Hack2O analysis

Andrew Wang

General feature table

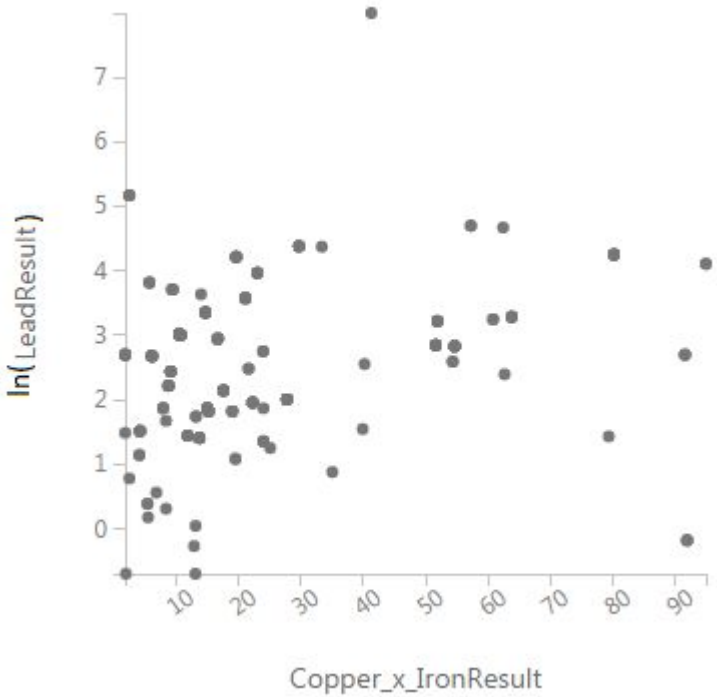
PropertyInfoDBID	ChlorideResult	ClResult	PhResult	CopperResult	IronResult	Copper_x_IronResult	LeadResult
8	9.04	0.34	1.412538	162	0.375	60.75	25.8
8	9.04	0.64	0.724436	162	0.375	60.75	25.8
15	9.46	0.7	0.645654	23.1	0.08	1.848	4.44
17	9.47	0.68	0.47863	218	0.552	120.336	25.1
17	9.47	0.78	0.512861	218	0.552	120.336	25.1
25	9.28	0.64	0.467735	75.9	0.278	21.1002	35.8
25	9.28	0.65	0.346737	75.9	0.278	21.1002	35.8
25	9.28	0.92	0.40738	75.9	0.278	21.1002	35.8
25	9.28	1.28	0.562341	75.9	0.278	21.1002	35.8
25	9.28	1.41	0.707946	75.9	0.278	21.1002	35.8
27	9.41	0.6	0.467735	43.1	0.212	9.1372	11.5
27	9.41	0.68	0.457088	43.1	0.212	9.1372	11.5
27	9.41	0.88	0.524807	43.1	0.212	9.1372	11.5

Processing the general feature table

- pH -> 10^{7-pH}, z-score
- Cu/Fe/Pb aggregation: highest recorded per house
- Pb outliers

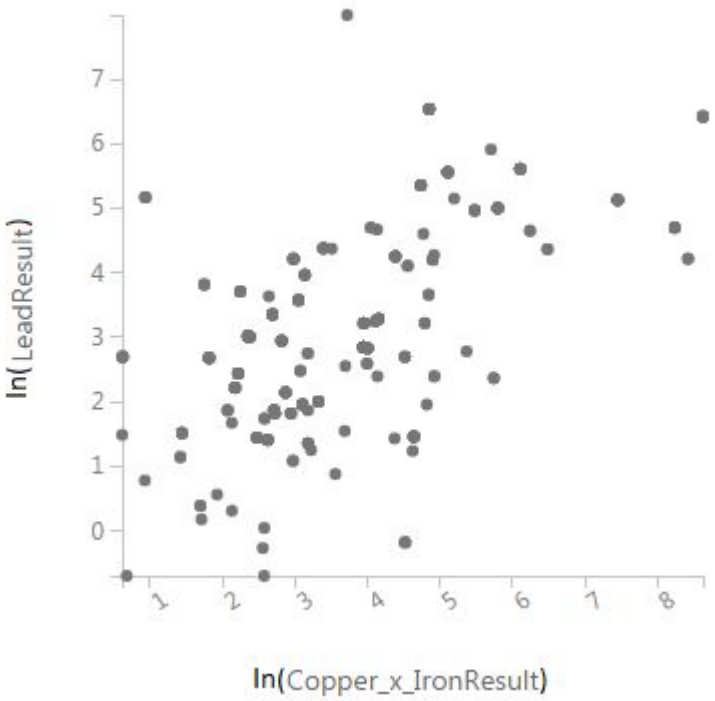
Cu * Fe, ln(Pb)

CopperResult	IronResult	Copper_x_IronResult	LeadResult
-0.045622	-0.090117	-0.093585	0.065356
0.013499	0.074789	0.062216	0.01336
-0.156592	-0.28656	-0.203495	-0.104994
NaN	NaN	NaN	NaN
1	0.351567	0.904722	0.385348
0.351567	1	0.591881	0.399933
0.904722	0.591881	1	0.397958
0.385348	0.399933	0.397958	1



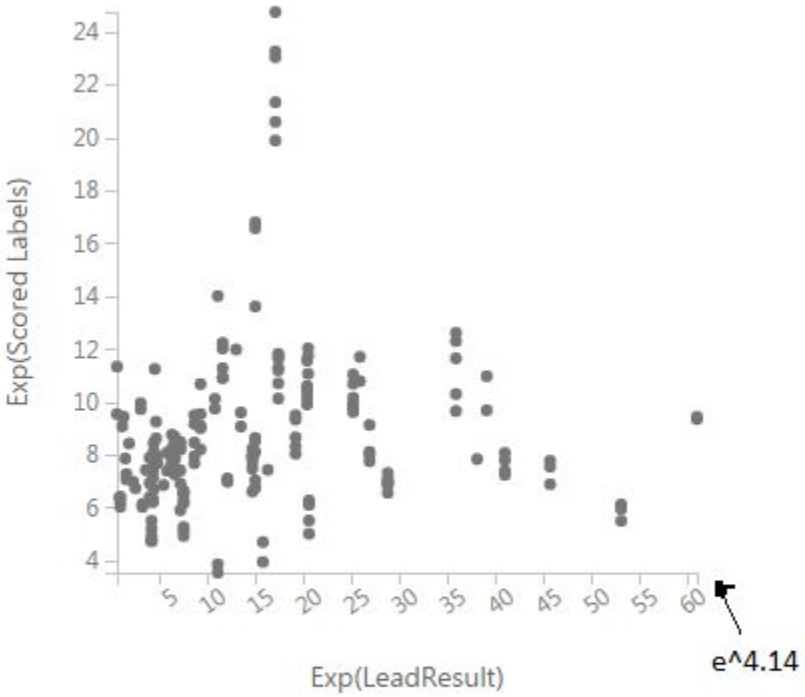
ln(Cu * Fe), ln(Pb)

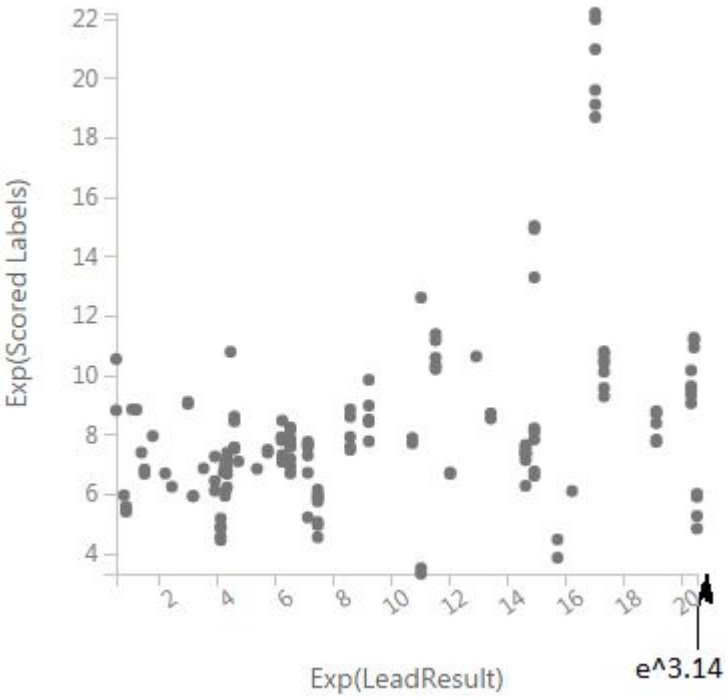
Copper_x_IronResult	LeadResult
-0.045722	0.065356
-0.001323	0.01336
-0.233502	-0.104994
NaN	NaN
0.679831	0.385348
0.6845	0.399933
1	0.593423
0.593423	1



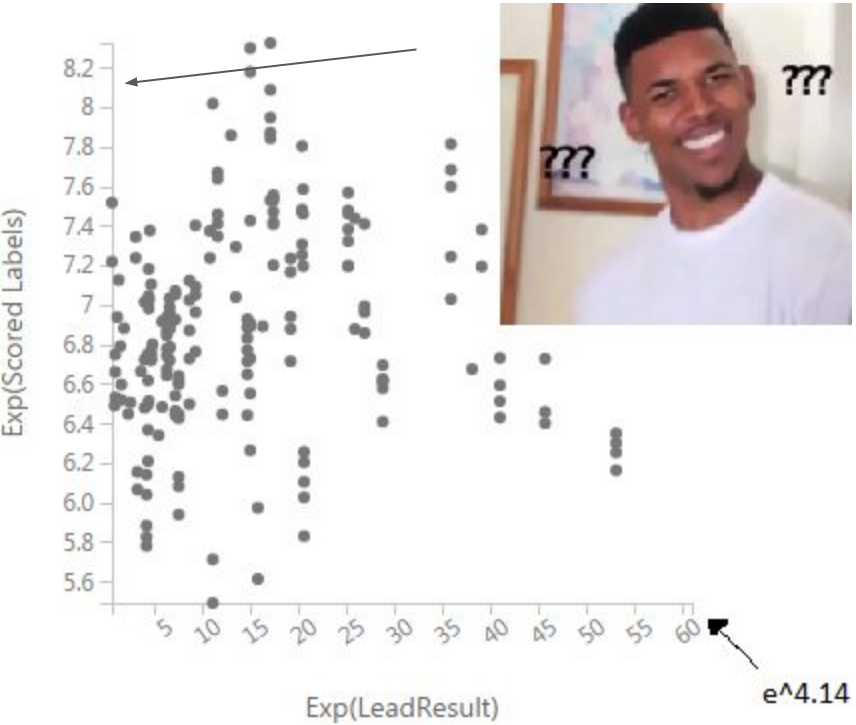
Train: Pb in [0, e^{π}]

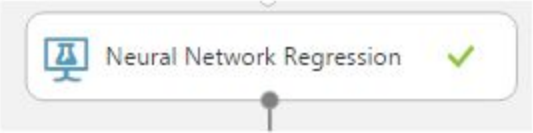
Linear regression (least-squares)



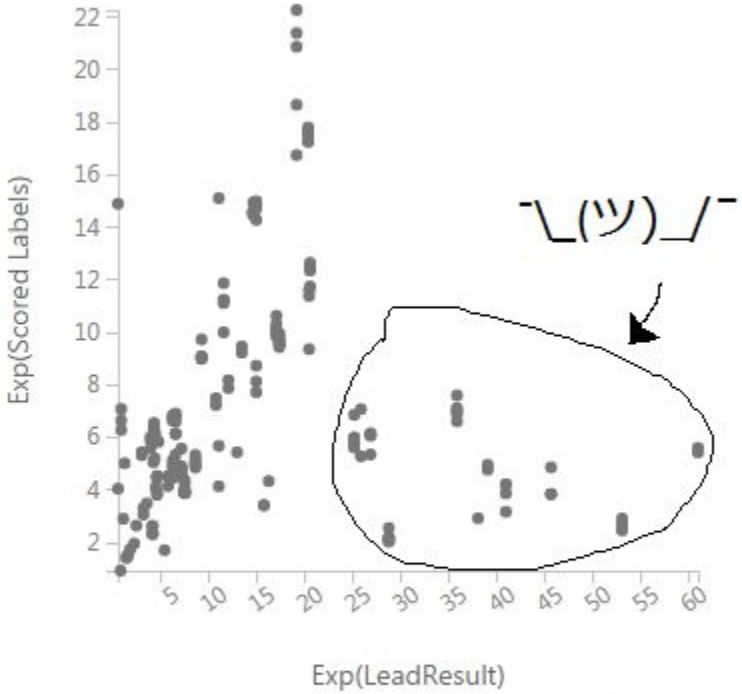


Neural networks

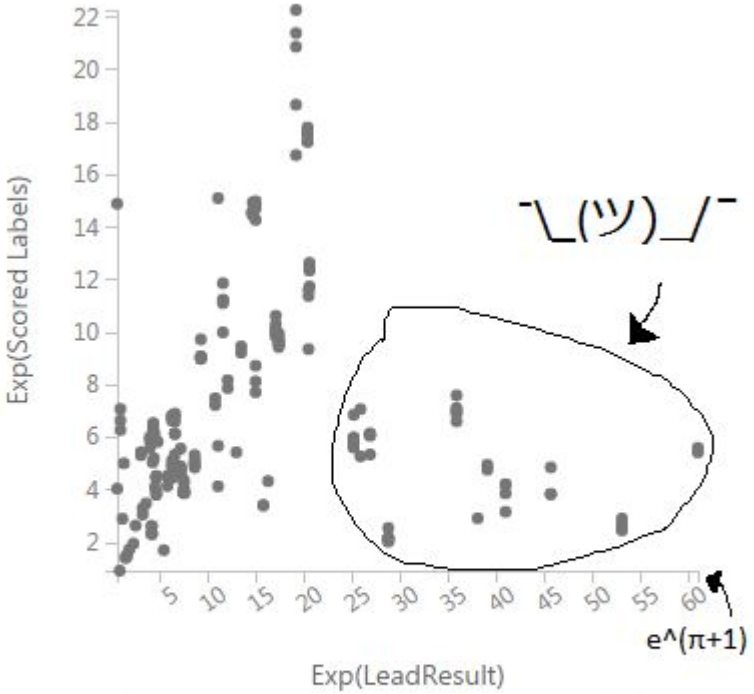


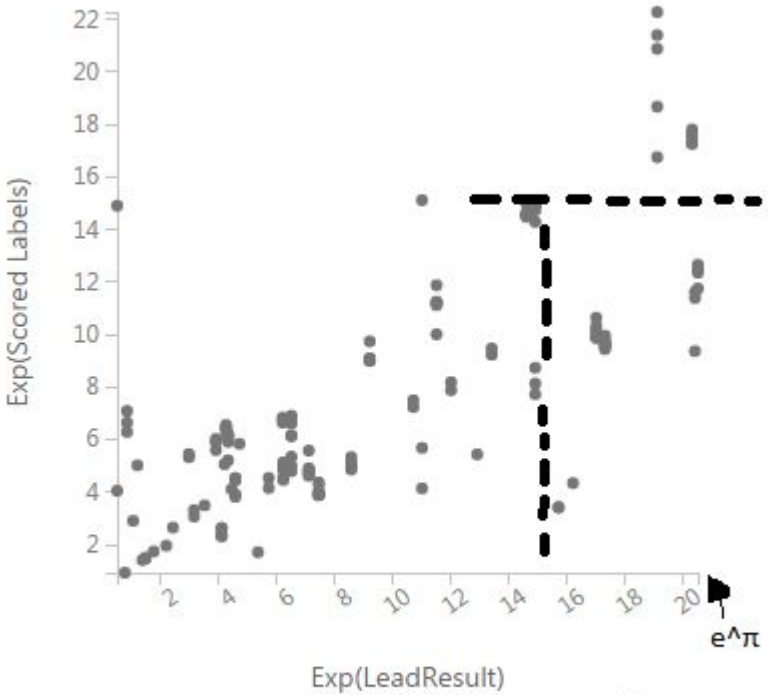


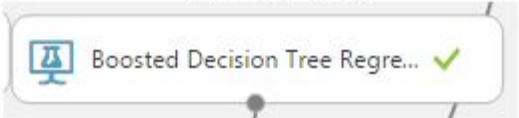
Boosted decision tree regression



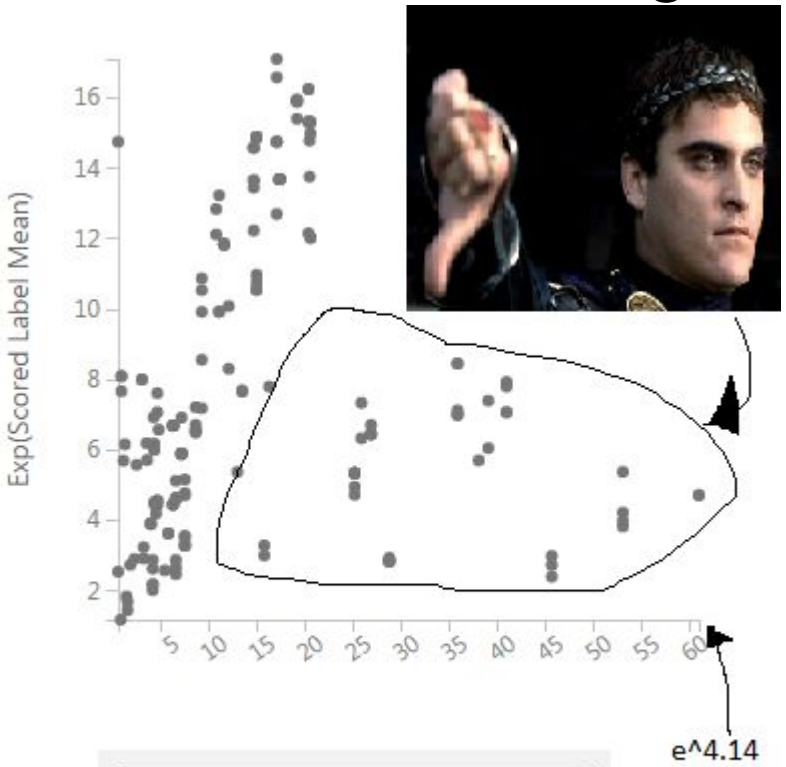
Boosted decision tree regression

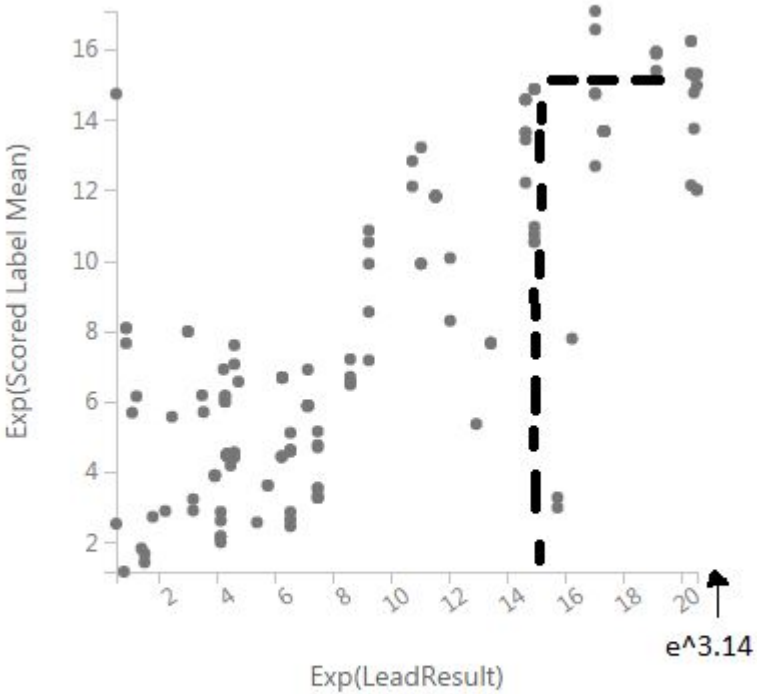


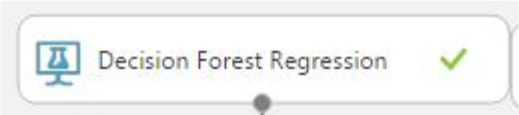


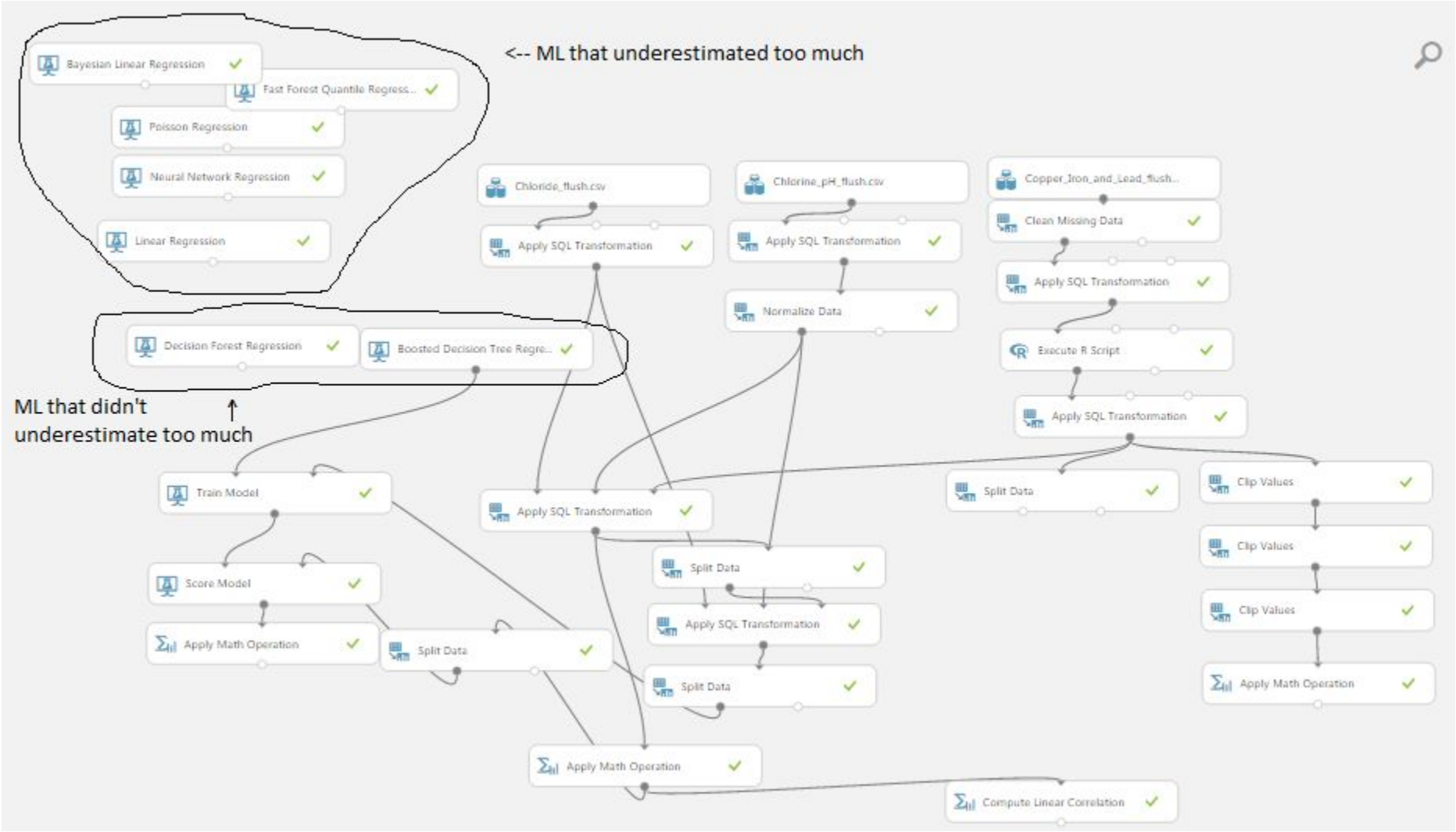


Decision forest regression








<-- ML that underestimated too much
Bayesian Linear Regression
Fast Forest Quantile Regress...
Poisson Regression
Neural Network Regression
Linear Regression
Decision Forest Regression
Boosted Decision Tree Regre...
ML that didn't underestimate too much
Chloride_flush.csv
Chlorine_pH_flush.csv
Copper_Iron_and_Lead_flush...
Clean Missing Data
Apply SQL Transformation
Apply SQL Transformation
Apply SQL Transformation
Normalize Data
Execute R Script
Apply SQL Transformation
Train Model
Apply SQL Transformation
Split Data
Clip Values
Clip Values
Clip Values
Score Model
Split Data
Apply SQL Transformation
Apply Math Operation
Split Data
Split Data
Apply Math Operation
Apply Math Operation
Compute Linear Correlation

Things that would have helped

- Precise location data
- City planning data
- Geotagged tweets

Moral of the story

- Bayesian linear regression
- Neural network linear regression
- Regular linear regression
- Fast forest quantile regression
- Poisson regression

- Decision forest regression
- Boosted decision tree regression
- Not trying to predict out of the sample range
- ln(x)