QSPR MODEL
Predict Compound Toxicity

Cheminformatics Project

Shayan Hatami

pLC50

The negative logarithm of the lethal concentration 50% for Pimephales promelas

Classify chemicals based on their toxicity and assess their environmental impact

$$pLC50=-\log10(LC50)$$

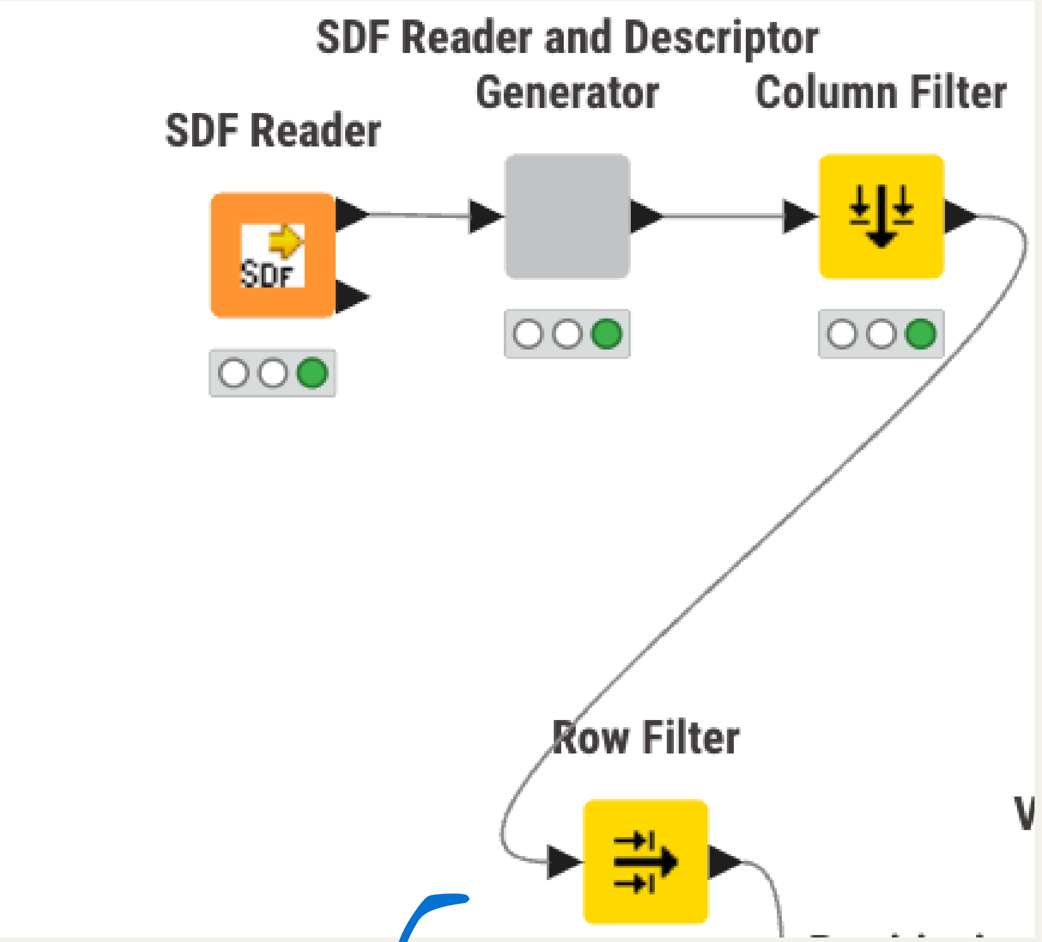
LC50 is expressed in moles per liter (M).

Pimephales promelas is a freshwater species sensitive to pollutants

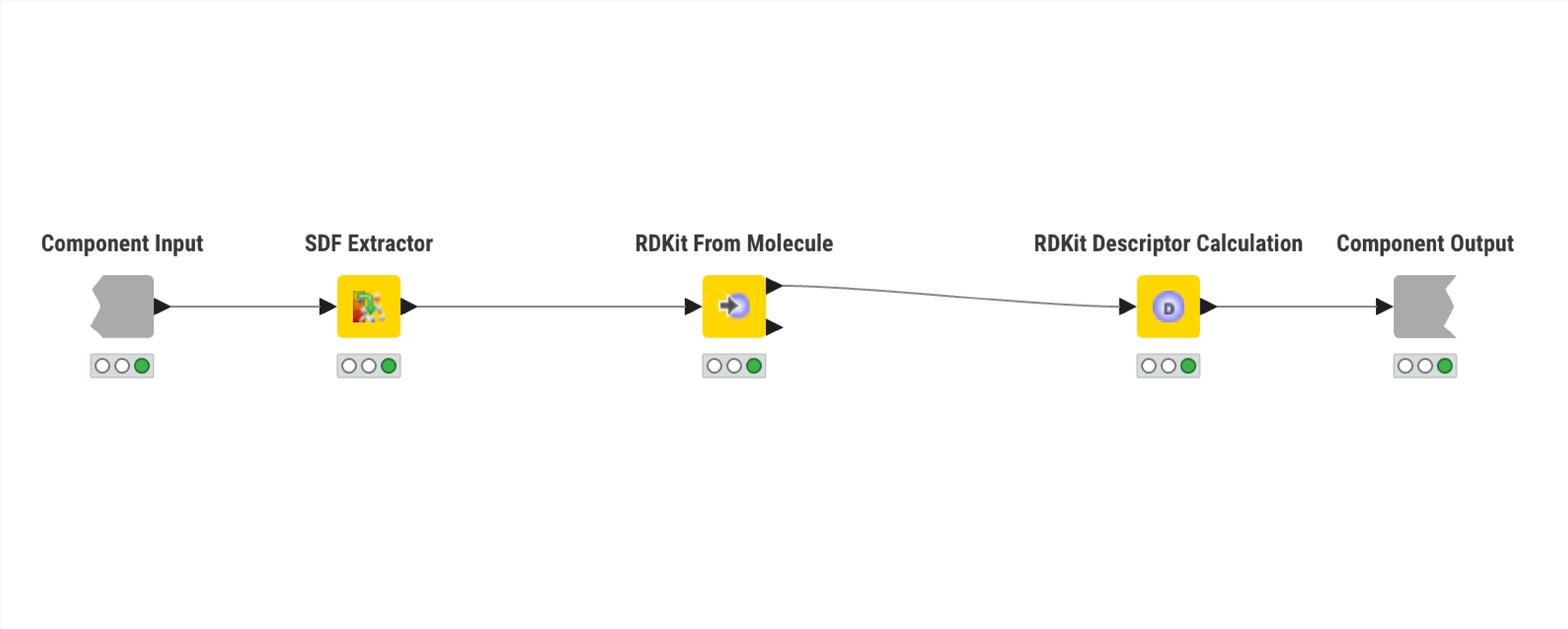
Reading The Dataset

The dataset is in SDF (Structure Data File) format

375 chemical compounds annotated with experimental pLC50 values

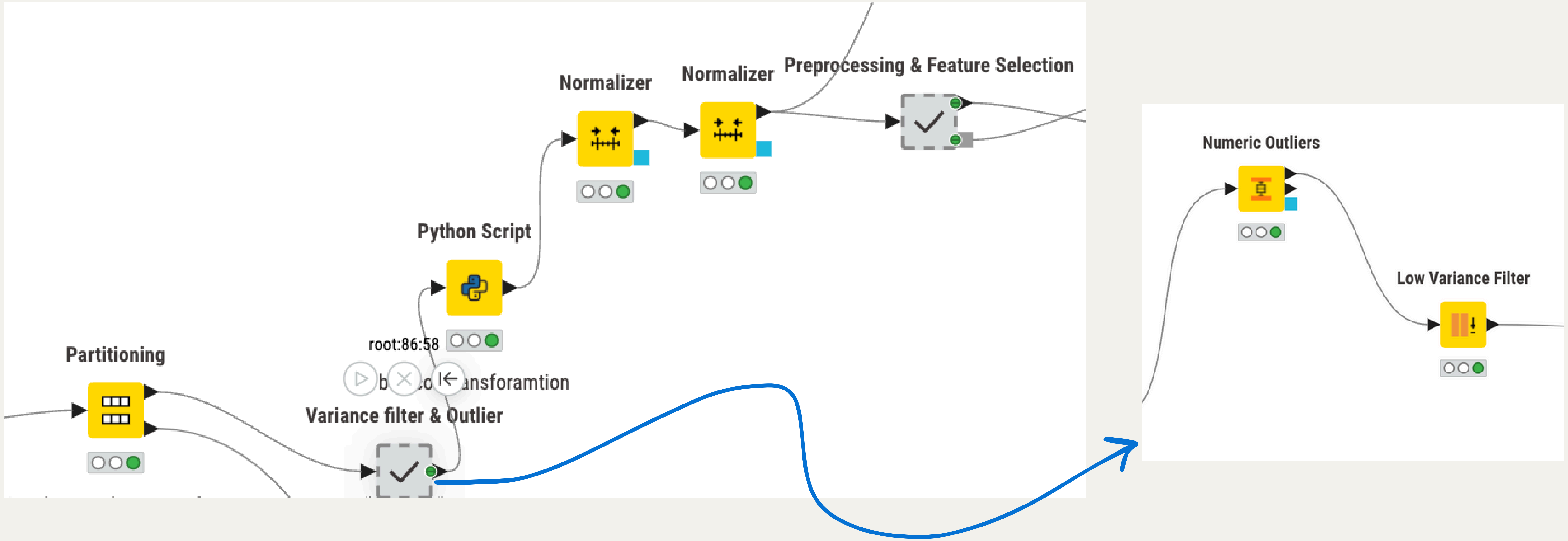


Target column outlier



Preprocessing

Closest permitted value
Box cox transformation
Z score, min max scaler



Feature Selection

Highly correlated features, More than 0.9

Backward selection

Scoring metric -> R²

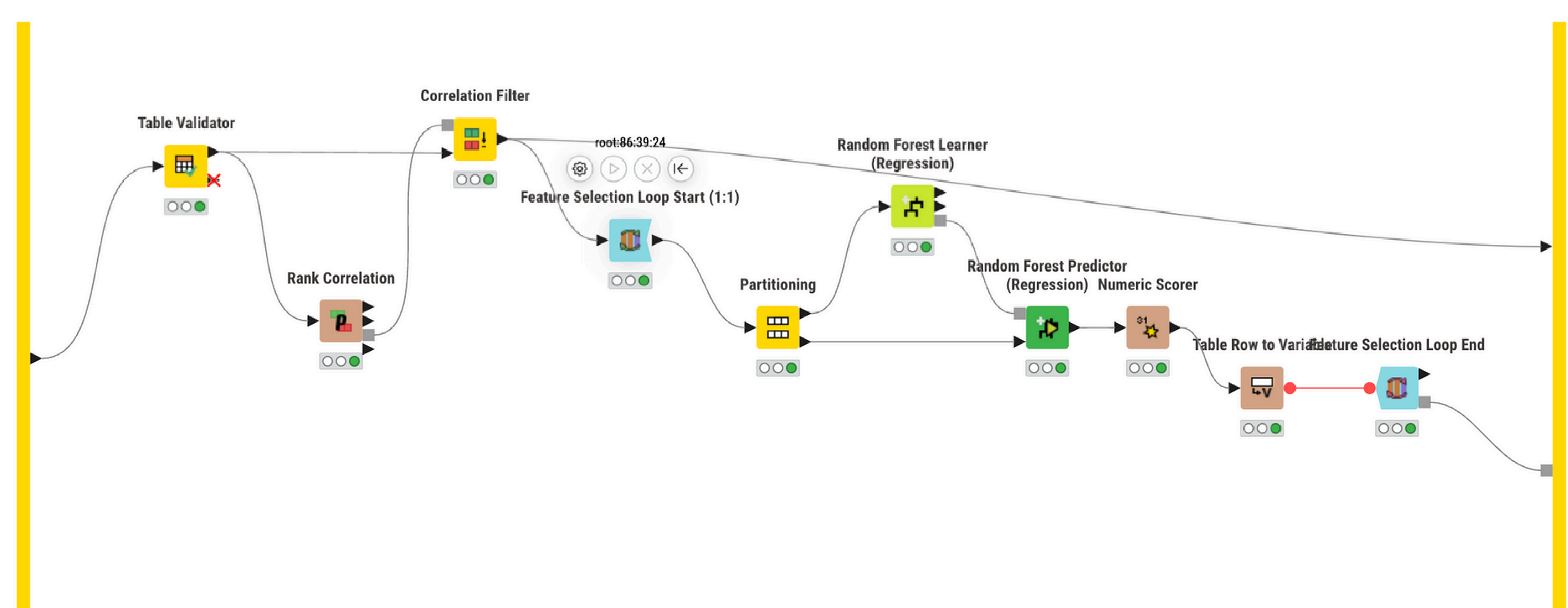
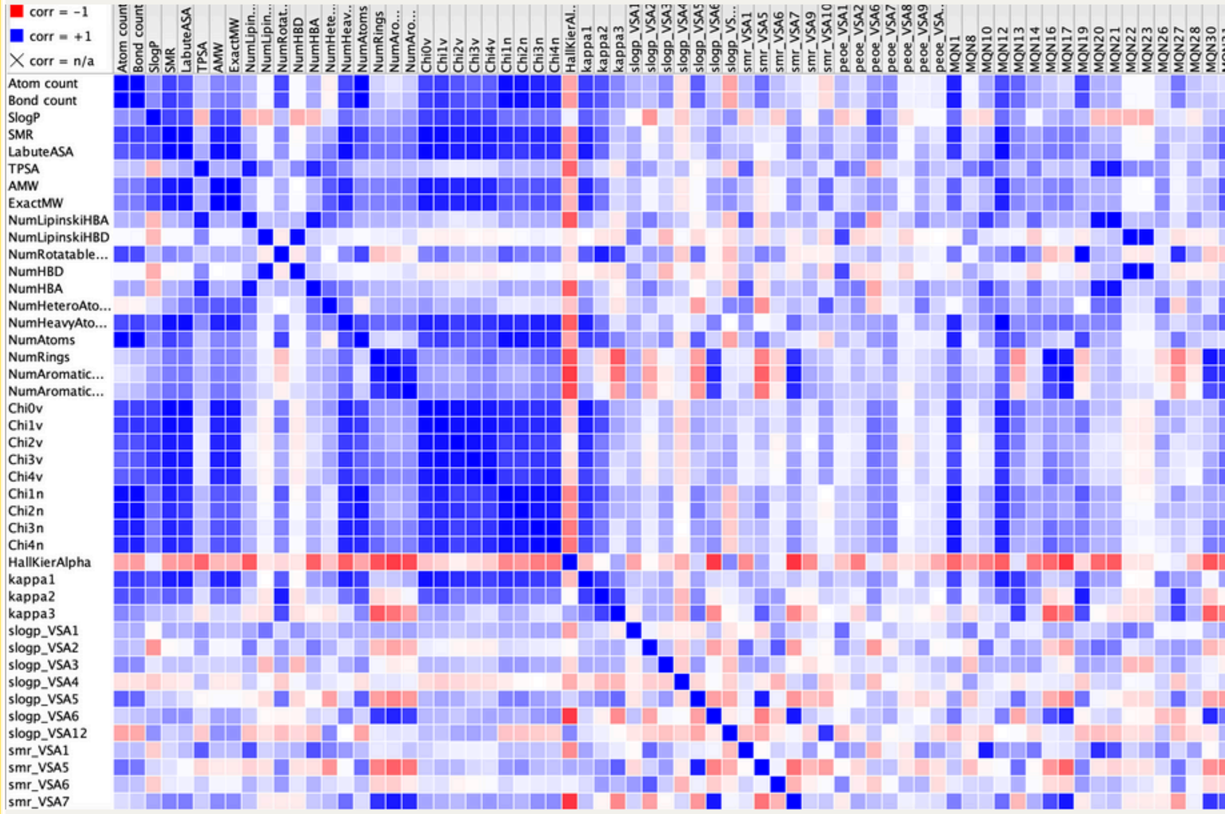


Table Validator
Rank Correlation
Correlation Filter
root:86:39:24
Feature Selection Loop Start (1:1)
Partitioning
Random Forest Learner (Regression)
Random Forest Predictor (Regression)
Numeric Scorer
Table Row to Variable
Feature Selection Loop End

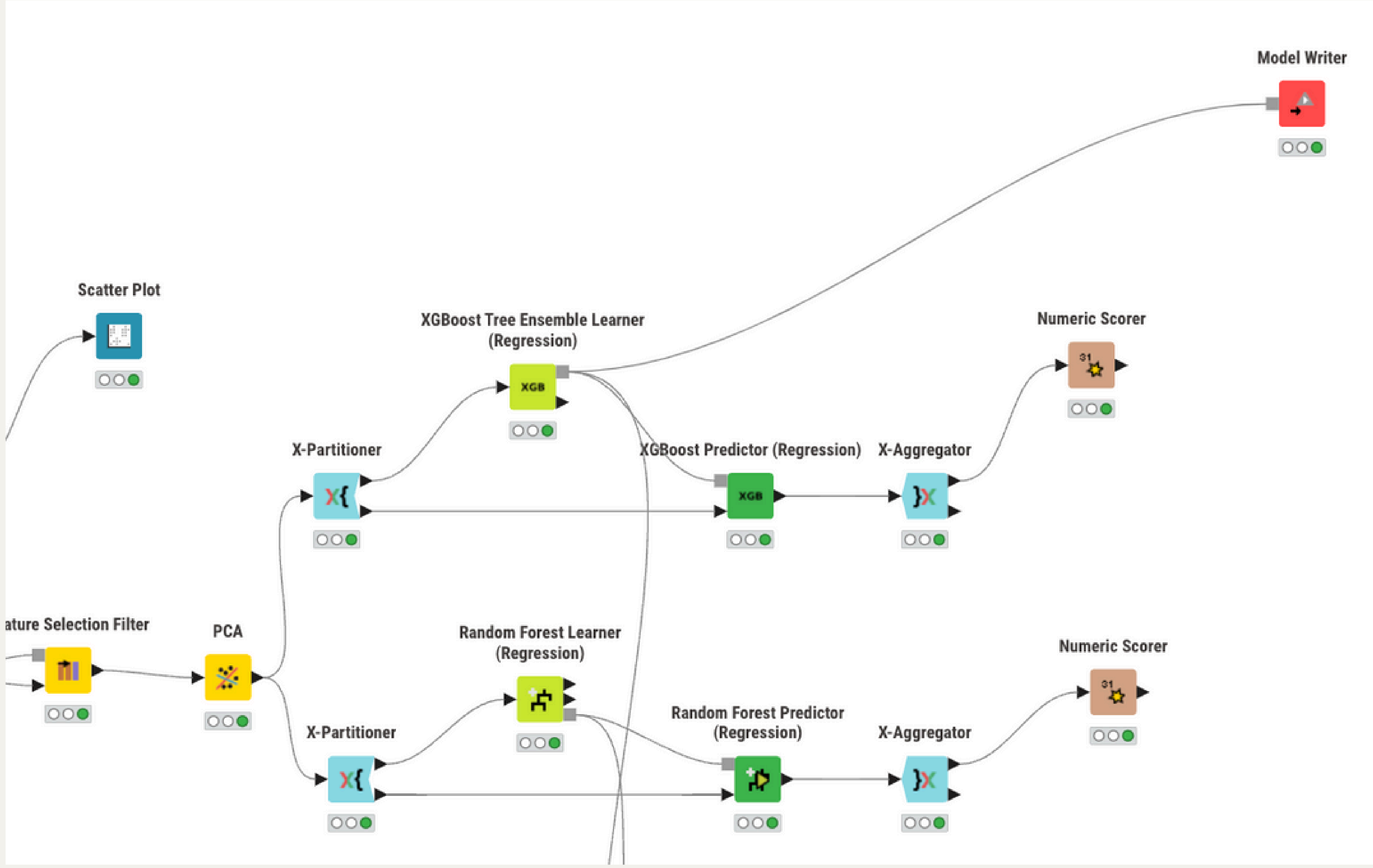


Training

XGboost, Random forest

5-Fold Cross Validation

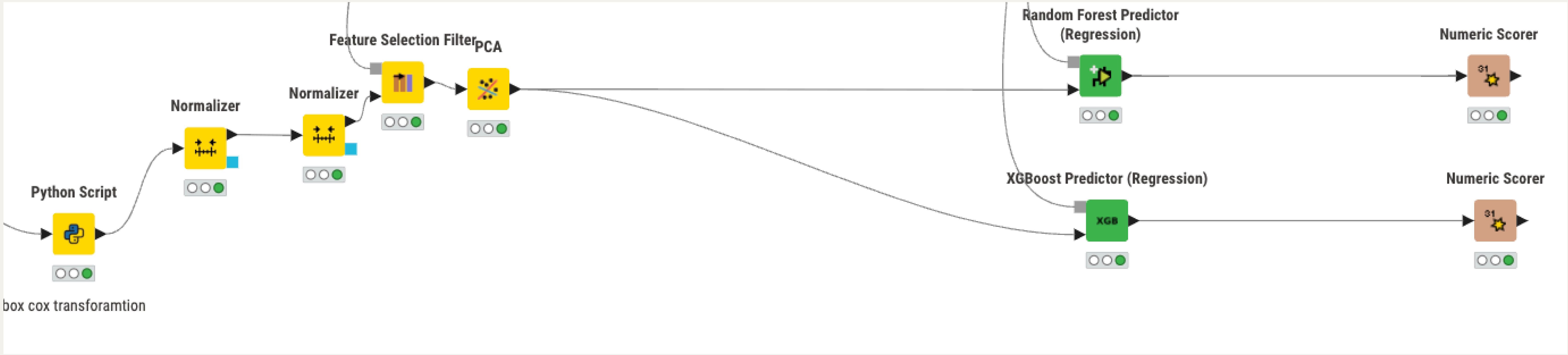
Added PCA


Model Writer
Scatter Plot
XGBoost Tree Ensemble Learner (Regression)
XGB
Numeric Scorer
X-Partitioner
XGBoost Predictor (Regression)
XGB
X-Aggregator
ature Selection Filter
PCA
Random Forest Learner (Regression)
Numeric Scorer
Random Forest Predictor (Regression)
X-Partitioner
X-Aggregator

Testing

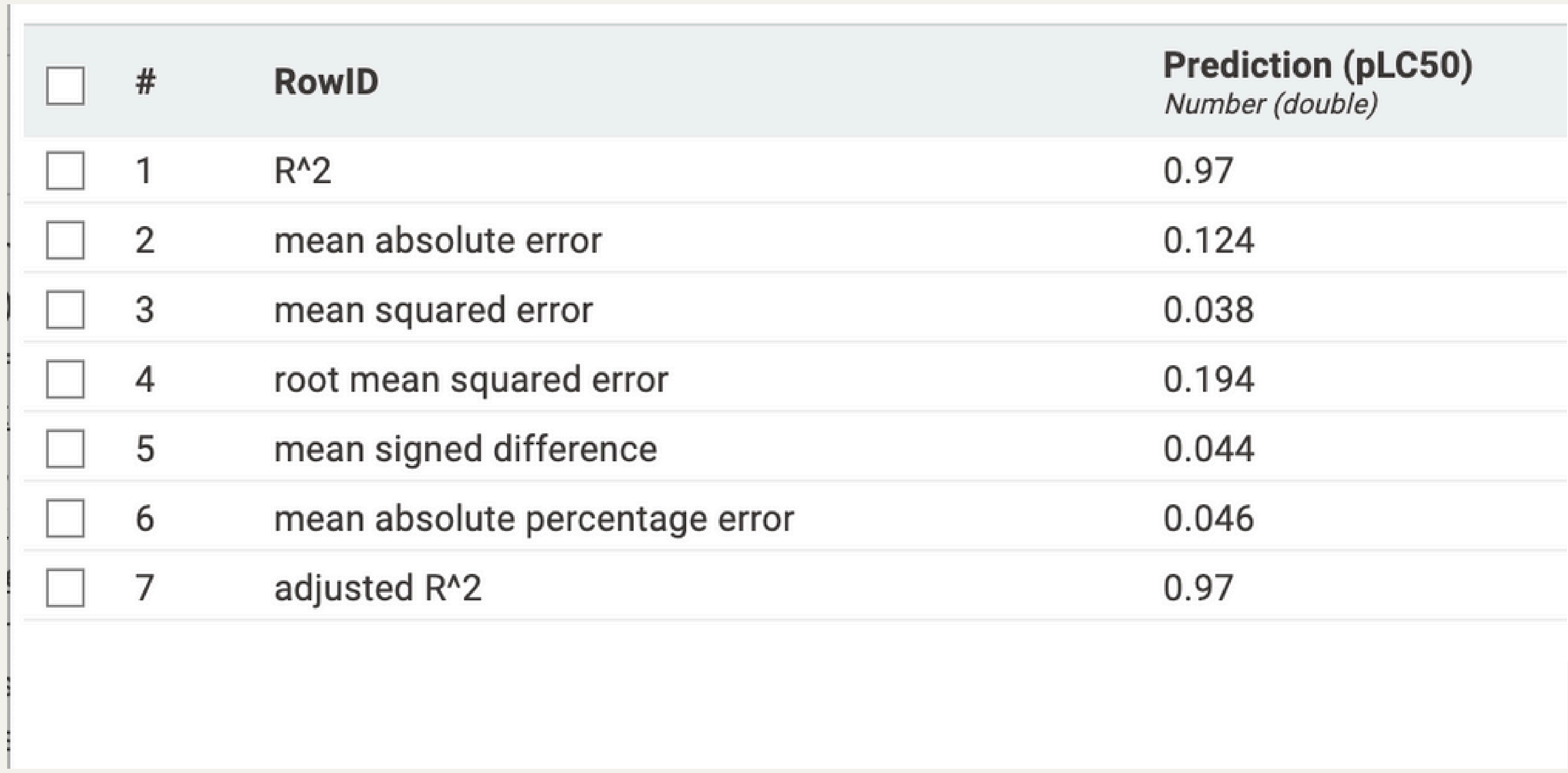
Same preprocessing as Training data

XGBoost had a higher score



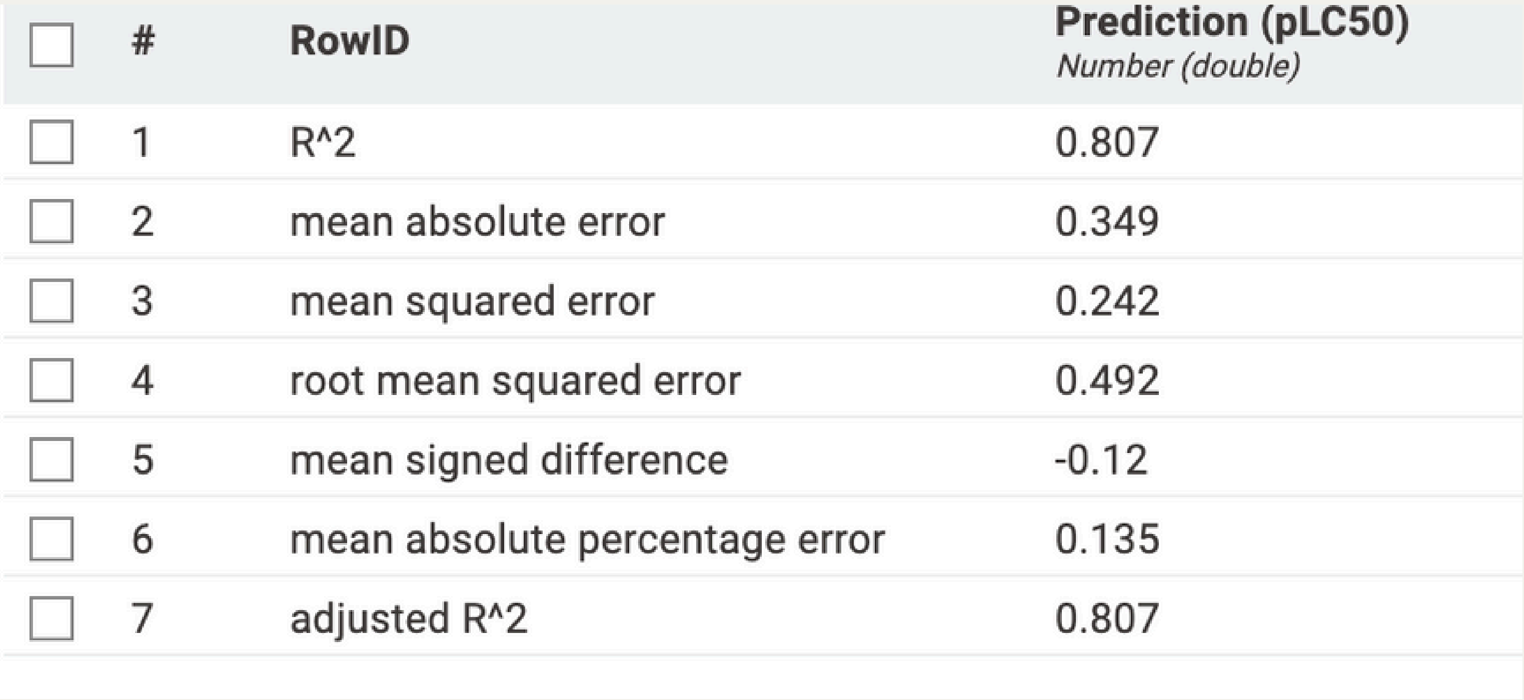
Results

XGboost



□	#	RowID	Prediction (pLC50) *Number (double)*
□	1	R^2	0.97
□	2	mean absolute error	0.124
□	3	mean squared error	0.038
□	4	root mean squared error	0.194
□	5	mean signed difference	0.044
□	6	mean absolute percentage error	0.046
□	7	adjusted R^2	0.97

Random forest



□	#	RowID	Prediction (pLC50) *Number (double)*
□	1	R^2	0.807
□	2	mean absolute error	0.349
□	3	mean squared error	0.242
□	4	root mean squared error	0.492
□	5	mean signed difference	-0.12
□	6	mean absolute percentage error	0.135
□	7	adjusted R^2	0.807