The other stuff: Turning machine learning into data science

Brandon Rohrer Senior Data Scientist



The Secret to Data Science

Know your data.

Know your question.

Know your domain.

Know how you're going to use the answer.

Machine Learning answers a question

Prediction: What value?

Classification: What class?

Clustering: Which group?

Anomaly detection: Is it weird?

Recommendation: Which option?

Machine Learning alone is not Data Science

- 1. Data quality: Is it garbage?
- 2. Data exploration: What does it look like?
- 3. Feature Engineering: What does it mean?
- 4. Operationalization: How do I harvest data and deliver answers?

Data exploration

Distribution

Structure

Most relevant features

Visualization

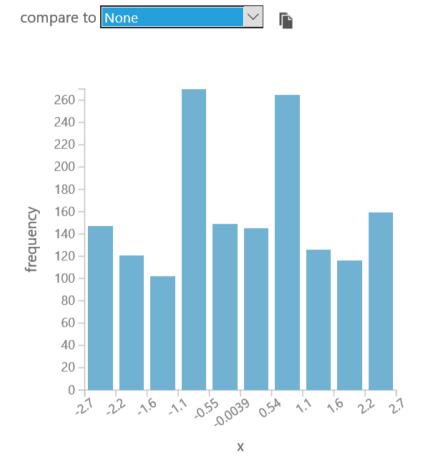
Know your data.

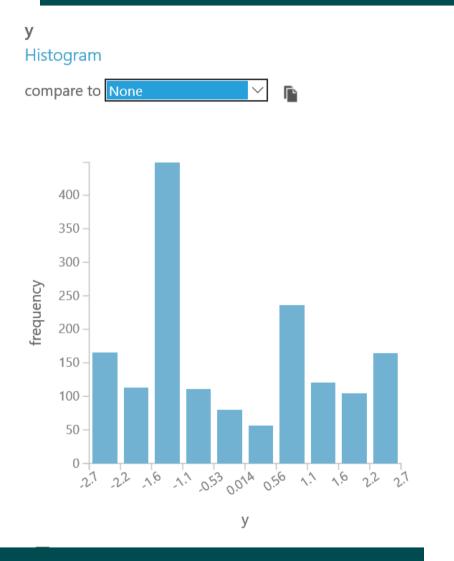
Mystery data

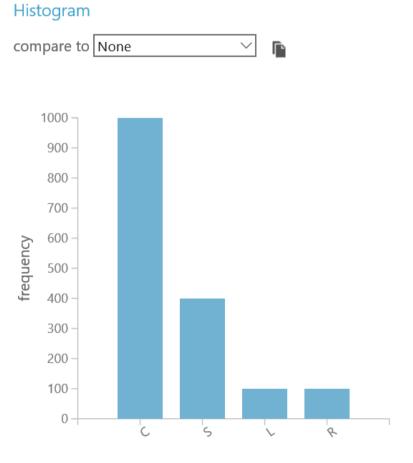
Two features
Four classes
1600 data points



Histogram



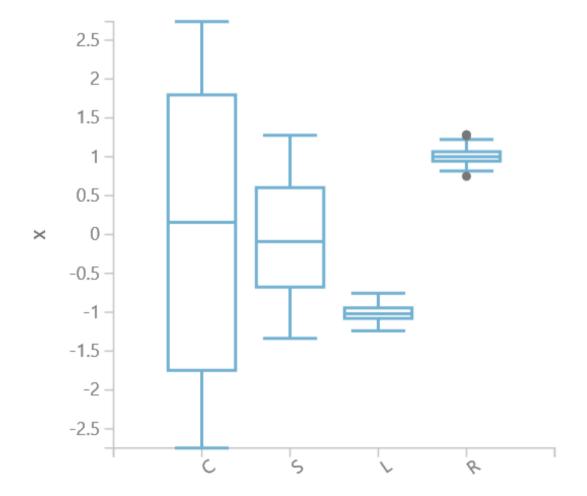




Х

MultiboxPlot

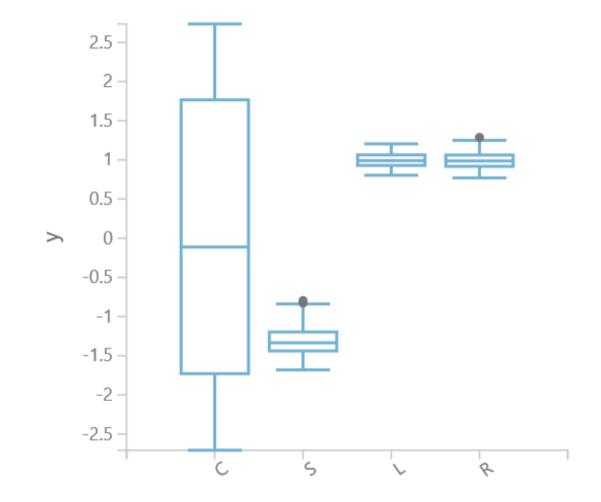
compare to

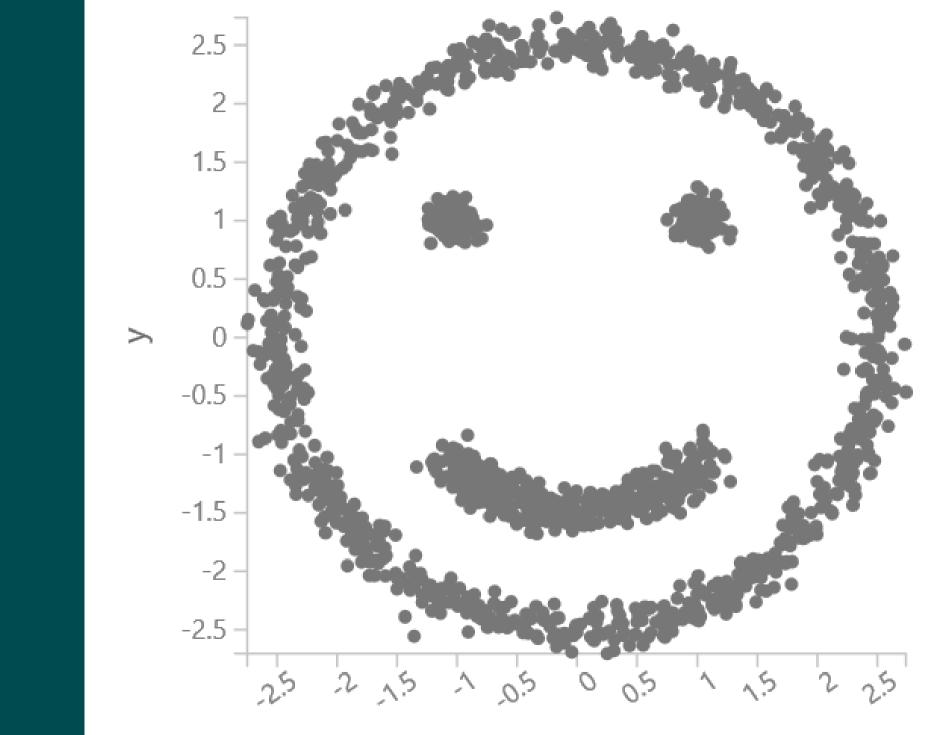


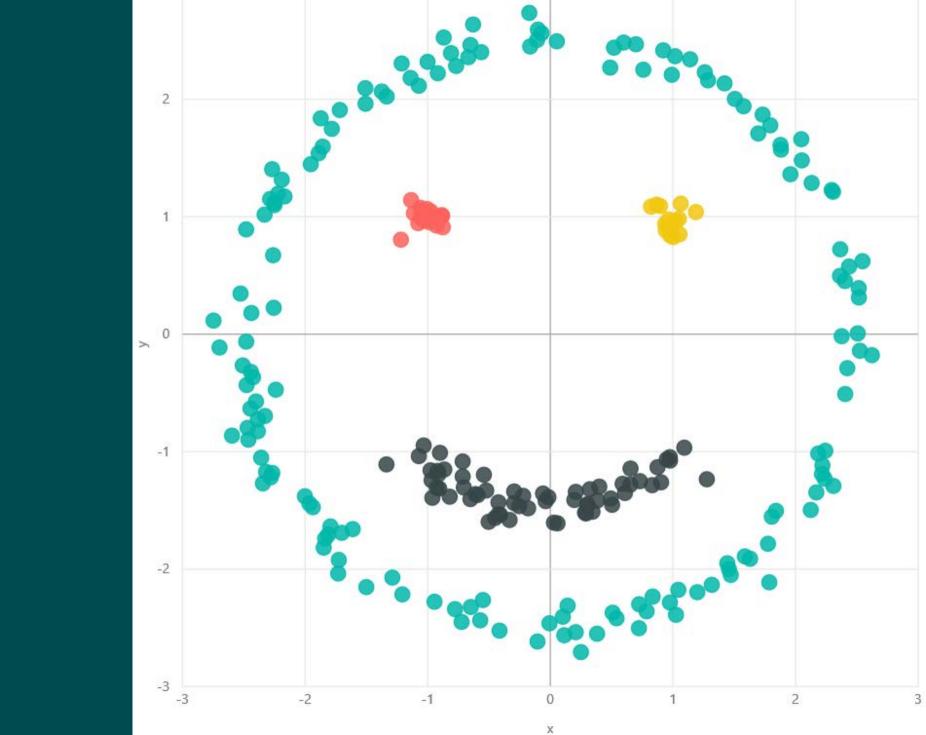


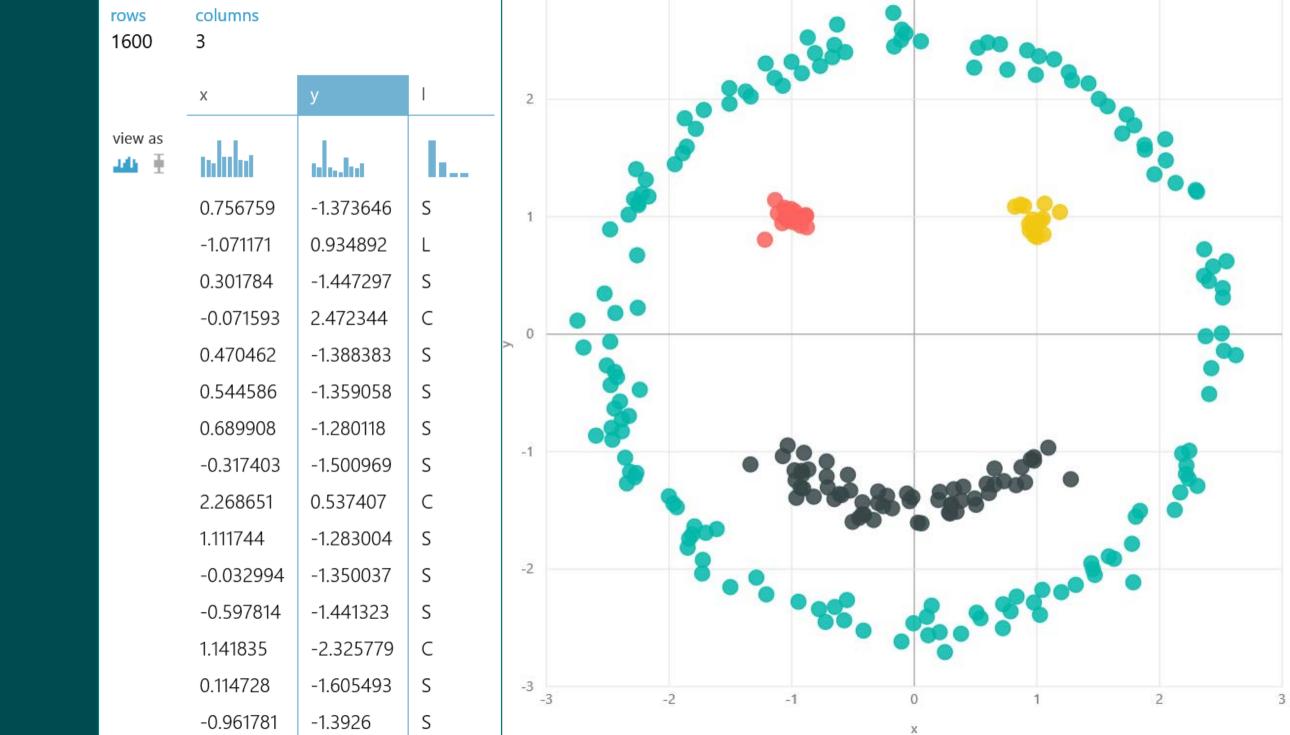
MultiboxPlot

compare to









Data quality

Missing values

Wrong values

Collection flaws

Feature dependence/leakage

Misinterpretation

Know your domain.

Missing values

There are lots of options

Column mean

Column median

Interpolation

Imputation

Not missing at random

Rank order

Missing means zero

Missing means not applicable

Missing means applicable, but lacking information

Mostly empty columns and rows

Multiple representations for "missing"

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
	durdr	han			l	llm	Li	
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick		17		4				
Bruce	37	14	63		1	veggie		NA
Steve	83		77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			_
Carol		3	127	11	1	veggie	1	
Mandy	44	2	68	8	1	chicken		null

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
	dundr	lian			l	llm	L	
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick		17		4				
Bruce	37	14	63		1	veggie		NA
Steve	83		77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			-
Carol		3	127	11	1	veggie	1	
Mandy	44	2	68	8	1	chicken		null

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
	باناسا	Пын			l	Ilm	1 1	
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick	48	17		4				
Bruce	37	14	63		1	veggie		NA
Steve	83	median	77	7	1	chicken		n/a
Clint	27	CUL	118	9		shrimp	3	None
Wanda	19	fill	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			_
Carol	48	3	127	11	1	veggie	1	
Mandy	44	2	68	8	1	chicken		null

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets		emergency_contact
	بالله	lian			l	llm	Ī	í	
Tony	48	27		1	5	shrimp			Pepper
Donald	67	25	86	10	2	beef			Jane
Henry	69	21	95	6	1	chicken	62		Janet
Janet	62	21	110	3	1	beef			Henry
Nick	46	17		4					
Bruce	37	14	63		1	veggie			NA
Steve	83	mean	77	7	1	chicken			n/a
Clint	27		118	9		shrimp	3		None
Wanda	19	fill	52	2	2	shrimp			empty
Natasha	26	4	162	5	3				_
Carol	46	3	127	11	1	veggie	1		
Mandy	44	2	68	8	1	chicken			null

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
	ada.	limi			l	llm	1 ,	
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick	46	17		4				
Bruce	37	14	63		1	veggie		NA
Steve	83	12	77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7 int	erpol	ation	2	shrimp		empty
Natasha	26	4	СГРОГ		3			_
Carol	46	3	127	11	1	veggie	1	
Mandy	44	2	68	8	1	chicken		null

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
	dalah	limi			l	llııı		
Tony	48	27	250	1	5	shrimp		Pepper
Donald	67	25	86	not mis	ssina	beef		Jane
Henry	69	21	95	at rand		chicken	62	Janet
Janet	62	21	110	attani		beef		Henry
Nick	46	17	250	4				
Bruce	37	14	63		1	veggie		NA
Steve	83	12	77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			_
Carol	46	3	127	11	1	veggie	1	
Mandy	44	2	68	8	1	chicken		null

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
	dinti	Hinti			l	llm	L	
Tony	48	27	78	1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick	34	17	60	imput	ation			
Bruce	37	14	63	imput		veggie		NA
Steve	83	16	77	(MI	LE)	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			_
Carol	27	3	127	11	1	veggie	1	
Mandy	44	2	68	8	1	chicken		null

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
	بالألبان	lmar	lii i		l	llm	ĺπ	
Tony	48	27	250	1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick	46	17	250	4				
Bruce	mic	scina ron		12	1	veggie		NA
Steve	11115	ssing ran	IK	7	1	chicken		n/a
Clint	4	order		9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			_
Carol	46	3	127	11	1	veggie	1	
Mandy	44	2	68	8	1	chicken		null

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact	parking_space_lsMissing
	بالله	lim	lii. i	. 1	l	llm	Lτ		
Tony	48	27	250	1	5	shrimp		Pepper	false
Donald	67	25	86	10	2	beef		Jane	false
Henry	69	21	95	6	1	chicken	62	Janet	false
Janet	62	21	110	3	1	beef		Henry	false
Nick	46	17	250	4	D O	infor	po o ti		false
Bruce	37	14	63	-99		infor	mau	on	true
Steve	83	12	77	7	1	chicken		n/a	false
Clint	27	9	118	9		shrimp	3	None	false
Wanda	19	7	52	2	2	shrimp		empty	false
Natasha	26	4	162	5	3			_	false
Carol	46	3	127	11	1	veggie	1		false
Mandy	44	2	68	8	1	chicken		null	false

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact	parking_space_lsMissing
	ada	limi	lui, i	. 1	di	llm	Li		
Tony	48	27	250	1	5	shrimp		Pepper	false
Donald	67	25	86	10	2	beef		Jane	false
Henry	69	21	95	6	1	chicken	62	Janet	false
Janet	62	21	110	3	1	beef		Henry	false
Nick	46	17	250	4	0				false
Bruce	37	14	63	-99	1		zero		true
Steve	83	12	77	7	1			nent	false
Clint	27	9	118	9	0	repi	acen	ICIIC	false
Wanda	19	7	52	2	2	shrimp		empty	false
Natasha	26	4	162	5	3			_	false
Carol	46	3	127	11	1	veggie	1	нин	false
Mandy	44	2	68	8	1	chicken		null	false

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets		emergency_contact	parking_space_lsMissing	attending_party_lsMissing
	ada	lim	lm, 1	. 1	dia.	llm		I			
Tony	48	27	250	1	5	shrimp			Pepper	false	false
Donald	67	25	86	10	2	beef			Jane	false	false
Henry	69	21	95	6	1	chicken	62		Janet	false	false
Janet	62	21	110	3	1	beef			Henry	false	false
Nick	46	17	250	4	0						true
Bruce	37	14	63	-99	1	zer	O	re	eplacem	ient (false
Steve	83	12	77	7	1				nissing 1		false
Clint	27	9	118	9	0	VVI	CII	11		iag	true
Wanda	19	7	52	2	2	shrimp			empty	false	false
Natasha	26	4	162	5	3				_	false	false
Carol	46	3	127	11	1	veggie	1			false	false
Mandy	44	2	68	8	1	chicken			null	false	false

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	_	emergency_contact	parking_space_IsMissing	attending_party_IsMissing	emergency_contact (2)
	بالألية	limi	lm, 1	. 1	dia.	llm		ı		I.	I.	
Tony	48	27	250	1	5	shrimp			Pepper	false	false	present
Donald	67	25	86	10	2	beef			Jane	false	false	present
Henry	69	21	95	6	1	chicken	62		Janet	false	false	present
Janet	62	21		_					Henry	false	false	present
Nick	46	17		mul	tiple				no	false	true	absent
Bruce	37	14			the state of the s				NA	true	false	absent
Steve	83	12	rer	orese	ntatic	ns			n/a	false	false	absent
Clint	27	9							None	false	true	absent
Wanda	19	7	10	ווו זכ	issing				empty	false	false	absent
Natasha	26	4							-	false	false	absent
Carol	46	3	127	11	1	veggie	1			false	false	absent
Mandy	44	2	68	8	1	chicken			null	false	false	absent

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets		emergency_contact	parking_space_IsMissing	attending_party_IsMissing
	ada.	lim	lii i	. 1	dia.	llm		í			I.
Tony	48	27	250	1	5	shrimp			Pepper	false	false
Donald	67	25	86	10	2	beef			Jane	false	false
Henry	69	21	95	6	1	chicken	62		Janet	false	false
Janet	62	21	ي مي ا		41	beef			Henry	false	false
Nick	46	17	arop	o mos	tiy				no	false	true
Bruce	37	14 e r	npty	/ colu	mns	veggie			NA	true	false
Steve	83	12	ניסוי	COIG	111115	chicken			n/a	false	false
Clint	27	9	118	9	0	shrimp	3		None	false	true
Wanda	19	7	52	2	2	shrimp			empty	false	false
Natasha	26	4	162	5	3				_	false	false
Carol	46	3	127	11	1	veggie	1			false	false
Mandy	44	2	68	8	1	chicken			null	false	false

Column 0	age	years_seniority	income	parking_space	attending_party	entree	emergency_contact	parking_space_lsMissing	attending_party_lsMissing	emergency_con (2)
dr	on	rows r	niss	sina	dia.	llm		I.	I.	li .
	_				5	shrimp	Pepper	false	false	present
	Crit	tical va	alue	S	2	beef	Jane	false	false	present
пенну	U J	۷1	95	U	1	chicken	Janet	false	false	present
Janet	62	21	110	3	1	beef	Henry	false	false	present
Nick	46	17	250	4	0		no	false	true	absent
Bruce	37	14	63	-99	1	veggie	NA	true	false	absent
Steve	83	12	77	7	1	chicken	n/a	false	false	absent
Clint	27	9	118	9	0	shrimp	None	false	true	absent
Wanda	19	7	52	2	2	shrimp	empty	false	false	absent
Natasha	26	4	162	5	3		_	false	false	absent
Carol	46	3	127	11	1	veggie		false	false	absent
Mandy	44	2	68	8	1	chicken	null	false	false	absent

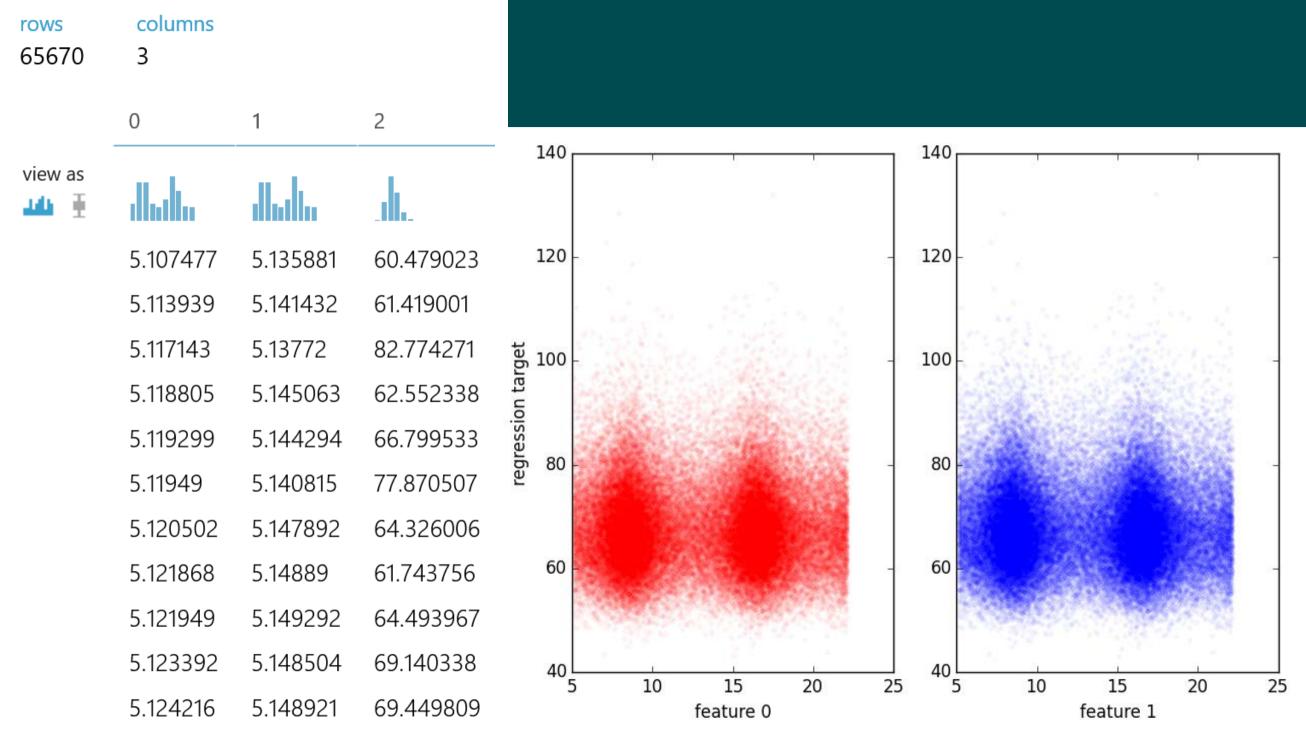
Column 0	age	years_seniority	income	parking_space	attending_party	entree	emergency_contact	parking_space_lsMissing	attending_party_IsMissing	emergency_cont (2)
	nd da	Ш	lm .	. 1	Jan.	llu		L	I.	h .
Tony	48	27	250	1	5	shrimp	Pepper	false	false	present
Donald	67	25	86	10	2	beef	Jane	false	false	present
Henry	69	21	95	6	1	chicken	Janet	false	false	present
Janet	62	21	110	3	1	beef	Henry	false	false	present
Bruce	37	14	63	-99	1	veggie	NA	true	false	absent
Steve	83	12	77	7	1	chicken	n/a	false	false	absent
Clint	27	9	118	9	0	shrimp	None	false	true	absent
Wanda	19	7	52	2	2	shrimp	empty	false	false	absent
Carol	46	3	127	11	1	veggie		false	false	absent
Mandy	44	2	68	8	1	chicken	null	false	false	absent

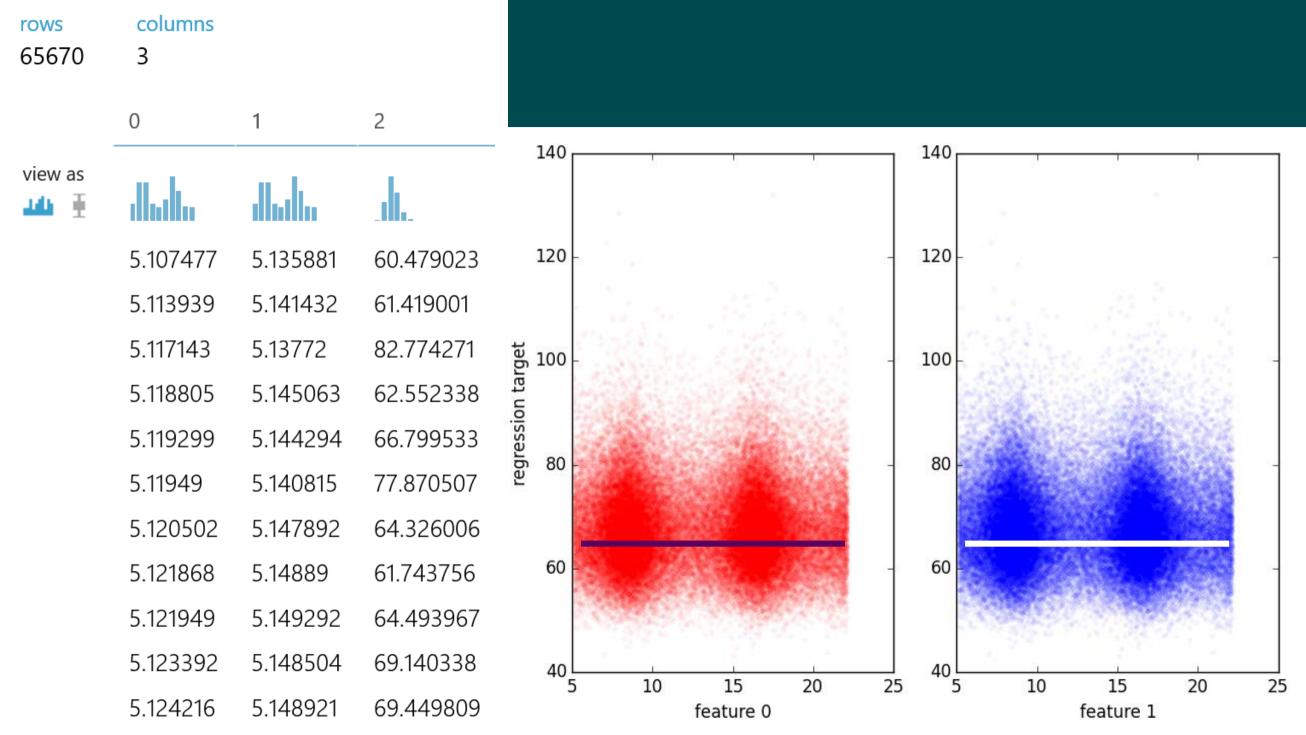
Feature Engineering

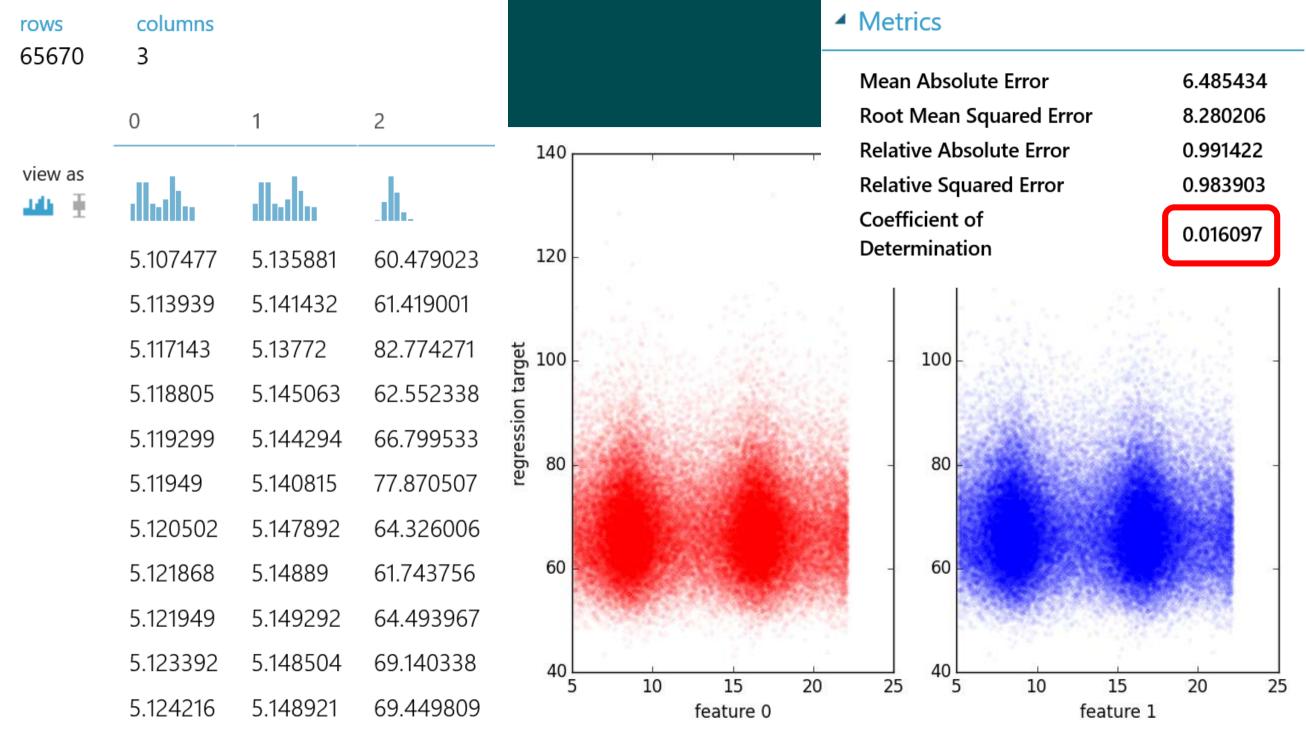
Domain knowledge
Combining features
Deep learning
Ignoring irrelevant features

Know your question.

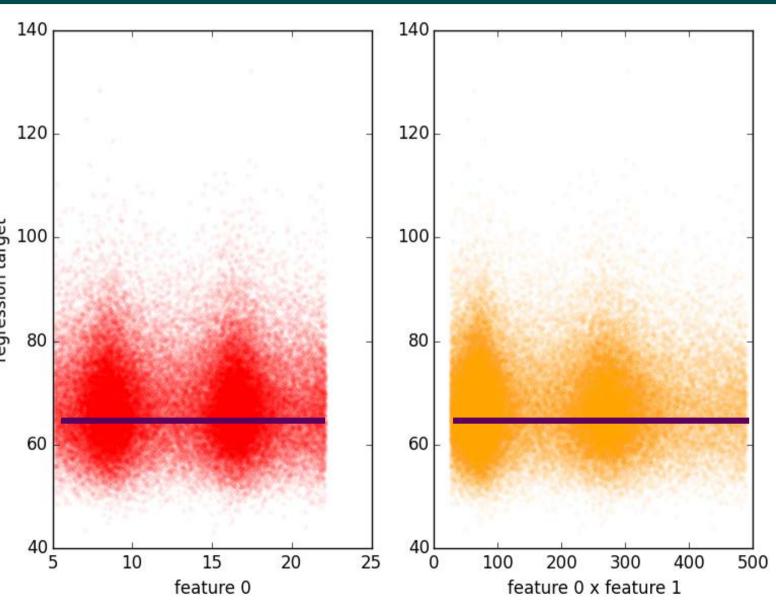
rows 65670	columns 3		
	0	1	2
view as			
	5.107477	5.135881	60.479023
	5.113939	5.141432	61.419001
	5.117143	5.13772	82.774271
	5.118805	5.145063	62.552338
	5.119299	5.144294	66.799533
	5.11949	5.140815	77.870507
	5.120502	5.147892	64.326006
	5.121868	5.14889	61.743756
	5.121949	5.149292	64.493967
	5.123392	5.148504	69.140338
	5.124216	5.148921	69.449809

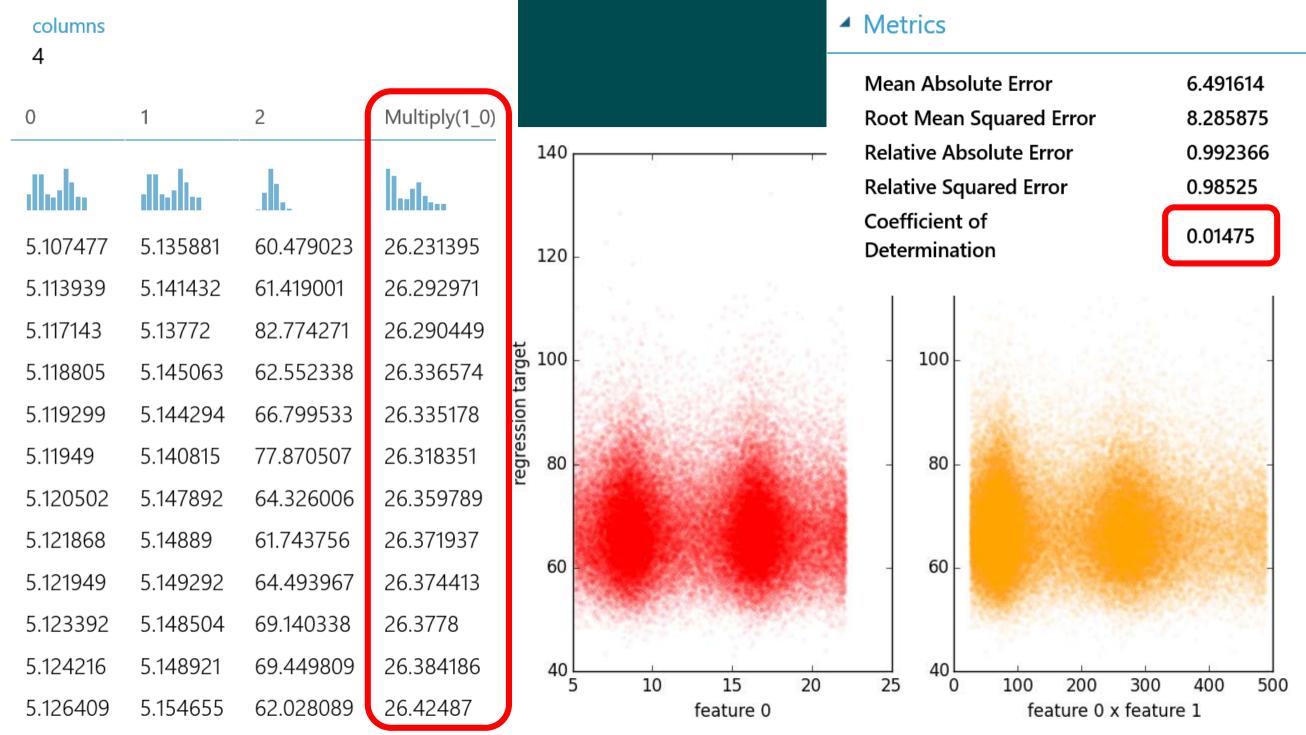


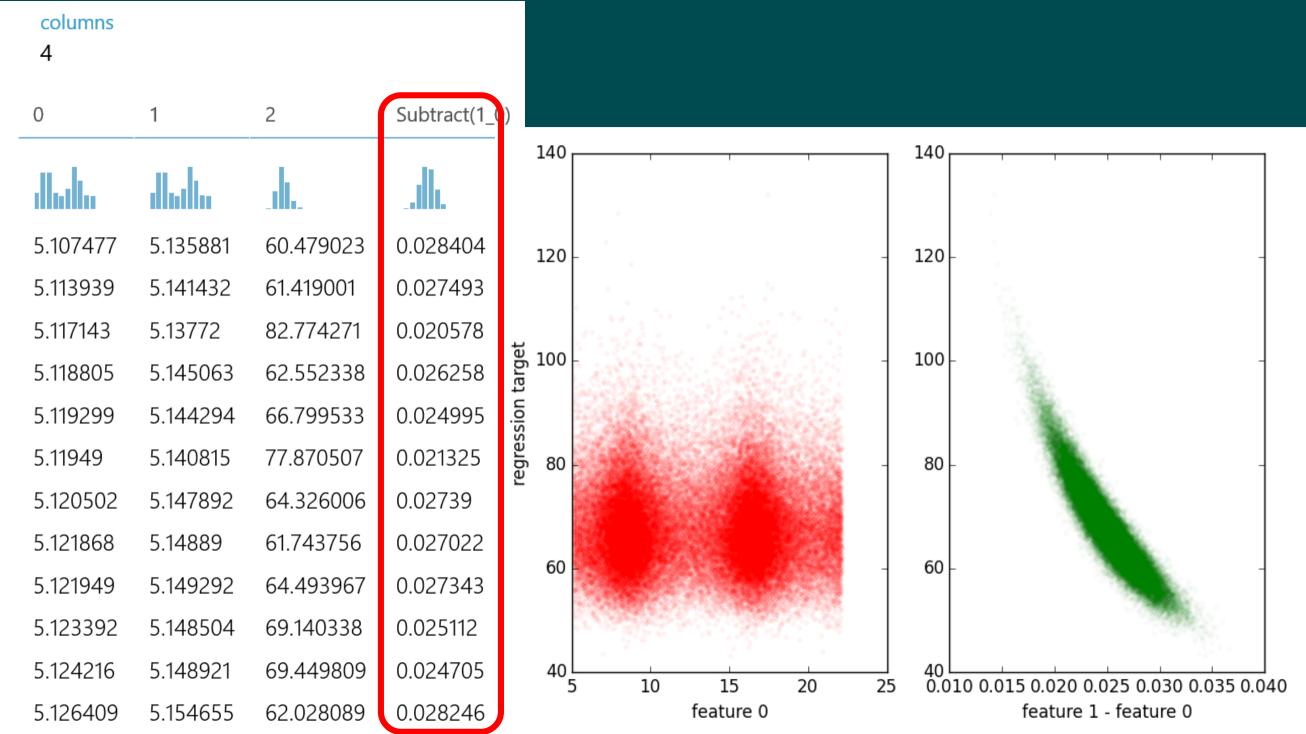


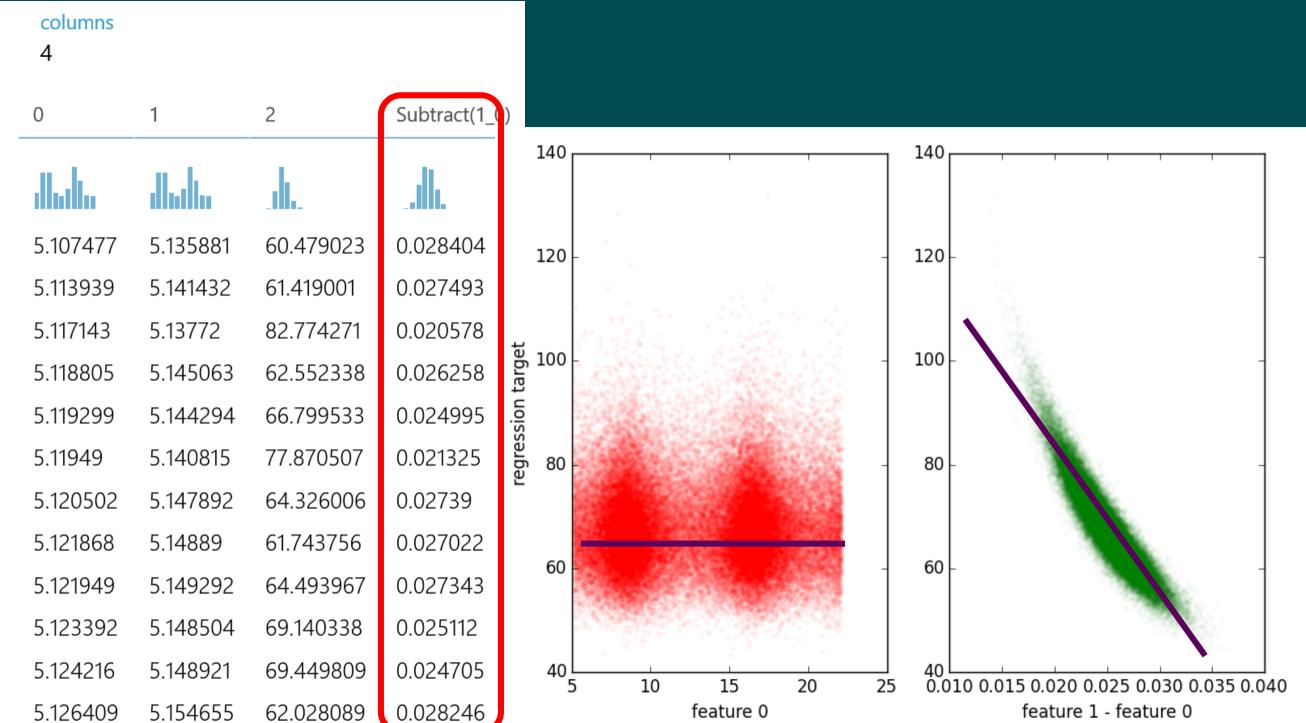


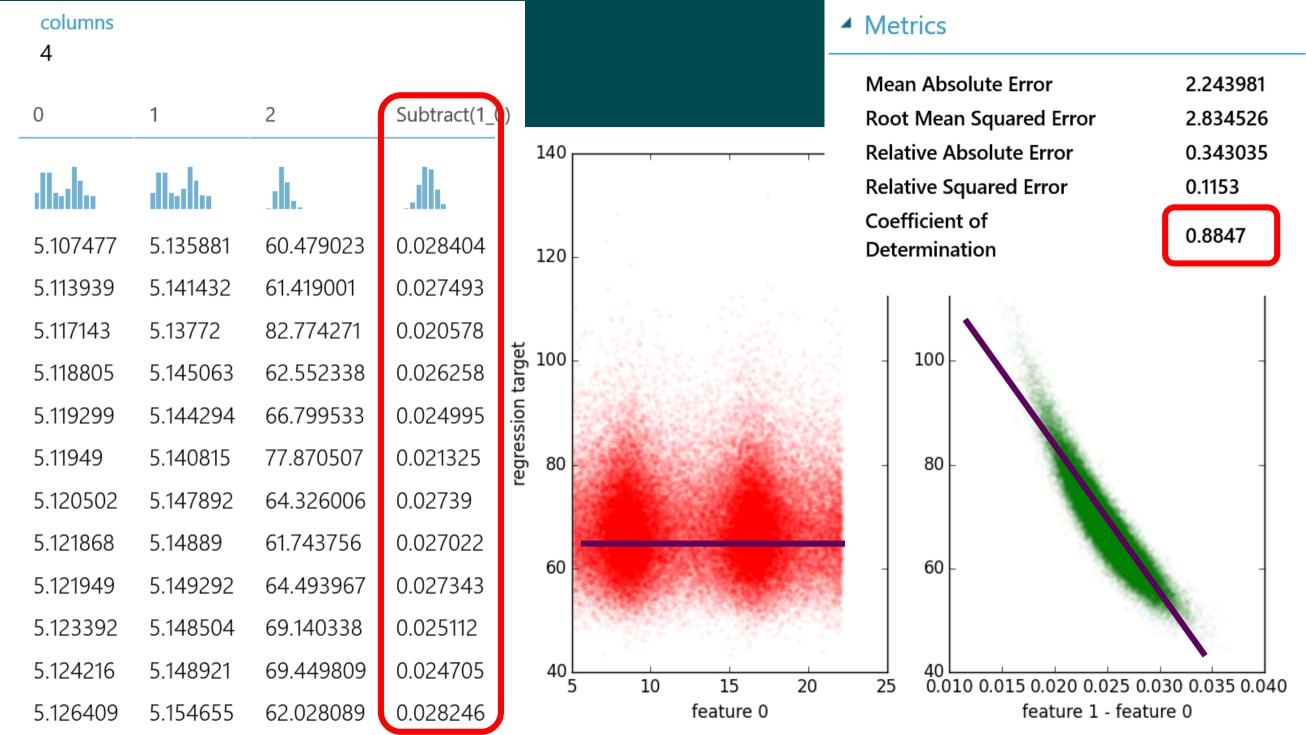
0	1	2	Multiply(1_0)	
dhdh	Illullu	.ll.	l _{uth}	1
5.107477	5.135881	60.479023	26.231395	1
5.113939	5.141432	61.419001	26.292971	
5.117143	5.13772	82.774271	26.290449	يب
5.118805	5.145063	62.552338	26.336574	regression target
5.119299	5.144294	66.799533	26.335178	sion
5.11949	5.140815	77.870507	26.318351	egres
5.120502	5.147892	64.326006	26.359789	_
5.121868	5.14889	61.743756	26.371937	
5.121949	5.149292	64.493967	26.374413	
5.123392	5.148504	69.140338	26.3778	
5.124216	5.148921	69.449809	26.384186	
5.126409	5.154655	62.028089	26.42487	











Operationalization

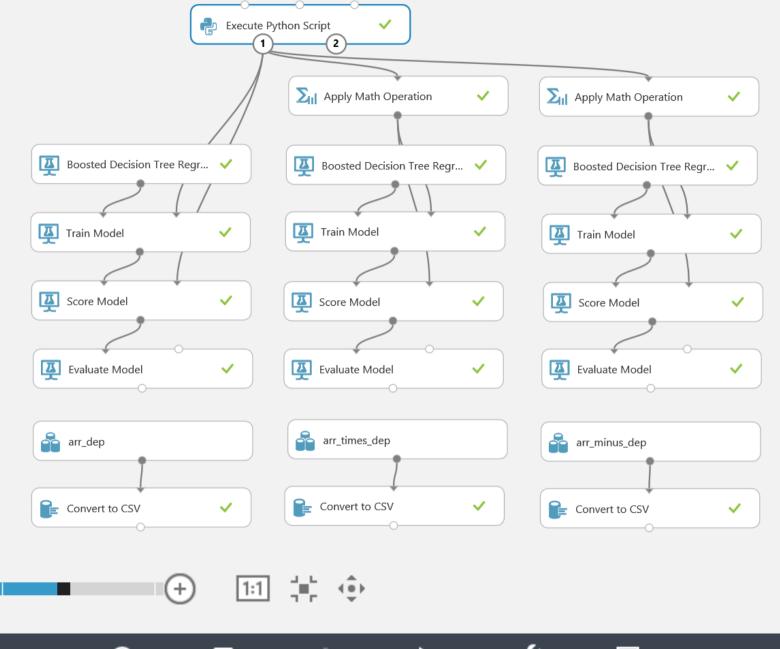
Getting the data to the algorithm

Getting the data cleaned

Getting the results to those who need them

Know how you are going to use the answer.

Creating a RESTful API







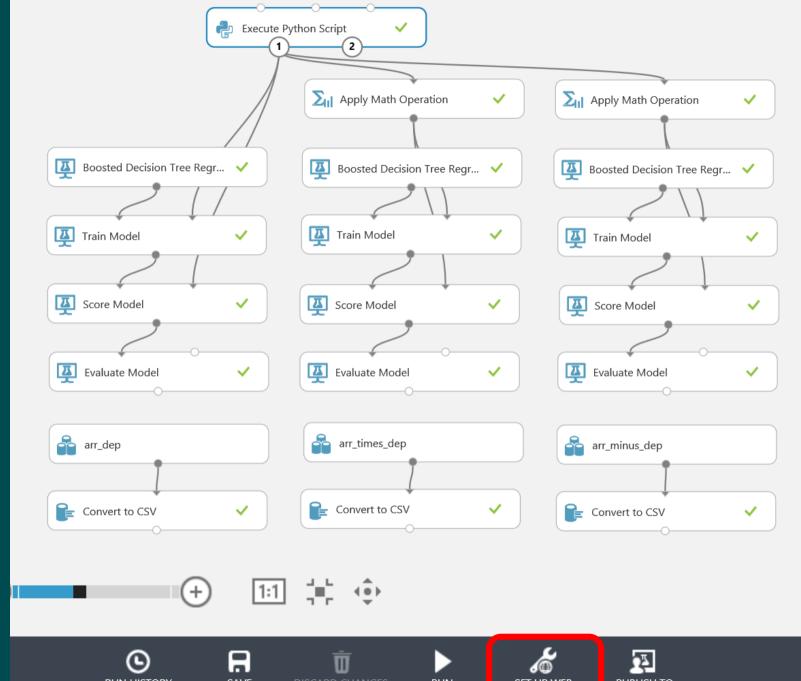








Creating a RESTful API





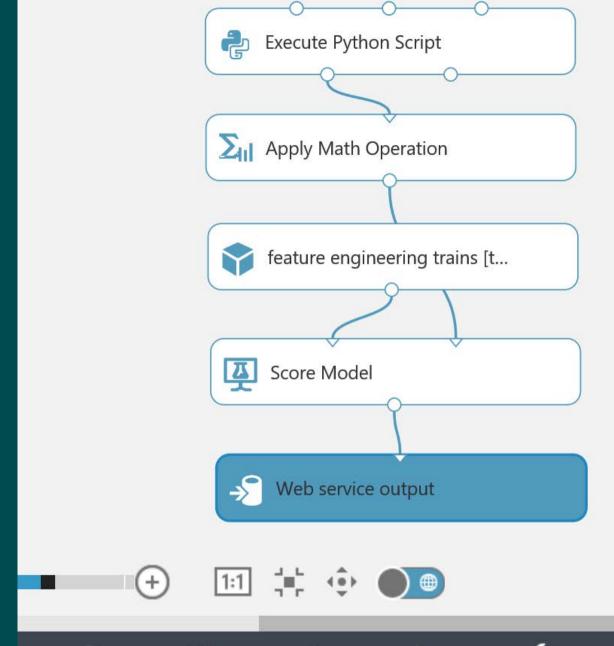














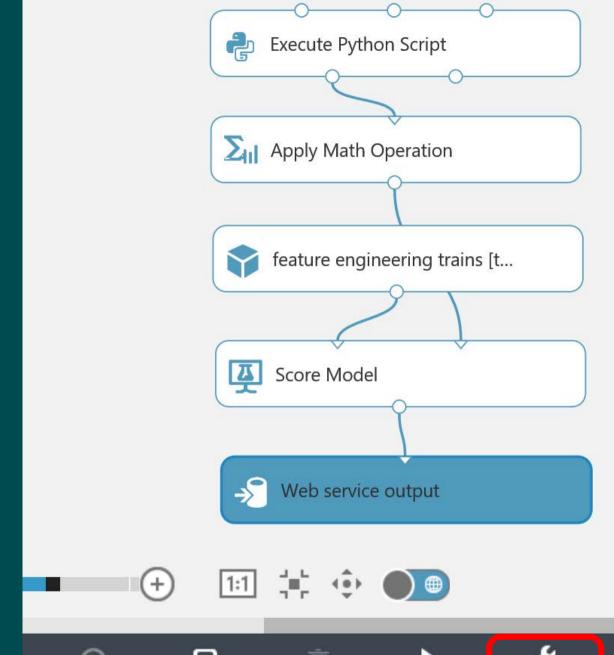
























feature engineering: trains [scoring exp.]

DASHBOARD CONFIGURATION

General

Published experiment

View snapshot View latest

Description

No description provided for this web service.

API key

oPYrmjj9qpuZlgdezr2PTPifG0D2YJ88v5gUOde2uGCNS9pNl6IoA2QHSEwSdcAUhEfD1QSU6SIzVc1BsqJiQQ==



Download Excel Workbook

Default Endpoint

API HELP PAGE TEST APPS

Test

REQUEST/RESPONSE

BATCH EXECUTION

feature engineering: trains [scoring exp.]

DASHBOARD CONFIGURATION

General

Published experiment

View snapshot View latest

Description

No description provided for this web service.

API key

oPYrmjj9qpuZlgdezr2PTPifG0D2YJ88v5gUOde2uGCNS9pNl6IoA2QHSEwSdcAUhEfD1QSU6SIzVc1BsqJiQQ==



Default Endpoint

API HELP PAGE TEST APPS

REQUEST/RESPONSE

Test

Download Excel Workbook

BATCH EXECUTION

feature engineering: trains [scoring exp.]

DASHBOARD CONFIGURATION

General

Published experiment

View snapshot View latest

Description

No description provided for this web service.

API key

oPYrmjj9qpuZlgdezr2PTPifG0D2YJ88v5gUOde2uGCNS9pNl6IoA2QHSEwSdcAUhEfD1QSU6SIzVc1BsqJiQQ==



Default Endpoint

API HELP PAGE TEST APPS

REQUEST/RESPONSE

Test

Download Excel Workbook

BATCH EXECUTION

Request Response API Documentation for feature engineering: ti [Scoring Exp.]

Updated: 10/08/2015 22:03

No description provided for this web service.

- Previous version of this API
- Submit a request
- Output Parameters
- ASP.Net App Template for RRS
- Sample Code

OData Endpoint Address

https://ussouthcentral.services.azureml.net/odata/workspaces/ed2d176323ff470b8de0fa312268b88e/services/58c261e5aa5b4c89b6a

Request

Method Request URI

POST

https://ussouthcentral.services.azureml.net/workspaces/ed2d176323ff470b8de0fa312268b88e/services/58c261e5aa5b4c89api-version=2.0&details=true

Request Response API Documentation for feature engineering: ti [Scoring Exp.]

Updated: 10/08/2015 22:03

No description provided for this web service.

- Previous version of this API
- Submit a request
- Output Parameters
- ASP.Net App Template for RRS
- Sample Code

OData Endpoint Address

https://ussouthcentral.services.azureml.net/odata/workspaces/ed2d176323ff470b8de0fa312268b88e/services/58c261e5aa5b4c89b6a

Request

Method Request URI

POST

https://ussouthcentral.services.azureml.net/workspaces/ed2d176323ff470b8de0fa312268b88e/services/58c261e5aa5b4c89api-version=2.0&details=true

Know your data.

Know your question.

Know your domain.

Know how you're going to use the answer.

https://gallery.cortanaanalytics.com/Collection/The-other-stuff-Getting-from-machine-learning-to-data-science-1

(http://bit.ly/1GBV3Gi)

Thoughts to share? Questions? Connect with me offline:
Brandon Rohrer on LinkedIn
@_brohrer_ on Twitter
brohrer@microsoft.com



