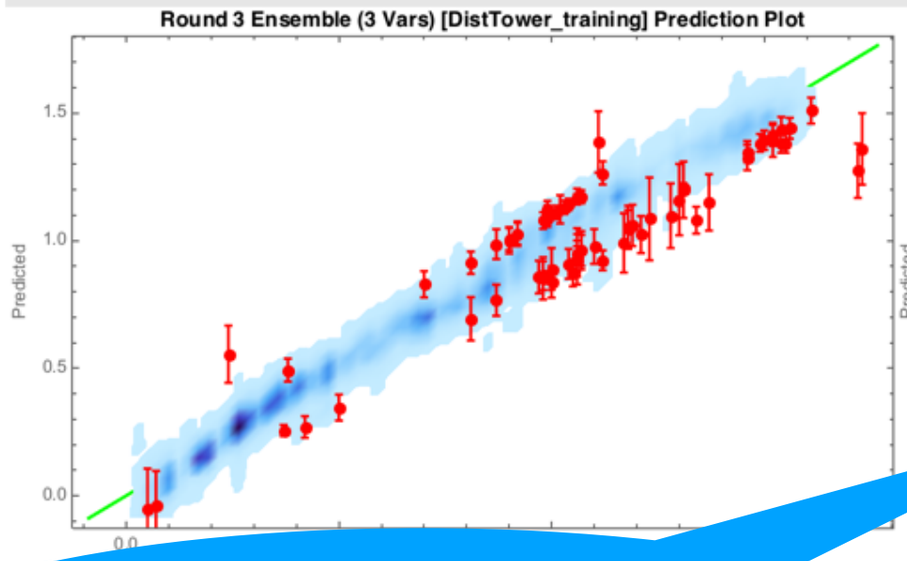
Target Response: Propylene 24

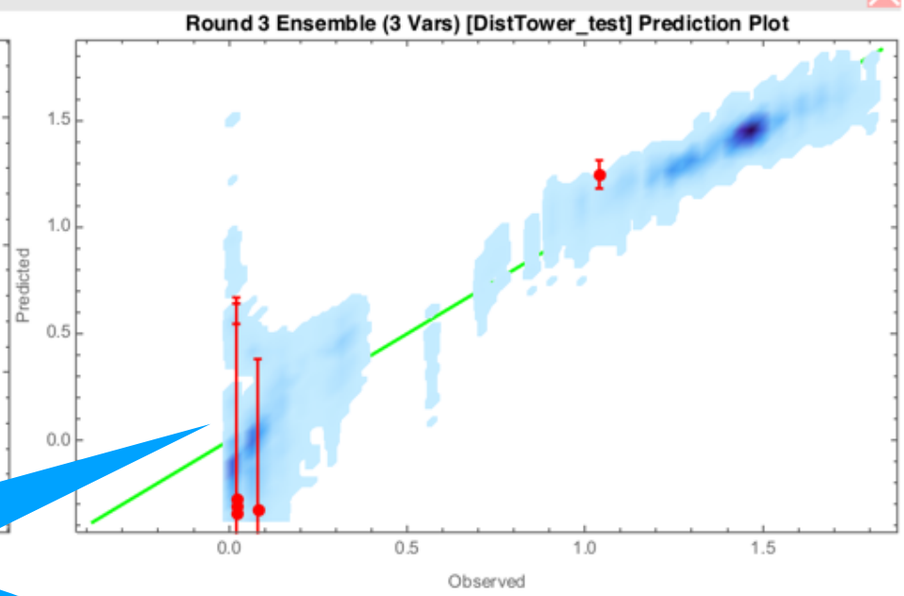
Data | Models | Explore

- Pick Models
  - Model Sets
    - No model sets retrieved.
  - Individual Models
    - Round 3 Ratio Model (1)
    - Round 3 Additive Model (1)
    - Best T&V Model (1)
  - Model Ensembles
    - Round 3 Ensemble (3 Vars)
- Select Target Data Set
  - Data Sets for Model Analysis
    - DistTower_training
    - DistTower_test
- Explore Models
  - Max Explorers: 2
  - Model Statistics
  - Response Plots
    - ResponsePlotExplorer
    - ResponseSurfacePlotExplorer
    - ResponseContourPlotExplorer
  - Prediction Plots (2)
    - ModelPredictionPlot
    - EnsemblePredictionPlot
    - ModelPredictionComparisonPlot
    - ModelResidualPlot
    - EnsembleResidualPlot
  - Data Outliers

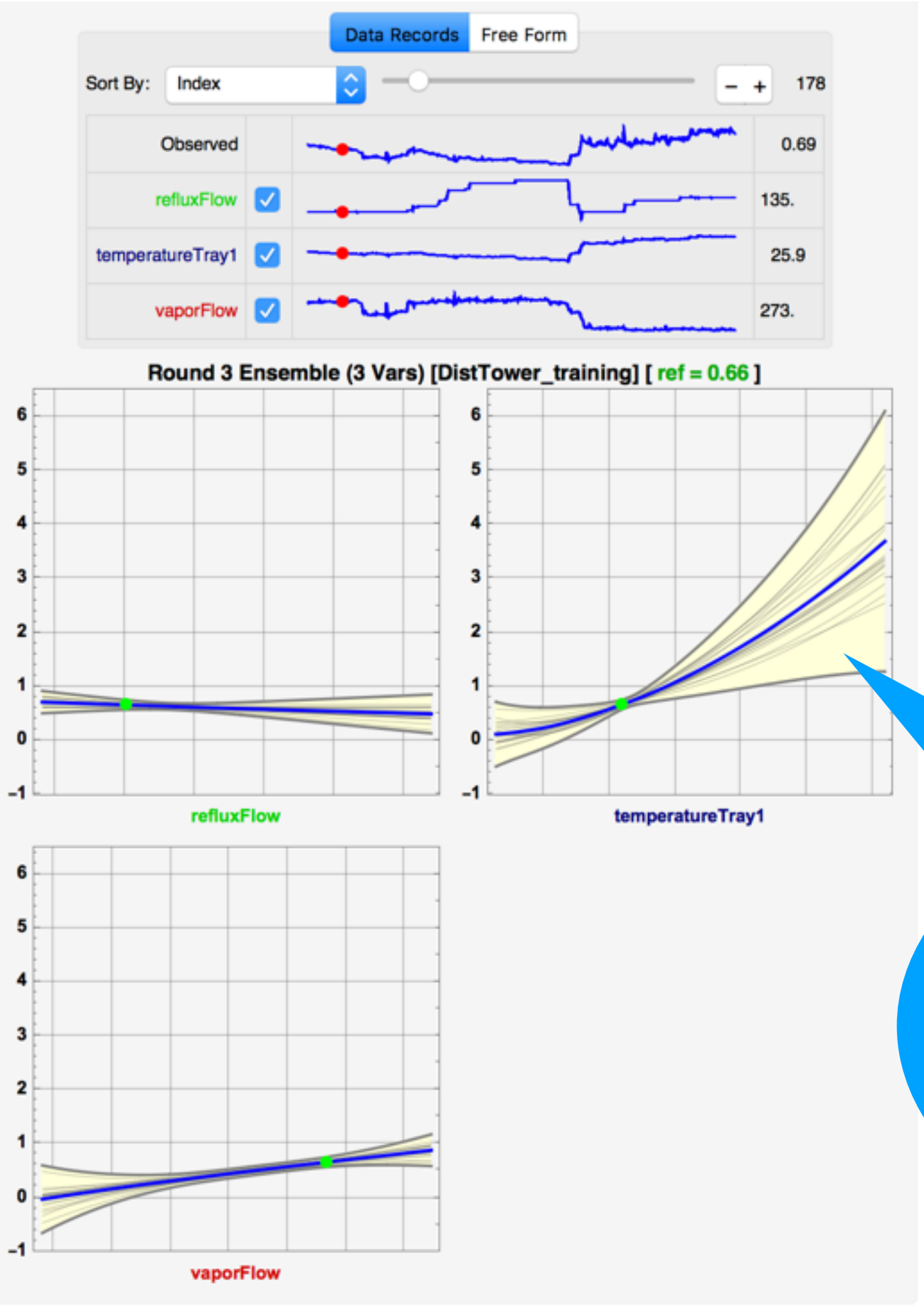


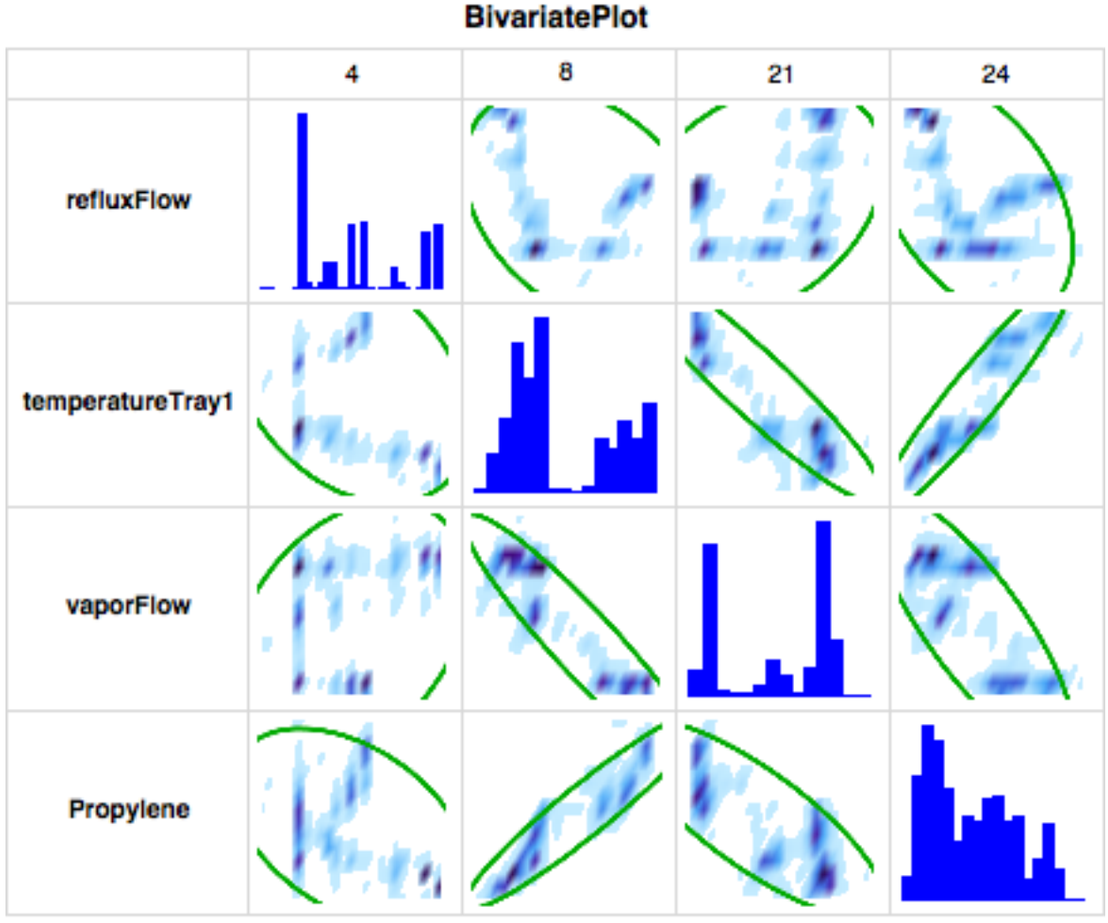




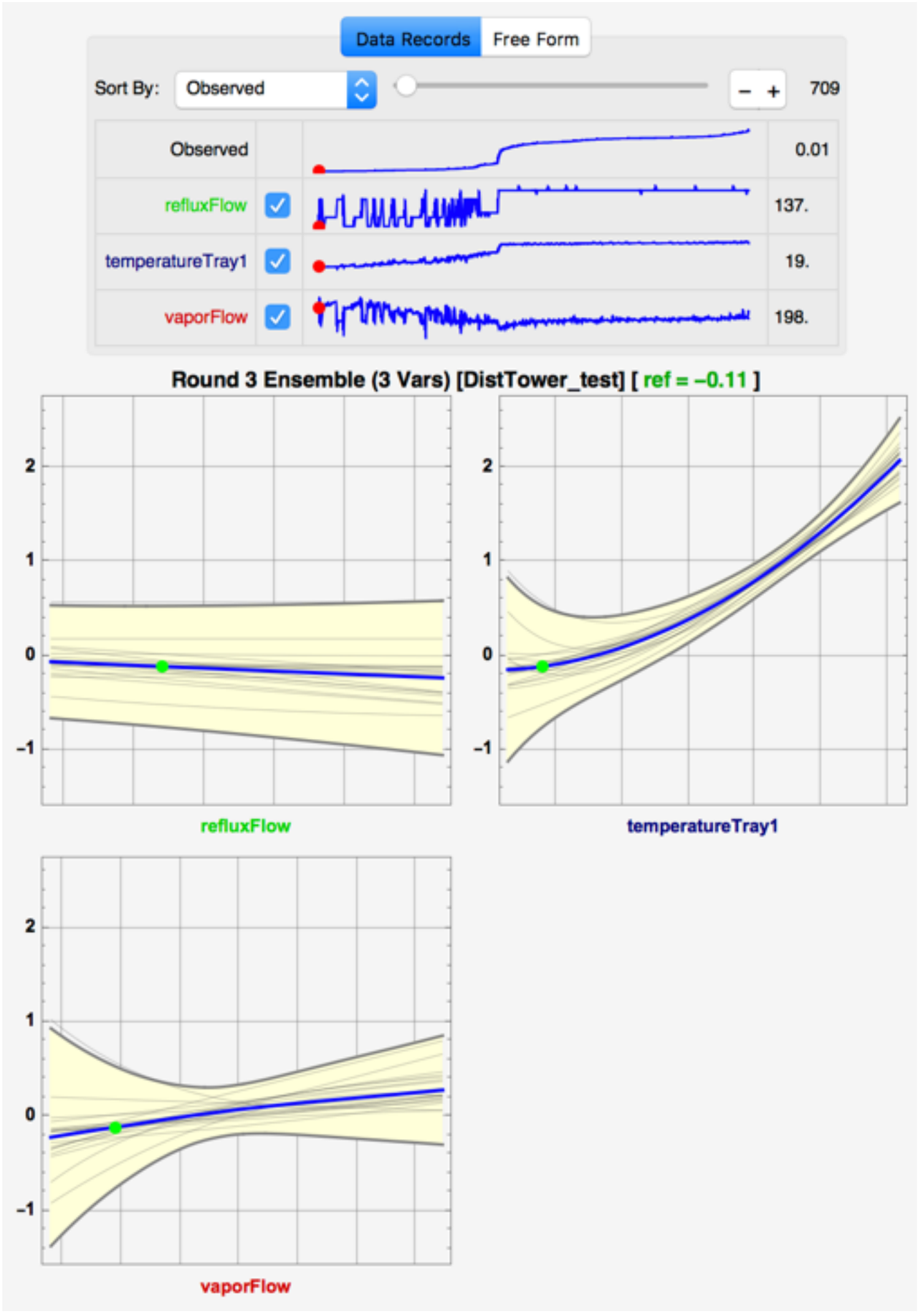


More importantly, the constituent models diverge when asked to extrapolate so have a heads-up on the risk associated with the prediction





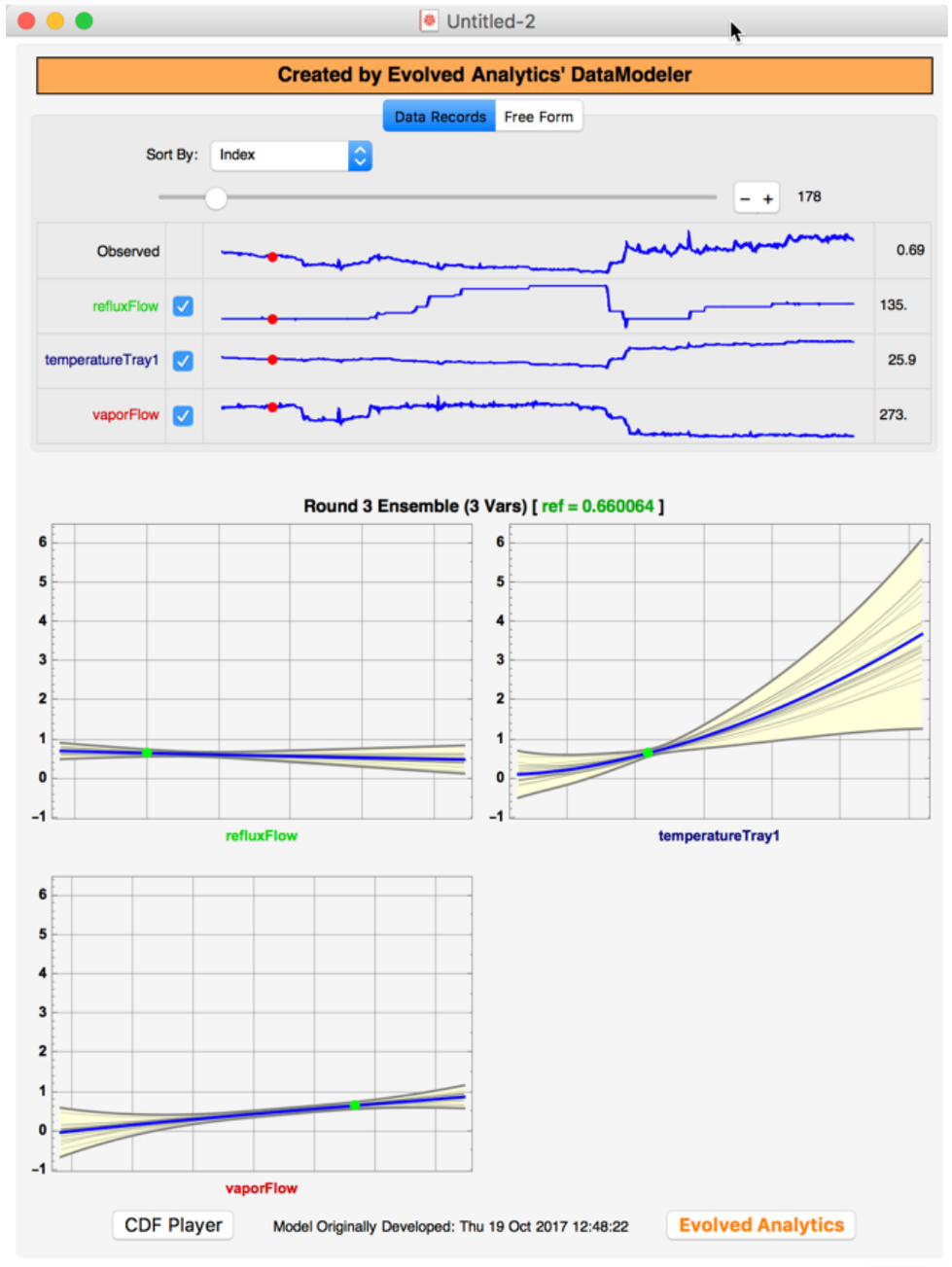
Since the variables in this case are coupled, changing one without a corresponding change of the others would venture into unknown regions of parameter space. This is detected by the ensemble since the constituent models (light gray lines) diverge if we try to change, for example, the tray temperature independently from the observed reference point (green dot)



- Here we are looking at the model prediction in the lowest observed value in the test data set.
- The constituent models diverge when away from known data records which is the desired behavior — even though the actual prediction degrades reasonably gracefully.
- Graceful degradation for extrapolation is a unique benefit of symbolic regression since other data-driven modeling techniques tend to fail spectacularly when asked to extrapolate
- Since these data points were explicitly chosen to test the ensemble ability to detect extrapolation, we can conclude that the ensemble is, in fact, a trustable model

Spiffy Feature

We can export models as CDFs with embedded data so they can be explored without a dependency on either DataModeler or Mathematica



AI ⇒ Augmented Intelligence

- DataModeler easily handles exploration and analysis of multivariate data with correlated inputs, missing values and (now) low-order categoricals

- Searching for simple & accurate models allows identifying driving variables as well as variable combinations useful for developing quality models for deployment

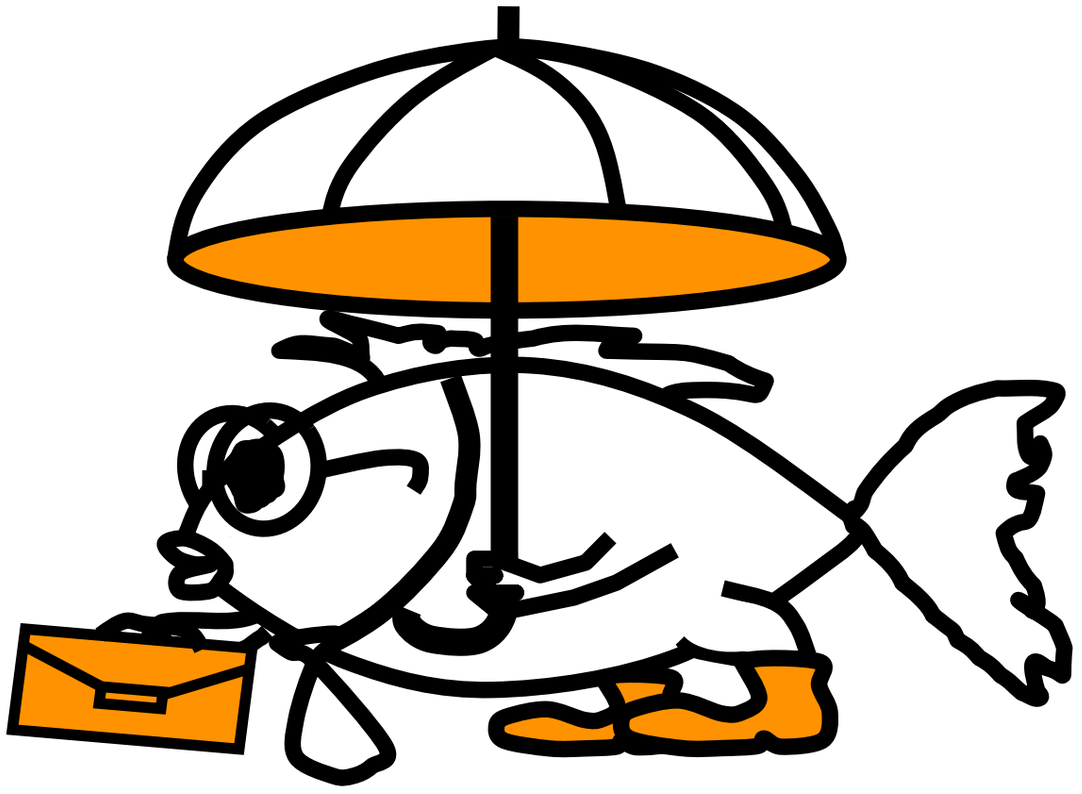
- Algebraic models allows for analyst inspection and insight development

- Ensembles of simple & accurate models can be used as trustable models and obviate the need for data set partitioning — which is important when data is sparse

- Ease-of-use is enhanced with graphical tools supporting an effective analysis workflow.

For More Information

- www.evolved-analytics.com
- mark@evolved-analytics.com



Evolved Analytics

Evolving Daily. Learning Always.

- GUI Development Acknowledgement:
  - Ariel Sepúlveda
  - ariel.sepulveda@prontoanalytics.com
  - www.prontoanalytics.com