Pitfalls Analyzer: Automatically Detecting Data Science Pitfalls in KNIME using KNIME

Gopi Krishnan Rajbahadur

Software Analysis & Intelligence Lab

Gustavo Ansaldi Oliva

Software Analysis & Intelligence Lab

Ahmed E. Hassan

Software Analysis & Intelligence Lab

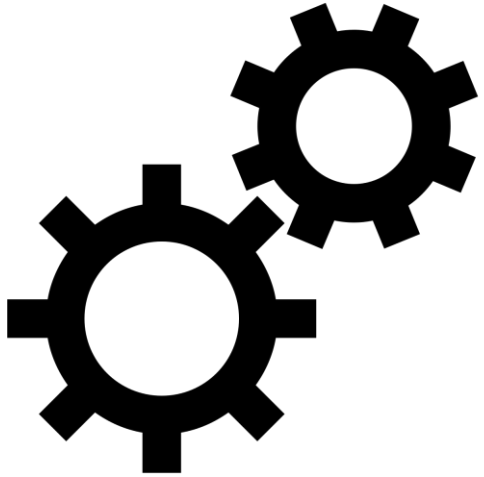
Juergen Dingel

MASE

Queen's University

There are advantages to using data science pipelines

Abstraction



Easy Automation

Fast Turnaround

Documentation

Easy access enables easy mistakes

"Simplistic studies comparing data intensive methods with linear regression will be scientifically valueless, if the regression techniques are applied incorrectly."

-Kitchenham and Mendes [PROMISE'09]

"Many users of such modelling toolkits have limited knowledge about many important details…often leads to major problems which in turn…lead to failure of analytics projects in practice"

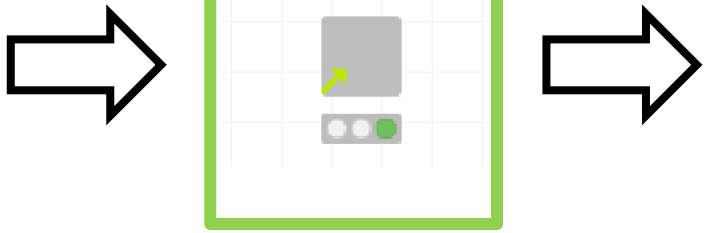
-Tantithamthavorn and Hassan [ICSE-SEIP'18]

Our Goal

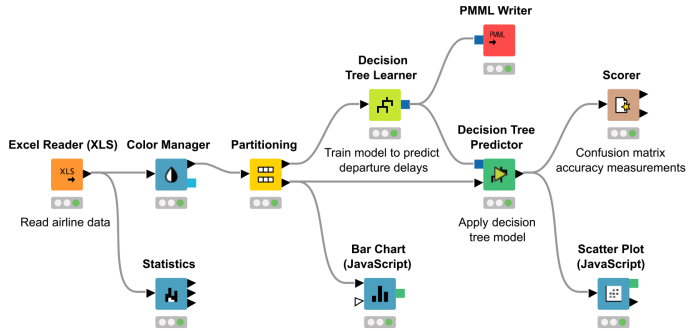
Help practitioners avoid common pitfalls in data science with low overhead

Our Goal

Help practitioners avoid common pitfalls in data science with low overhead



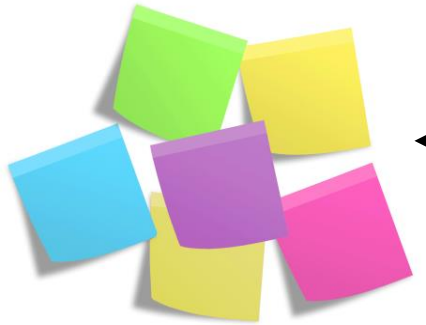
Let’s see how our pitfalls analyzer helps Minion understand exactly what it needs to make the banana plant grow

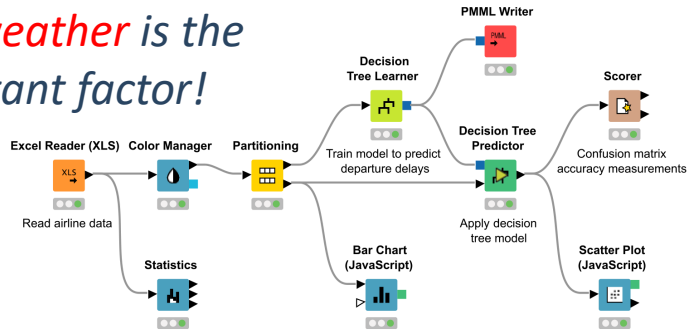


Finds that manure is the most important factor!

Changes

Data order



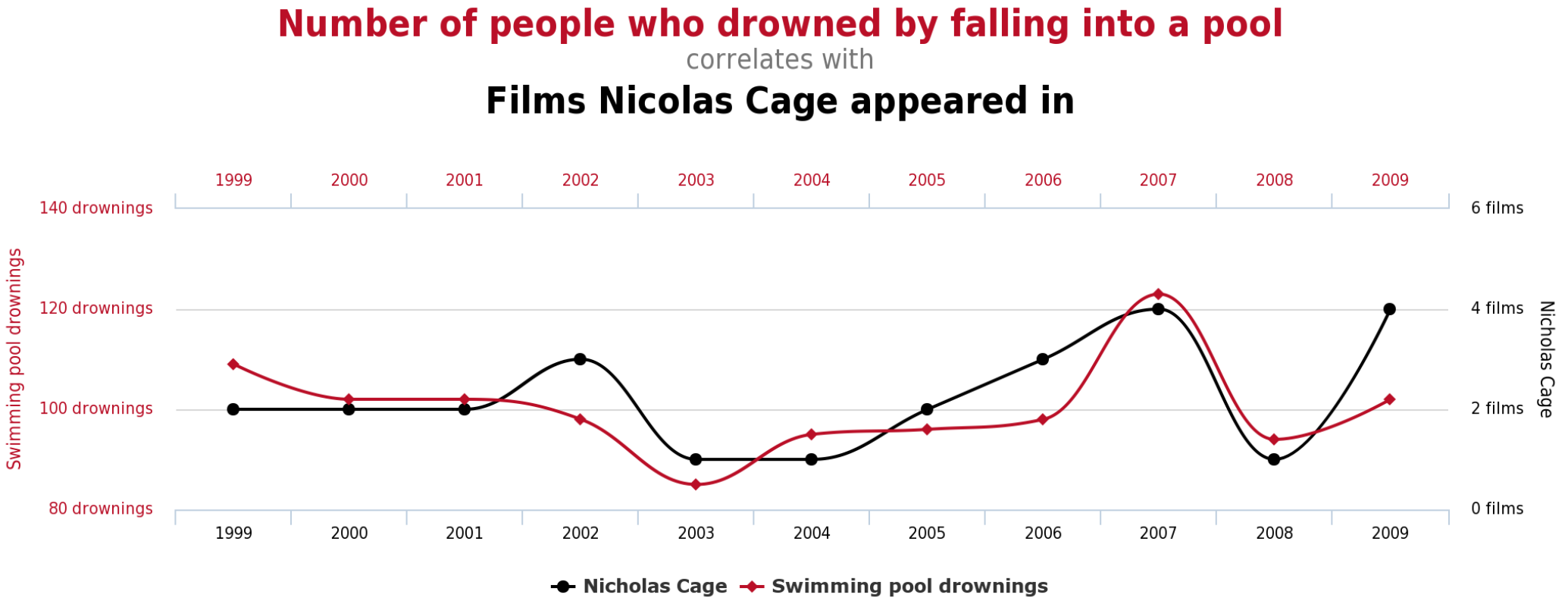


Finds that weather is the most important factor!

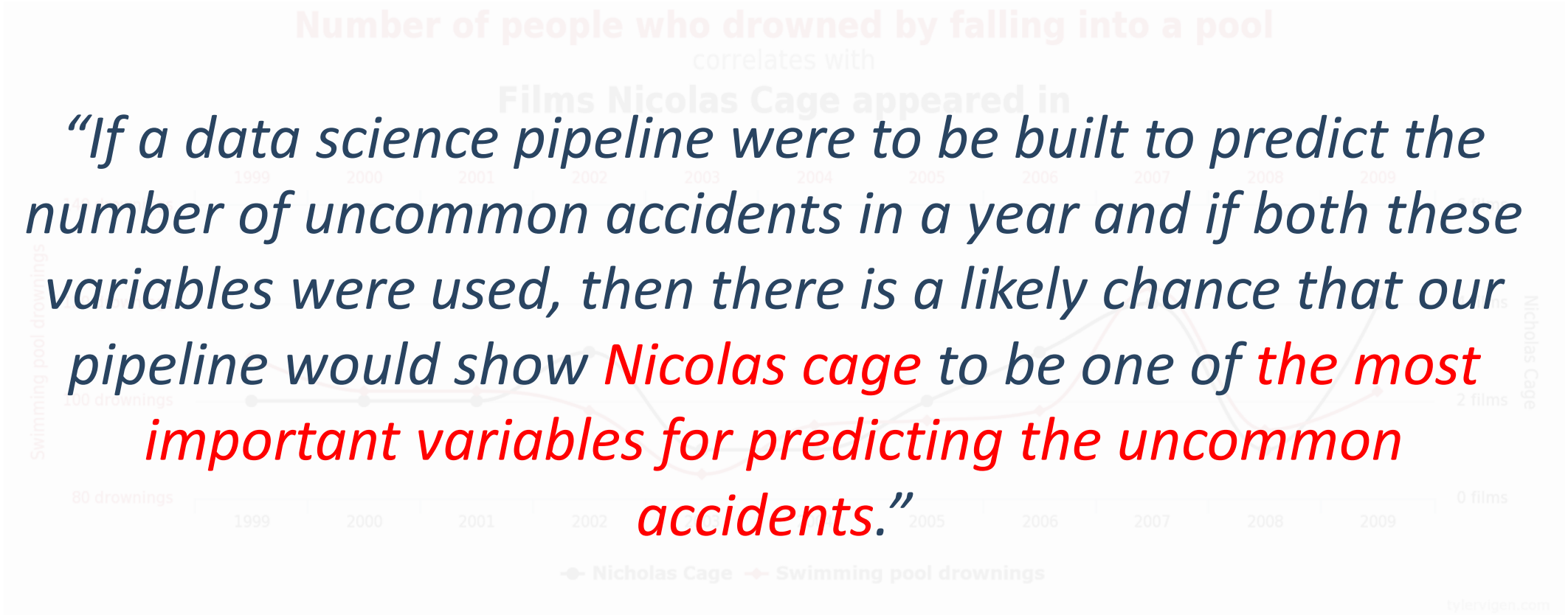
Let’s see how our pitfalls analyzer helps Minion understand exactly what it needs to make the banana plant grow

“Our Minion forgot to remove correlated variables from the dataset before model construction!”

Quick aside: Why is presence of correlated variables bad?

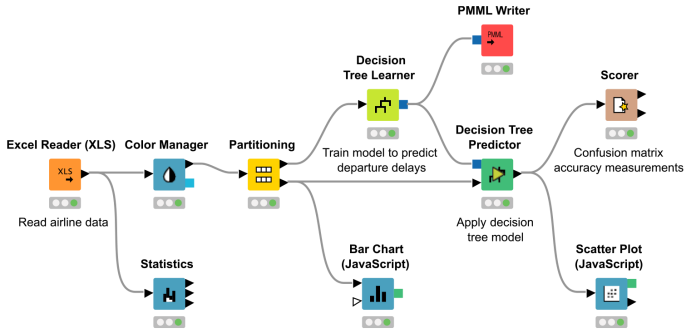


Quick aside: Why is presence of correlated variables bad?

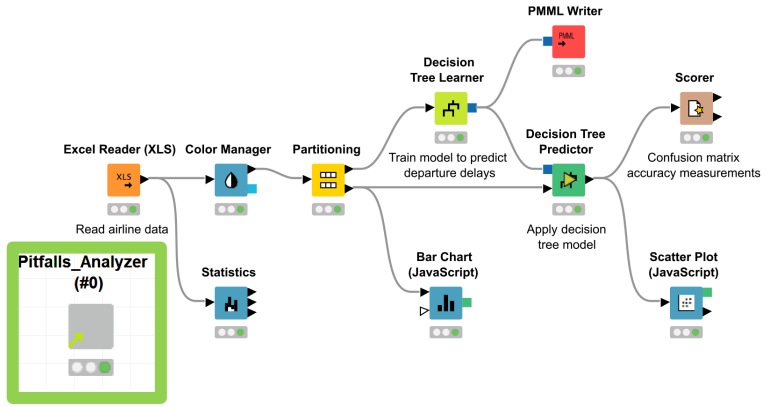


"If a data science pipeline were to be built to predict the number of uncommon accidents in a year and if both these variables were used, then there is a likely chance that our pipeline would show Nicolas cage to be one of the most important variables for predicting the uncommon accidents."

Let’s see how our pitfalls analyzer helps Minion understand exactly what it needs to make the banana plant grow

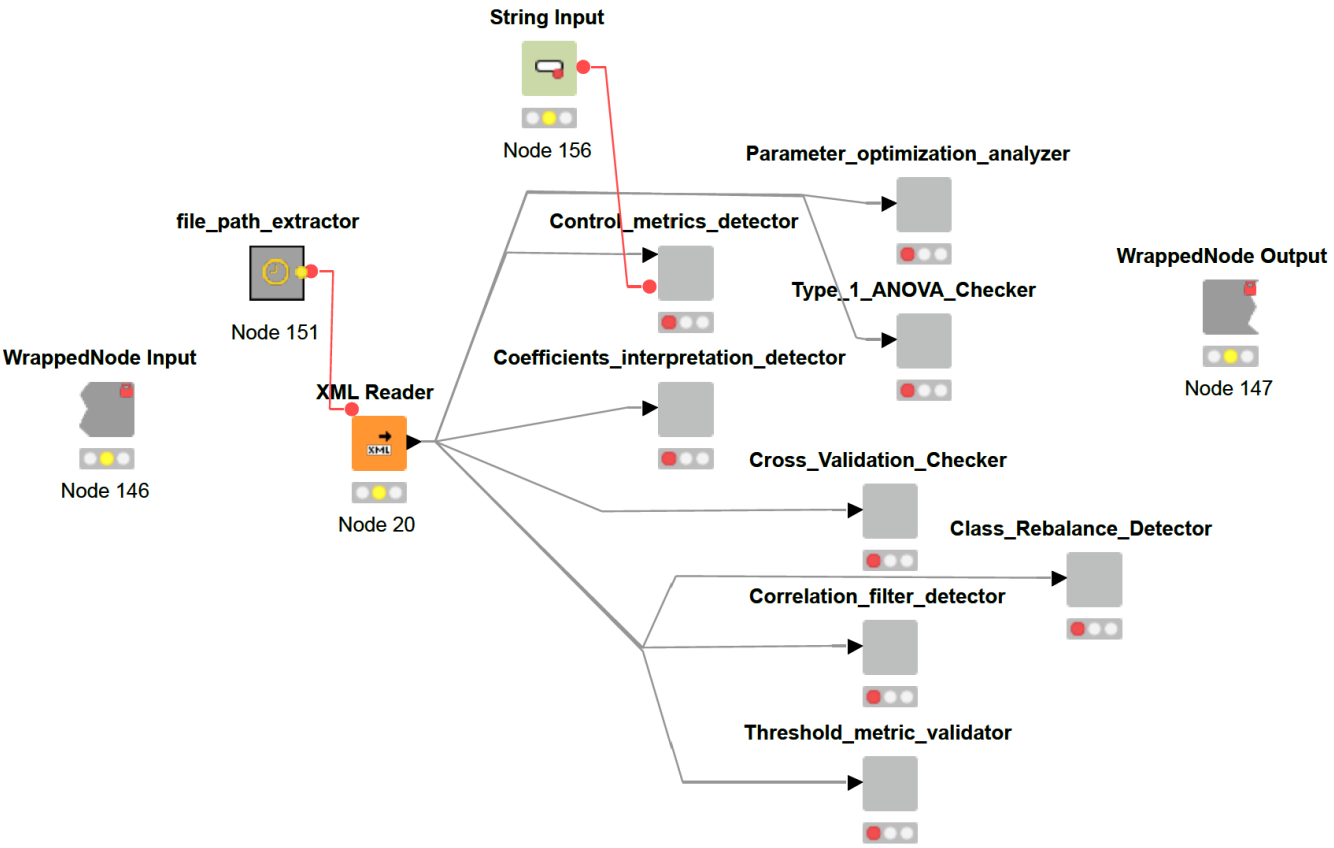


Add our pitfall analyzer component by dragging and dropping it to the current workflow!



[Correlation Analyzer] You do not have a correlation filter. Please be careful when interpreting your features and consider adding correlation filter to your workflow

We enable the identification of 8 common data science pitfalls in a pipeline



Pitfall 1: Absence of control variables

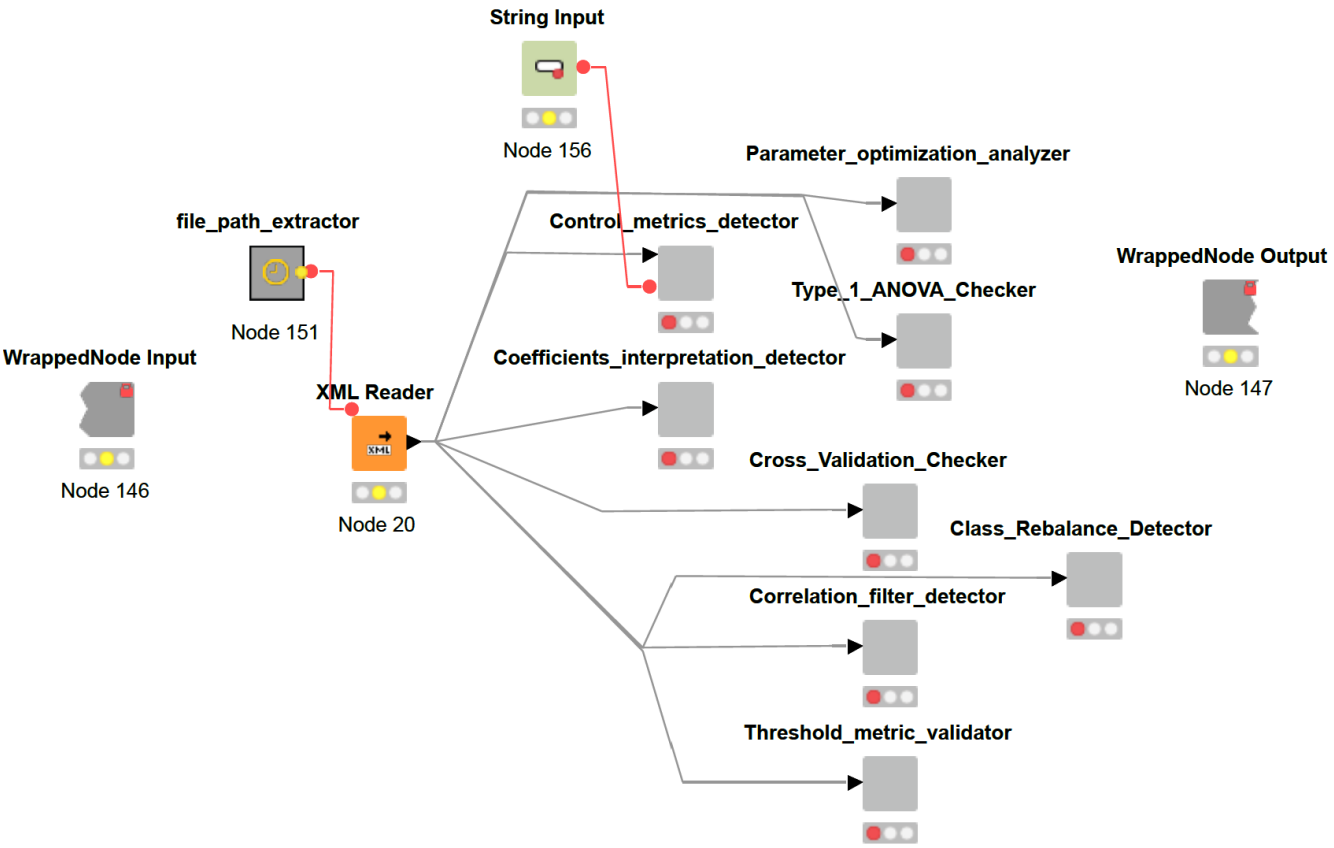
Pitfall 2: Not accounting for the impact of Correlated variables

Pitfall 3: Not accounting for the impact of data rebalancing techniques

Pitfall 4: Not experimenting with different learners or using default settings

Source: Tantithamthavorn and Hassan [ICSE-SEIP'18]

We enable the identification of 8 common data science pitfalls in a pipeline



Pitfall 5: Using threshold dependent performance measures

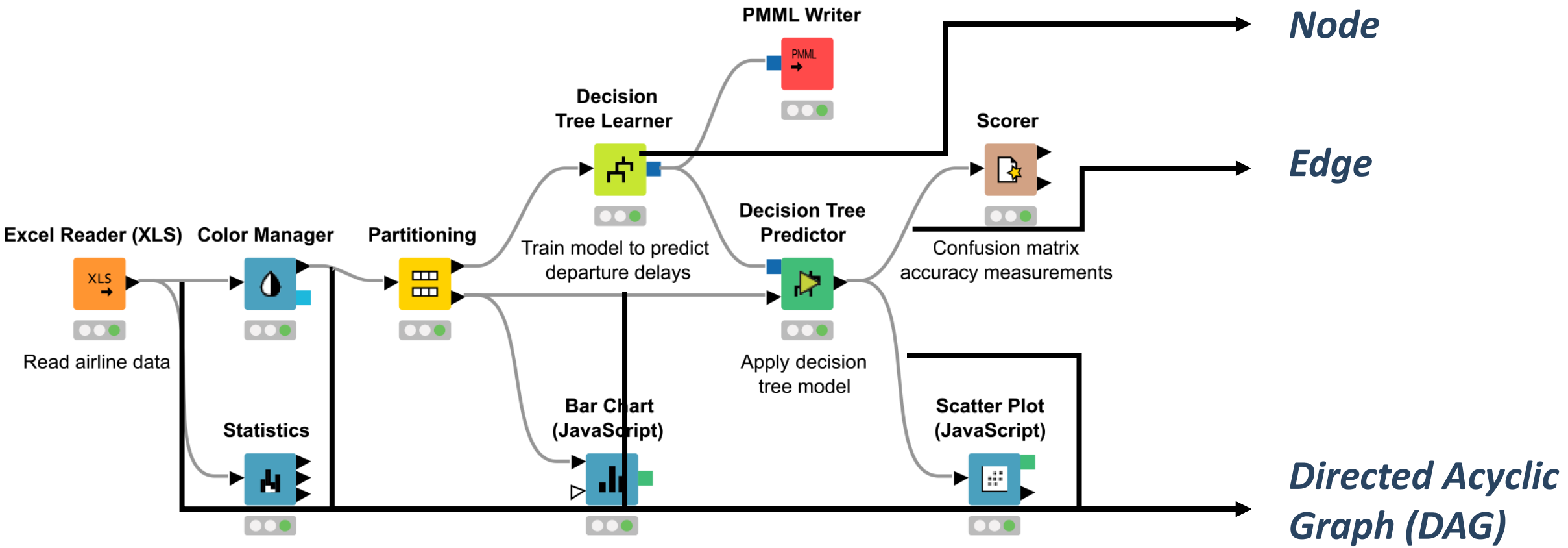
Pitfall 6: Using 10-fold cross validation to estimate the performance

Pitfall 7: Using ANOVA Type-I to interpret the results of a learner

Pitfall 8: Interpreting a regression learner using its coefficients

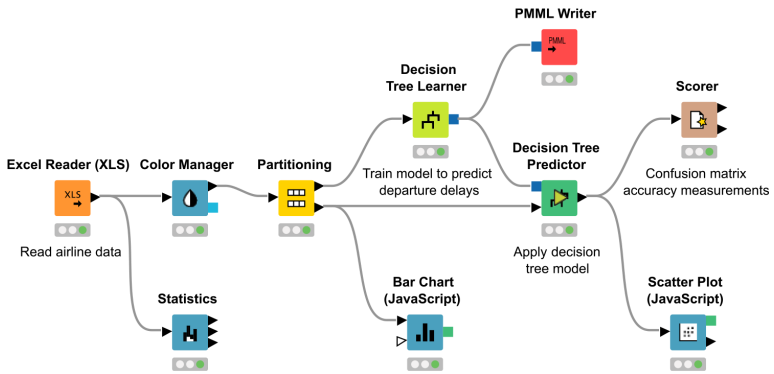
Source: Tantithamthavorn and Hassan [ICSE-SEIP'18]

An Example data science pipeline

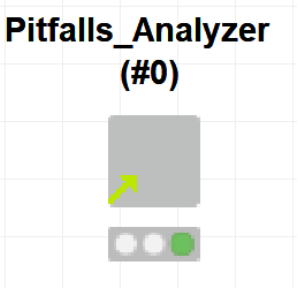


Our Pitfall Analyzer works by detecting data science anti-patterns on the DAG of the target pipeline

Open for Innovation ®
KNIME

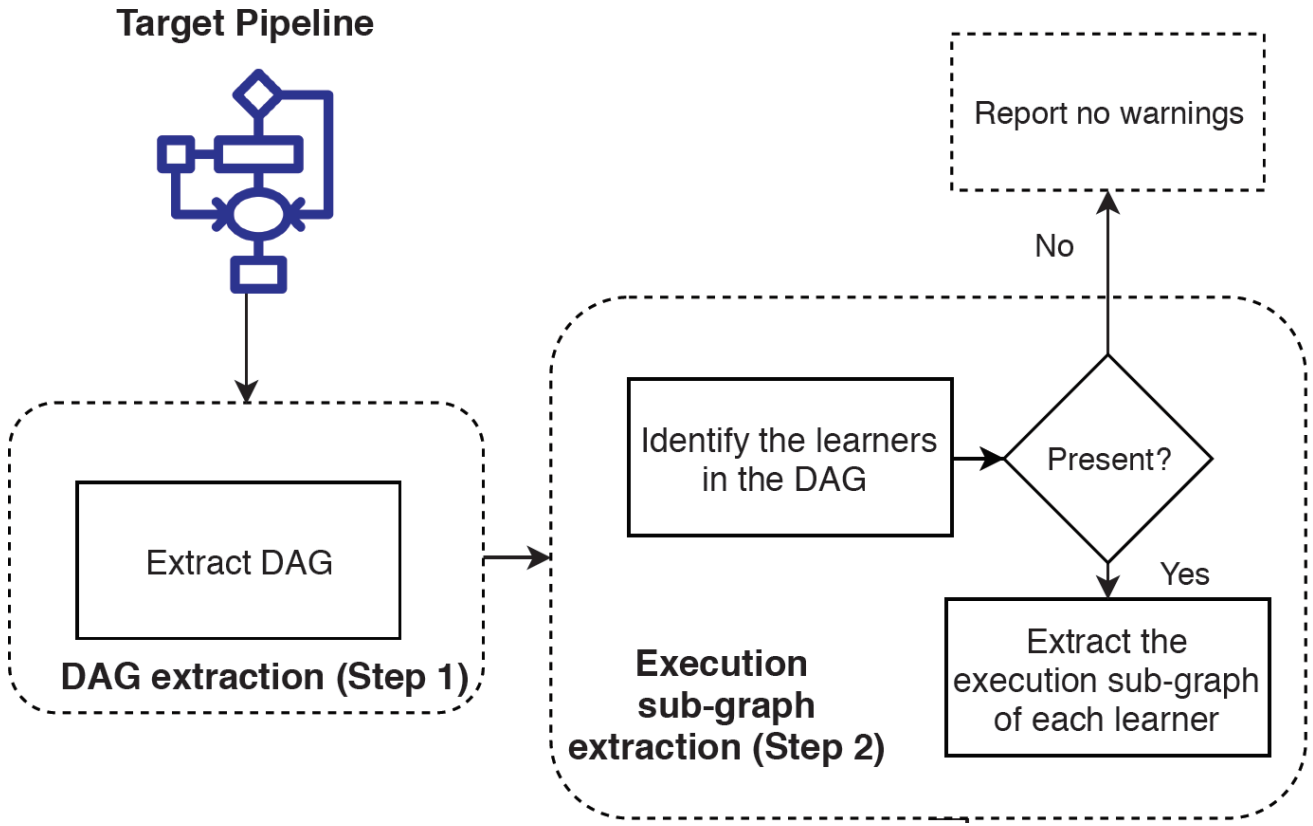


Create pipeline



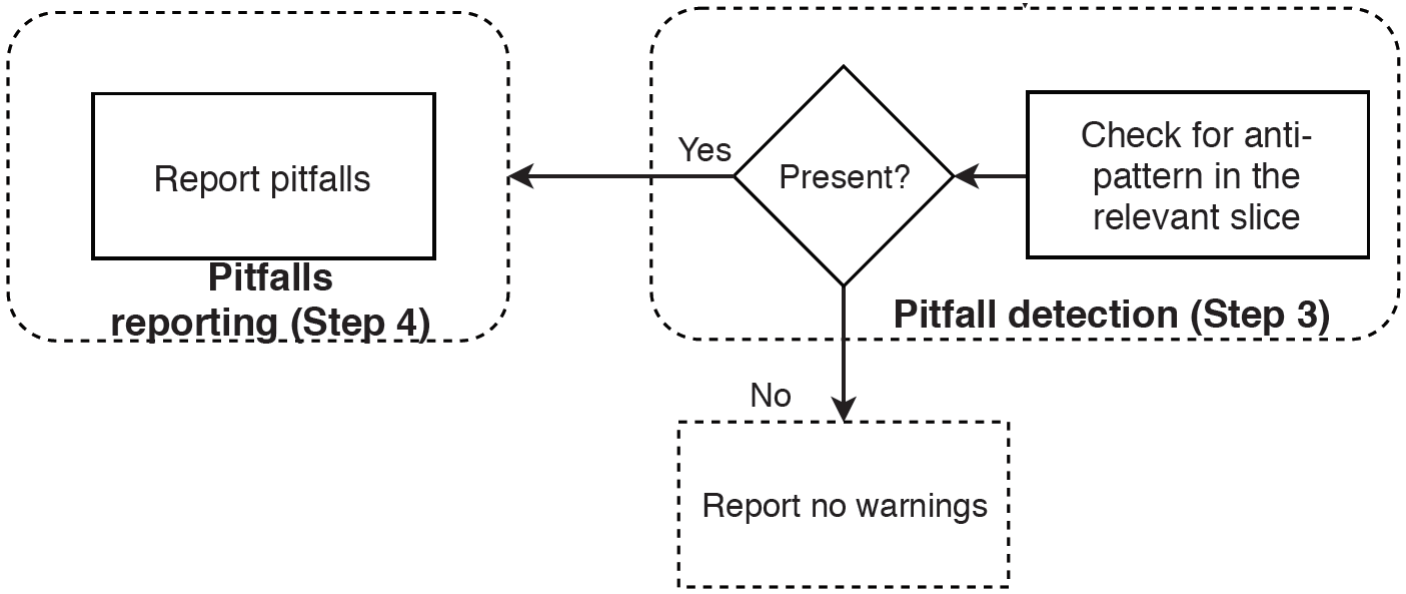
Add our component

DAG and Execution Sub-graph extraction



Extracts the backward and forward execution slice of each learner

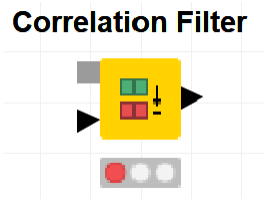
Data science anti-patterns are searched for in a target data science pipeline to identify pitfalls



Checks for the anti-pattern nodes in the relevant slice (either the forward or backward slice or both) of the execution sub-graph of the learner node

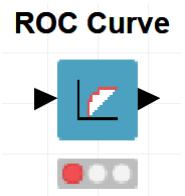
Example data science anti-patterns that are searched for in the pipeline

(Pitfall 2)



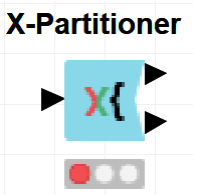
Absence of Correlated variables removal node

(Pitfall 5)



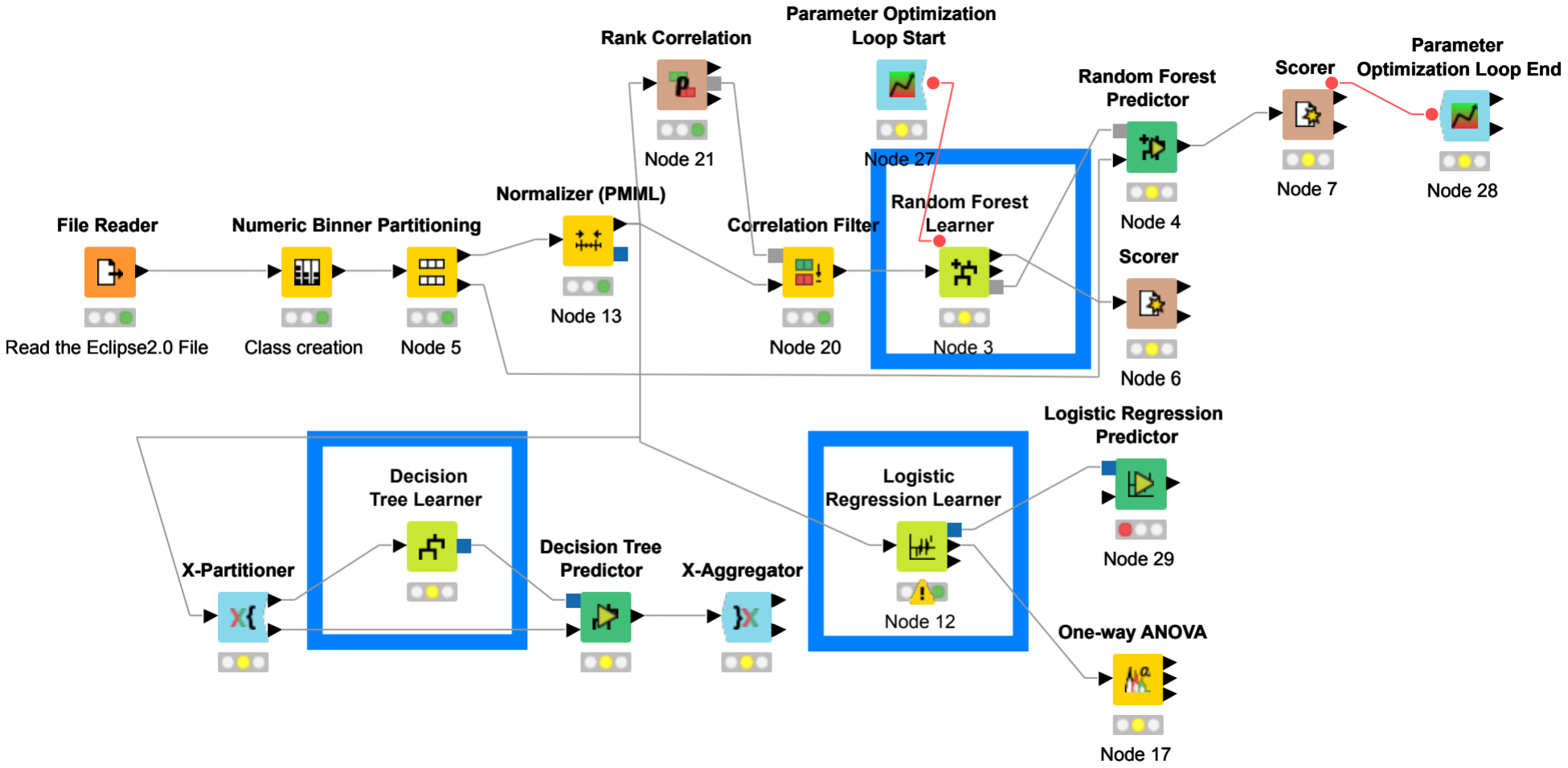
Threshold independent metrics being present and absence of threshold independent metrics

(Pitfall 6)

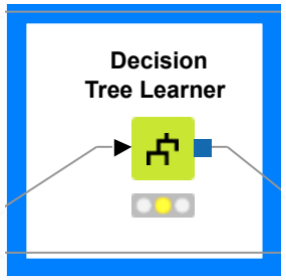


Presence of cross-validation nodes and the absence of bootstrap nodes

A quick example of Pitfall 2 (non removal of correlated variables) identification with our pitfalls analyzer approach

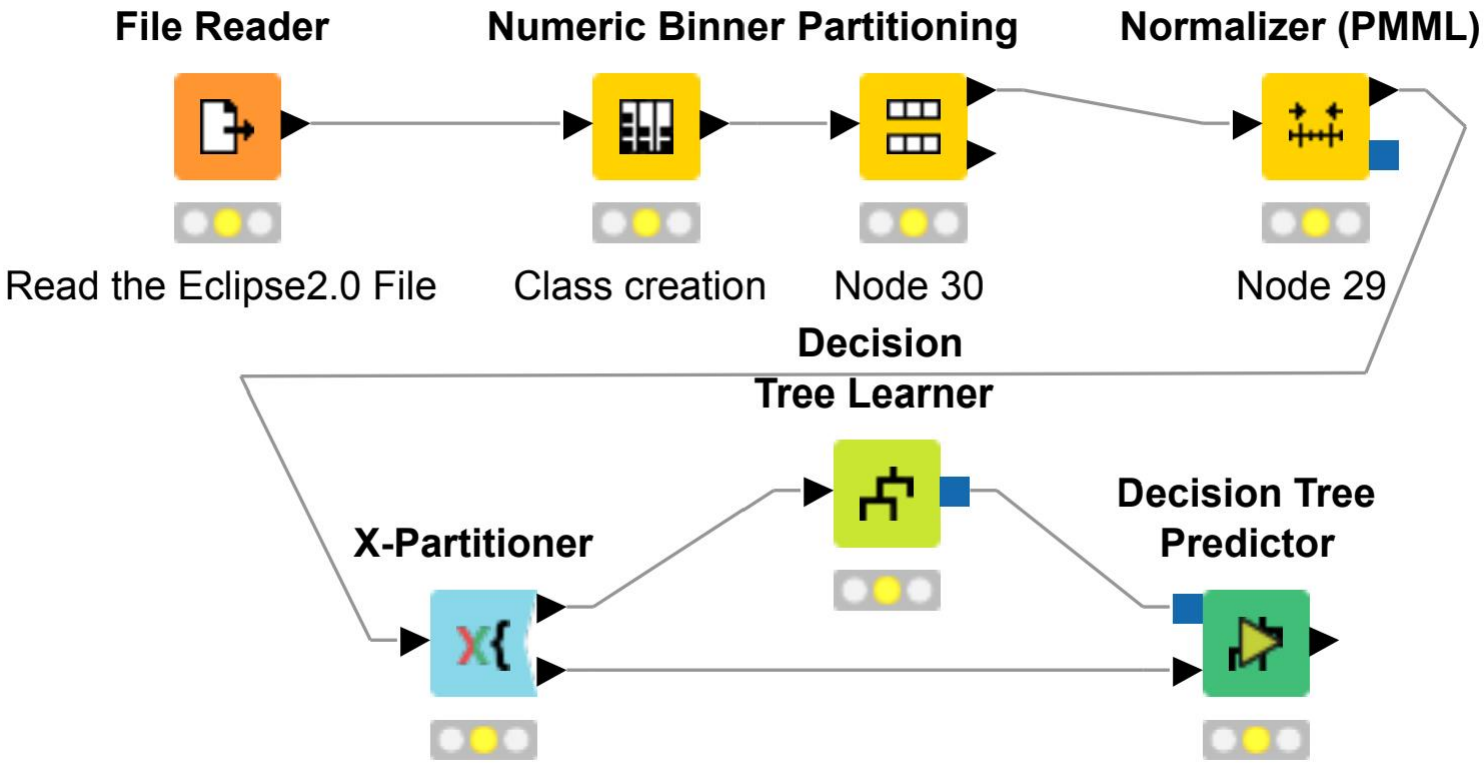


A quick example of Pitfall 2 (non removal of correlated variables) identification with our pitfalls analyzer approach



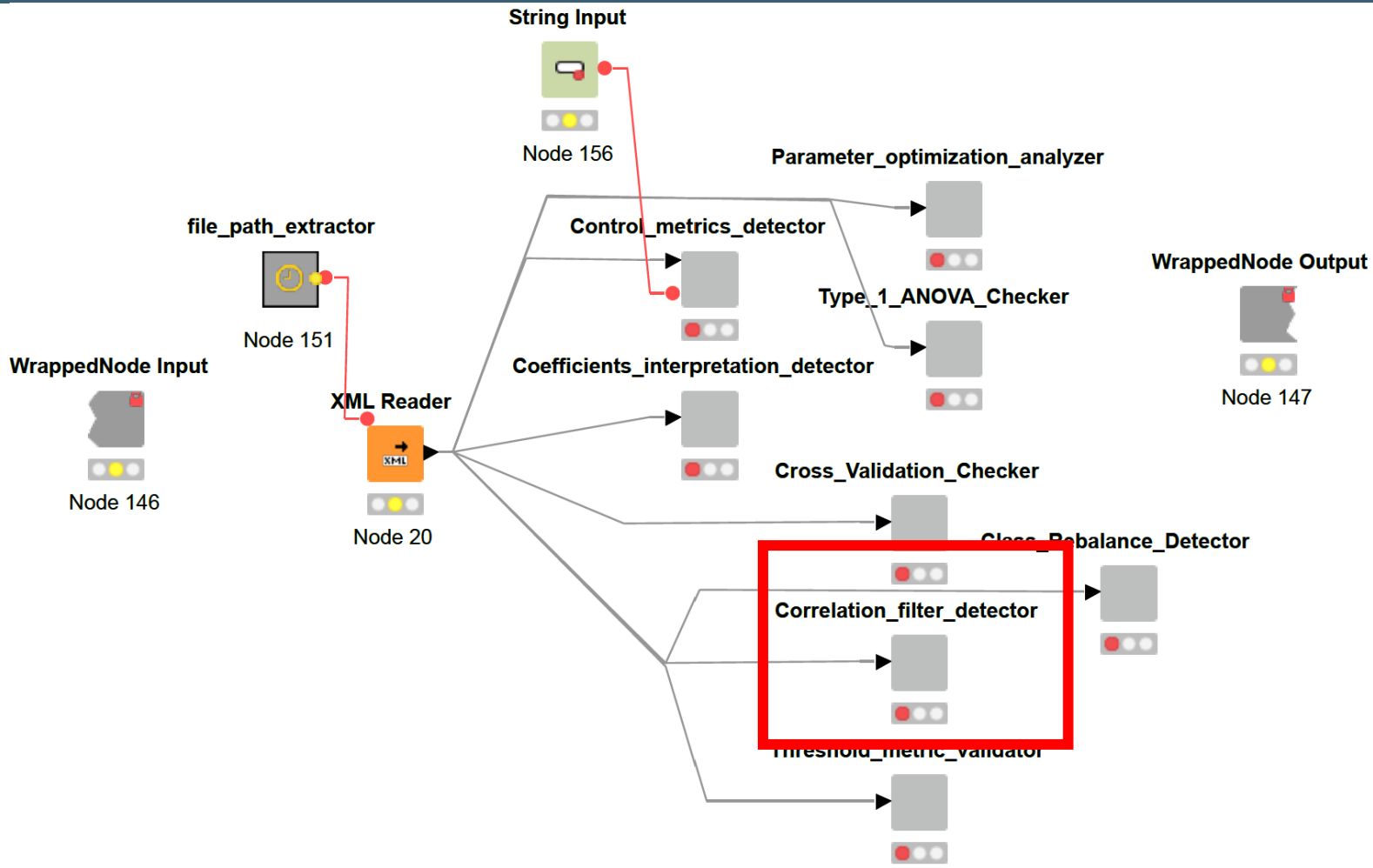
Let's focus on detecting Pitfall 2 for the decision tree learner in the pipeline

Checking for correlation anti-pattern in backward slice of the learner

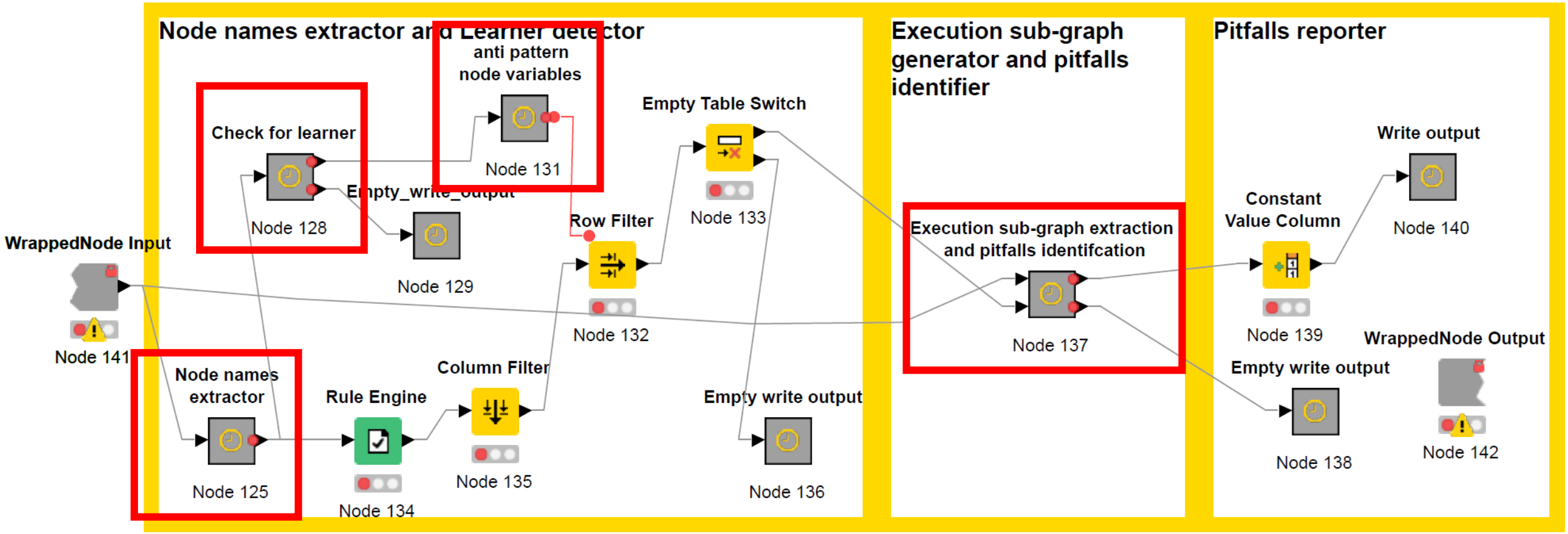


No Correlated variable removal node is present! – The pipeline exhibits the correlation pitfall (Pitfall 2)

Implementation of our Pitfalls analyzer in KNIME



Implementation of our Pitfalls analyzer's Pitfall 2 (correlation pitfall) detector in KNIME



Evaluation of the efficiency of our pitfalls analyzer – Evaluated open data science pipelines

Electricity Consumption Prediction (ECP)	**7 pipelines, 490 nodes**
Rotor Failure Detection (RFD)	**3 pipelines, 126 nodes**
Bikeshare Predictive Analytics (BPA)	**1 pipeline, 29 nodes**

2 of the authors of the paper manually identified the pitfalls present in the evaluated pipelines

Project	#Pipelines	Nodes	#Pitfalls	Pitfalls present
ECP	7	490	4	P2, P4, P5, P8
RFD	3	126	4	P2, P4, P5, P8
BPA	1	29	4	P2, P3, P4, P5

RFD- Rotor Failure Detection, ECP - Electricity Consumption Prediction, BPA - Bikeshare Predictive Analytics. Pitfalls detected via manual analysis.

It took the authors 1 hour to identify all the pitfalls

Our Pitfalls analyzer was able to identify all the pitfalls in seconds

PITFALLS DETECTED IN THE EVALUATED PROJECTS

Project	P1	P2	P3	P4	P5	P6	P7	P8	Total
ECP		×		×	×			×	4
RFD		×		×	×			×	4
BPA		×	×	×	×				4

RFD- Rotor Failure Detection, ECP - Electricity Consumption Prediction, BPA - Bikeshare Predictive Analytics

Our pitfalls analyzer is both scalable and efficient

Easy access enables easy mistakes

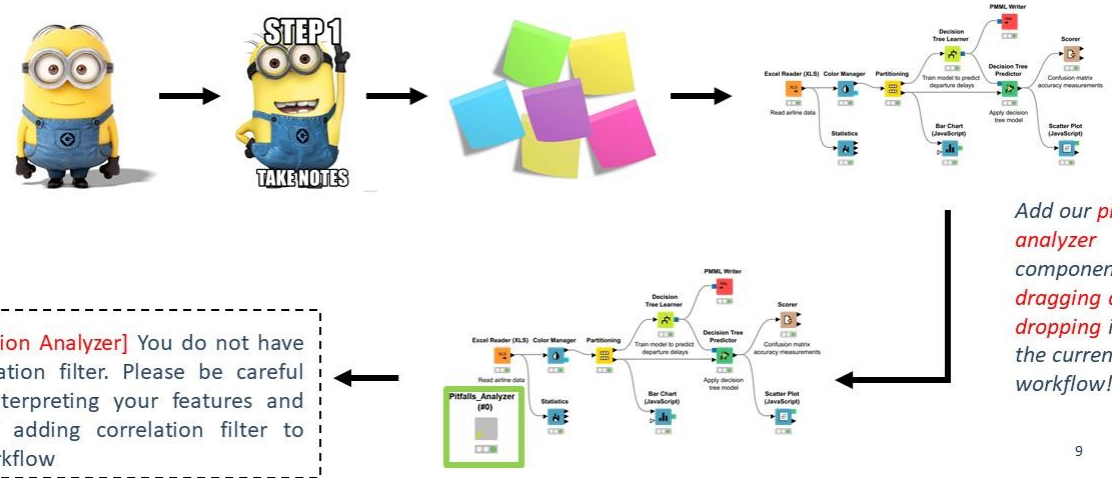
"Simplistic studies comparing data intensive methods with linear regression will be scientifically valueless, if the regression techniques are applied incorrectly."

-Kitchenham and Mendes [PROMISE'09]

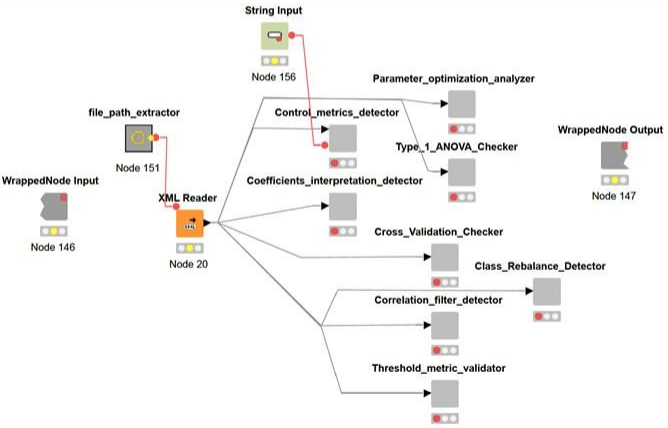
"Many users of such modelling toolkits have limited knowledge about many important details…often leads to major problems which in turn…lead to failure of analytics projects in practice"

-Tantithamthavorn and Hassan [ICSE-SEIP'18]

Let's see how our pitfalls analyzer helps Minion understand exactly what it needs to make the banana plant grow



We enable the identification of 8 common data science pitfalls in a pipeline



Pitfall 1: Absence of control variables

Pitfall 2: Not accounting for the impact of Correlated variables

Pitfall 3: Not accounting for the impact of data rebalancing techniques

Pitfall 4: Not experimenting with different learners or using default settings

Source: Tantithamthavorn and Hassan [ICSE-SEIP'18]

Our Pitfalls analyzer was able to identify all the pitfalls in seconds

PITFALLS DETECTED IN THE EVALUATED PROJECTS

Project	P1	P2	P3	P4	P5	P6	P7	P8	Total
ECP		×		×	×			×	4
RFD		×		×	×			×	4
BPA		×	×	×	×				4

RFD- Rotor Failure Detection, ECP - Electricity Consumption Prediction, BPA - Bikeshare Predictive Analytics

Our pitfalls analyzer is both scalable and efficient

Easy access enables easy mistakes

"Simplistic studies comparing data intensive methods with linear regression will be scientifically valueless, if the regression techniques are applied incorrectly."

-Kitchenham and Mendes [PROMISE'09]

"Many users of such modelling toolkits have limited knowledge about many important details…often leads to

Let's see how our pitfalls analyzer helps Minion understand exactly what it needs to make the banana plant grow

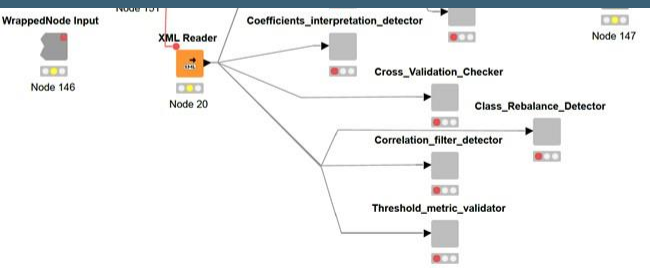
Add our pitfall analyzer component by dragging and dropping it to

[Correlation Analyzer] You do not have

Gopi Krishnan Rajbahadur

Krishnan@cs.queensu.ca

Paper: Pitfalls Analyzer: Quality Control for Model-Driven Data Science Pipelines



Pitfall 3: Not accounting for the impact of data rebalancing techniques

Pitfall 4: Not experimenting with different learners or using default settings

Source: Tantithamthavorn and Hassan [ICSE-SEIP'18]

ECP		×		×	×		×	4
RFD		×		×	×		×	4
BPA		×	×	×	×			4

RFD- Rotor Failure Detection, ECP - Electricity Consumption Prediction, BPA - Bikeshare Predictive Analytics

Our pitfalls analyzer is both scalable and efficient