自動化特徴量エンジニアリングの RapidMinerプロセスでの実装

―AutomaticFeatureEngineering―

株式会社KSKアナリティクス

rapidminer
KSK ANALYTICS JAPAN PARTNER

特徴量エンジニアリング

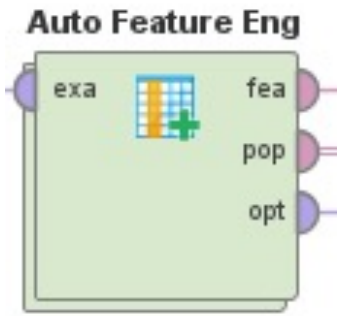
日本企業においてもデータ分析、機械学習を利用する機会が広がりつつあり、モデル開発・運用は容易になってきました。その結果、モデルの出来がビジネスに与える影響も大きくなり、モデルの改善が課題となることも少なくありません。

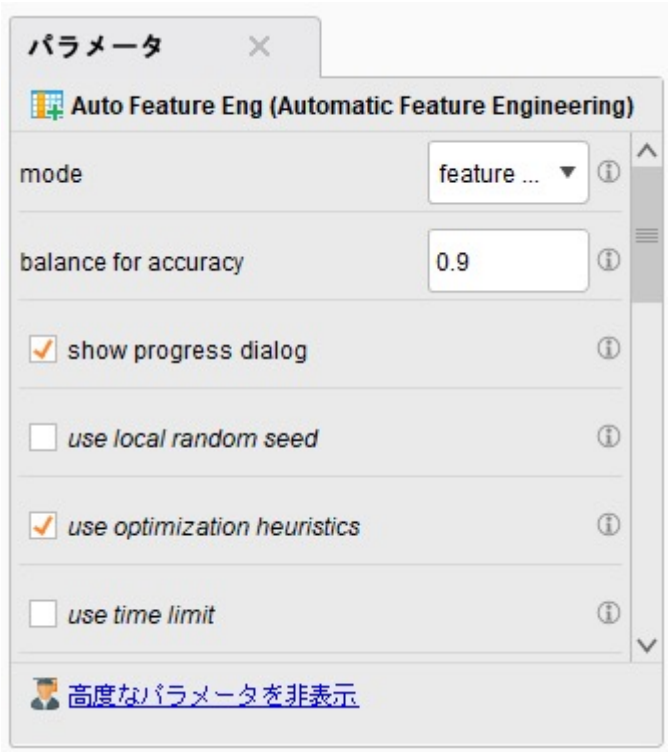
その策の一つとして「特徴量エンジニアリング」が挙げられます。Kaggleなどのコンペでも特徴量エンジニアリングを活用する案は多数見受けられます。一方、特徴量エンジニアリングはデータサイエンスだけでなく業務知識も必要とされることが多く、データ分析を始めたての方では活用が難しいことも多いです。

RapidMinerの”AutomaticFeatureEngineering”を使えば、簡単に特徴量エンジニアリングを取り入れることが出来ますので、是非お試しください。

※ ”AutomaticFeatureEngineering”は有償オペレータとなっております。

AutomaticFeatureEngineering





AutomaticFeatureEngineeringオペレータ
特徴量生成と特徴量選択の双方、
あるいは特徴量選択のみでも実行可能です。
内部にアルゴリズムやPerformanceをセットにしたValidationオペレータをセットします。

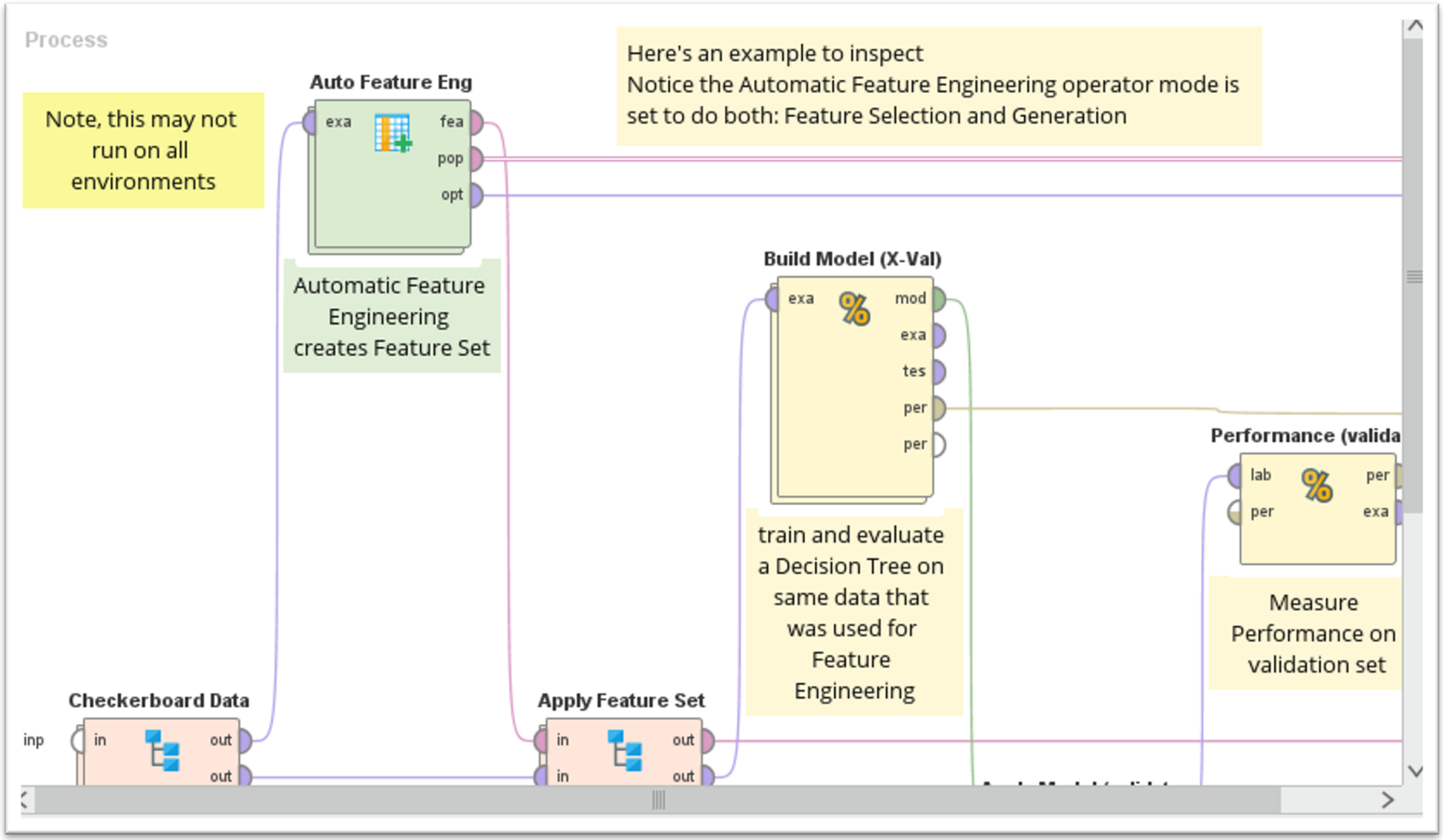
パラメータ設定のbalance for accuracyから複雑さと精度のバランスを調整することが出来、多目的特徴量エンジニアリングにも対応しています。

その他の詳細はヘルプ欄及びヘルプ欄のチュートリアルプロセスでご確認ください。

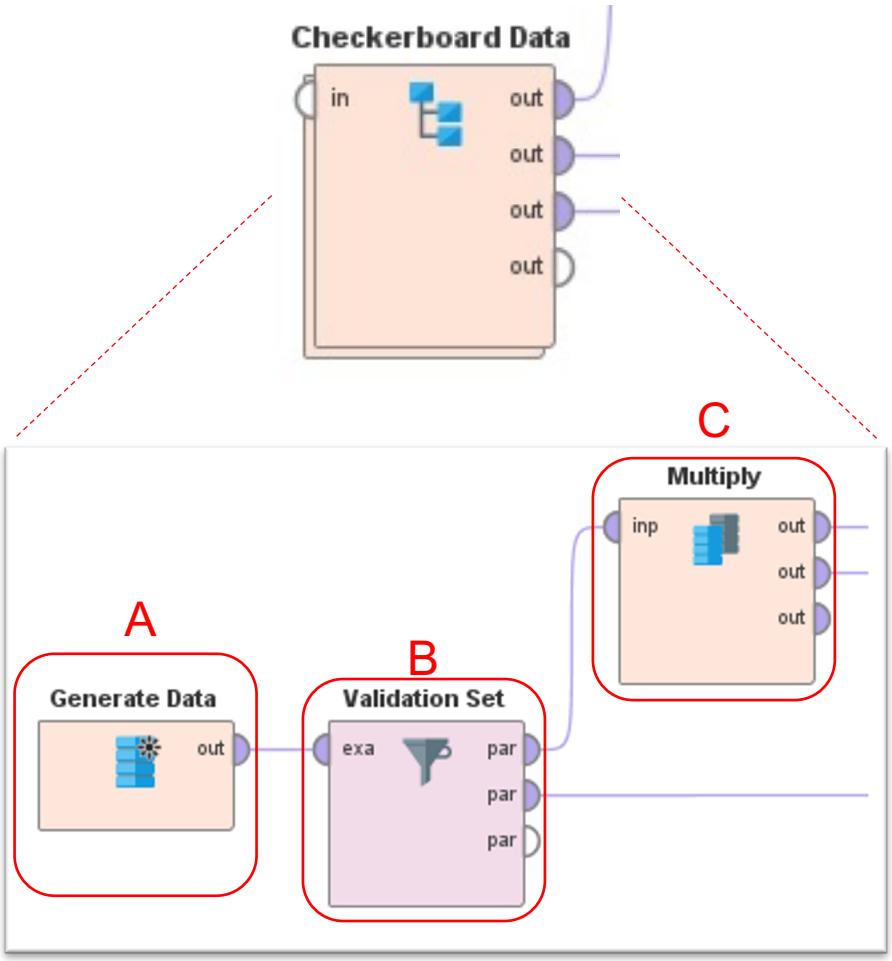
サンプルプロセス

RapidMinerの下記ディレクトリに収録されているサンプルプロセスを例にご紹介します。

//Training Resources/Model/Optimize/Feature Selection/Checkerboard Automatic Feature Engineering Solution



プロセス内容説明①



〇データ準備

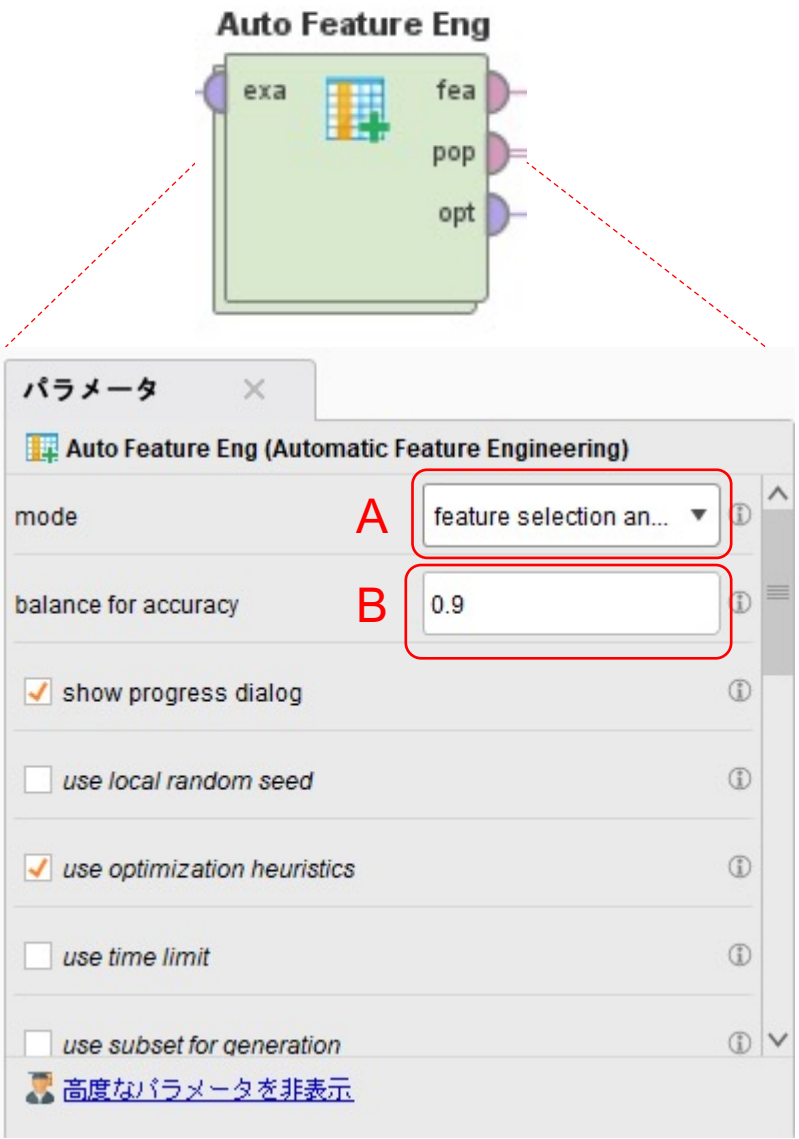
まずこのサブプロセスではチェッカーボードデータ（1000行の二項分類問題）を作成し(A)、

作成したデータを学習データと検証データとして、それぞれ80%、20%に分割します（B）。

分割したデータの内、上から出力されているトレーニングデータはMultiplyオペレータによって二つに複製されてサブプロセス外に出力されています(C）。

サブプロセスの一番上のoutポートはAutoFeatureEngに接続されており、その他の二つはApplyFeatureSetに接続します。

プロセス内容説明②



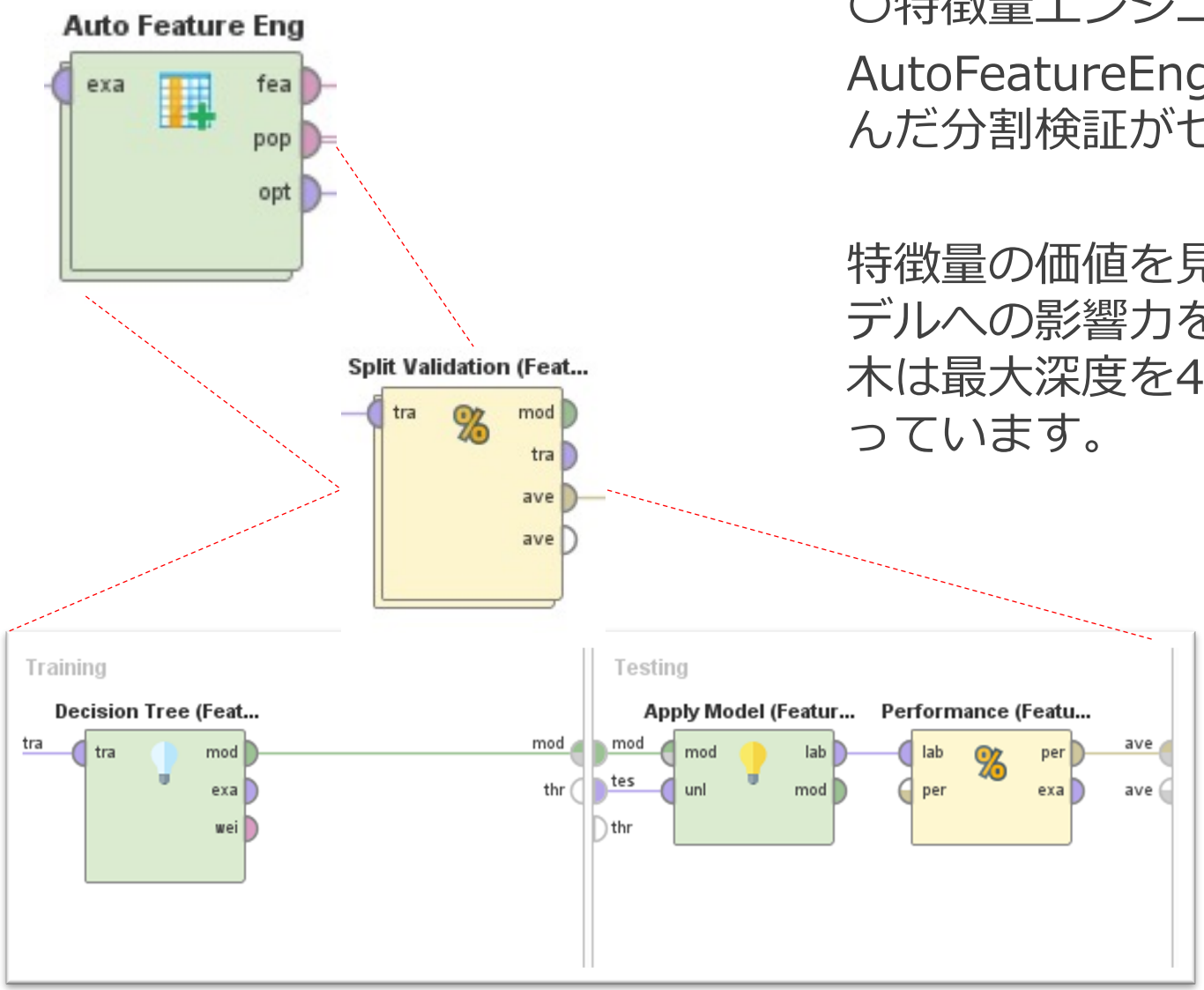
〇特徴量エンジニアリング-1
パラメータ設定を行います。

modeは”feature selection and generation”とし、特徴量の生成と選択を実行させます(A)。

balance for accuracyを今回は精度向上が第一目的にしているので、0.9としています(B）。ほかの問題の場合、精度と複雑性の兼ね合いを取るため、一般的には0.5程度の設定が良いと思われます。

popとoptポートは結果ポートに接続し、feaポートはApplyFeartureSetに接続します。

プロセス内容説明③



○特徴量エンジニアリング-2

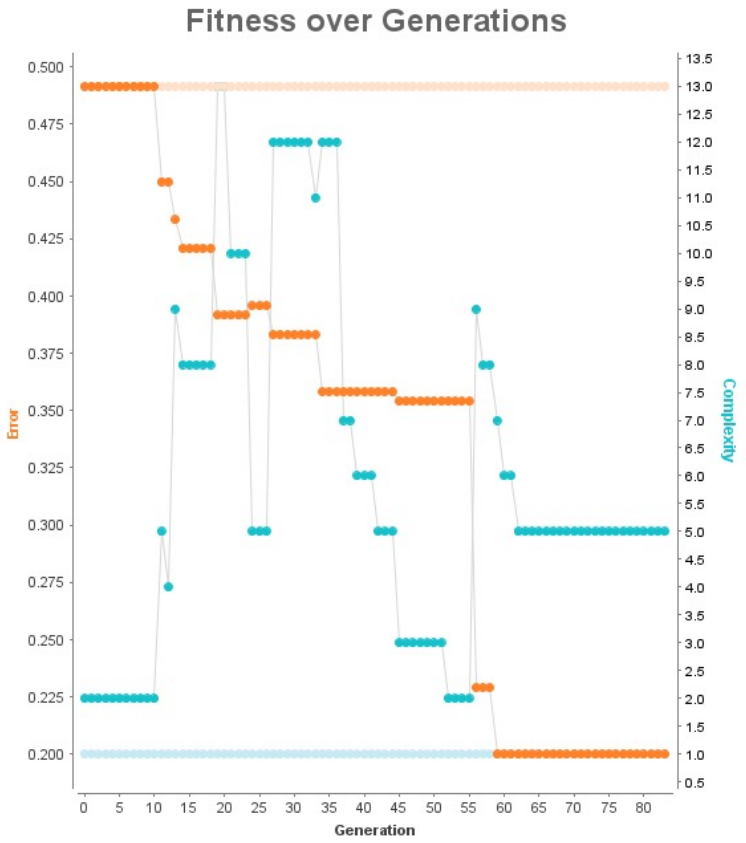
AutoFeatureEngの内部には決定木を含んだ分割検証がセットされています。

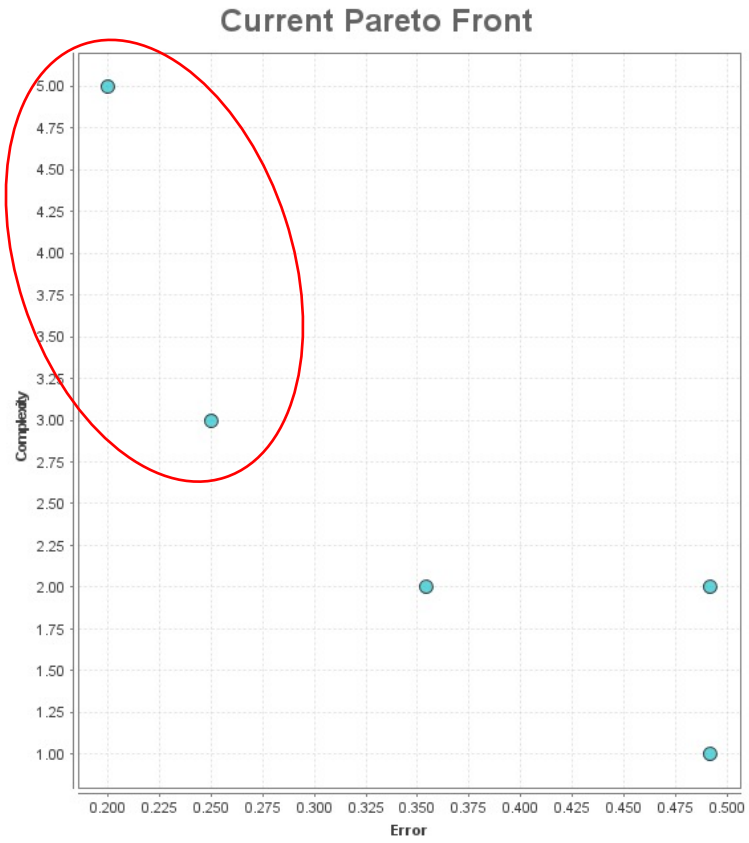
特徴量の価値を見る為、数回の分割でモデルへの影響力を測定できるよう、決定木は最大深度を4とする単純な設定になっています。

プロセス内容説明④

〇特徴量エンジニアリングの確認-1

AutoFeatureEngにブレークポイントを設定し、途中実行すると、
複雑さと精度のトレードオフ関係を可視化したグラフが表示されます。
今回は精度を重視し、バランスを0.9にしたので下の円付近の値を探索します。





プロセス内容説明⑤

〇特徴量エンジニアリングの確認-2

試行のログとそのパフォーマンス、複雑性を示す表(A)、パレートフロントに沿って良い結果の出た5つのポイントのコレクション(B)、最終的に選択された2つの特徴量(C)がそれぞれ結果に表示されます。

CのExpressionではその特徴量がどのように生成されたのか、式を確認できます。

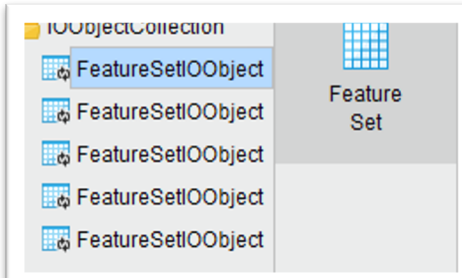
次はこの式を新規データに適用します。

A

Row No.	Generation	Min Error	Max Error	Min Comple...	Max Comple...
1	1	0.492	0.492	1	2
2	2	0.492	0.492	1	2
3	3	0.492	0.492	1	2
4	4	0.492	0.492	1	2
5	5	0.492	0.492	1	2
6	6	0.492	0.492	1	2
7	7	0.492	0.492	1	2
8	8	0.492	0.492	1	2
9	9	0.492	0.492	1	2
10	10	0.492	0.492	1	2
11	11	0.492	0.492	1	2
12	12	0.450	0.492	1	5
13	13	0.450	0.492	1	4
14	14	0.433	0.492	1	9
15	15	0.421	0.492	1	8

ExampleSet (84 行,0 特別属性,5 通常属性)

B



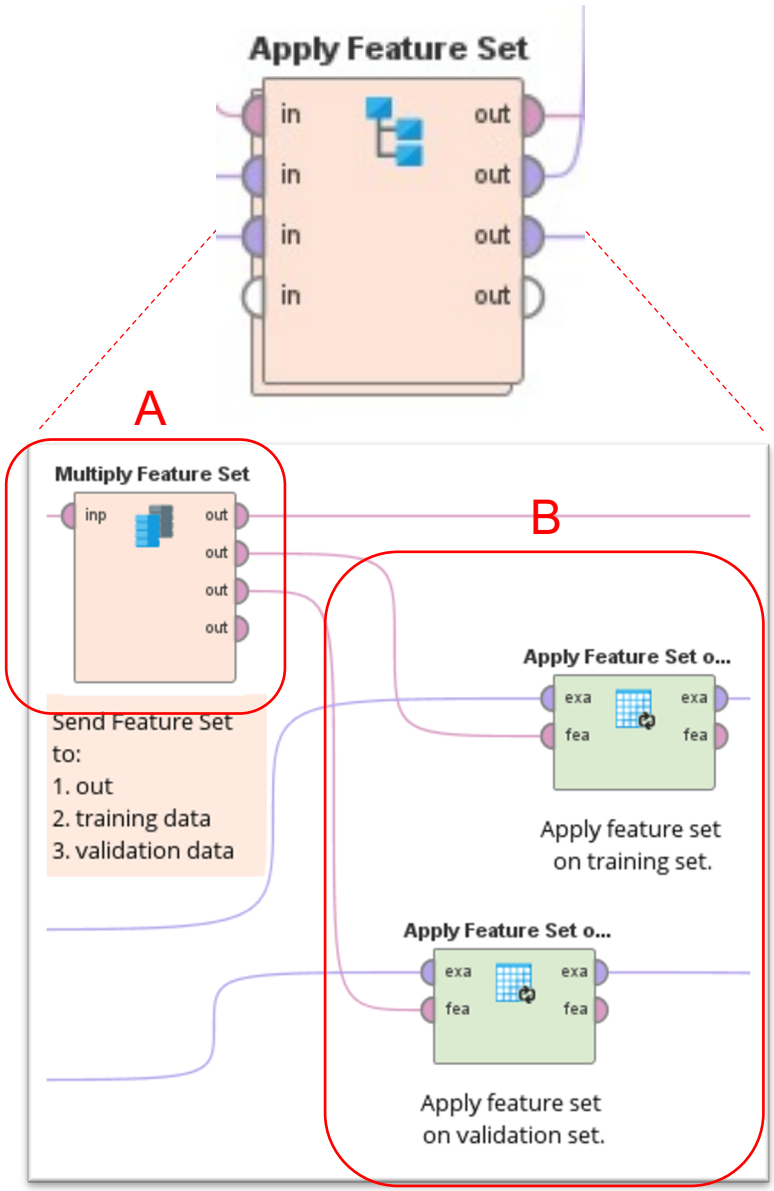
Feature Set (fitness: 0.200; complexity: 5; original: no)

Name	Expression	Complexity
GenSym169	abs([att2]+[att1])	3
GenSym95	[att2]*[att1]	2

C

Name	Expression	Complexity
GenSym169	abs([att2]+[att1])	3
GenSym95	[att2]*[att1]	2

プロセス内容説明⑥



〇特徴量の適用

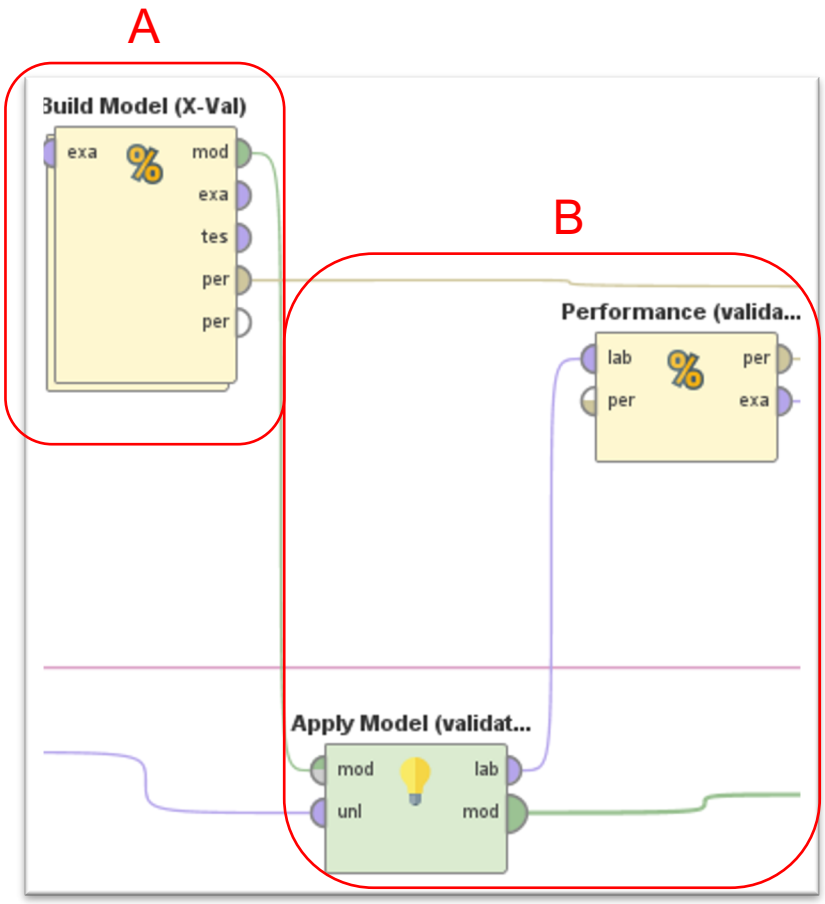
特徴量エンジニアリングで得られた特徴量をMultiplyで複製し(A)、

それぞれ先ほど分割した学習データと検証データに適用します（B）。

ApplyFeartureSetオペレータはAutoFeatureEngで得られた特徴量セットと新規データセットを接続すると、データセットに特徴量セットと同じ構造・同じ式を当てはめることが出来ます。

結果を詳しく見たい場合はここにブレークポイントを設置してご確認ください。

プロセス内容説明⑦



〇モデリング、検証

ここでは特徴量セットを適用した学習データで再び決定木を使いモデリング、交差検証を行い(A)、

そのモデルを検証データに適用し、精度検証を行っています(B)。

再度、決定木を利用していますが、この決定木は先ほどと異なり、最大深度10とする多少、複雑なモデルを作成します。

プロセス内容説明⑧

〇結果確認

最後に精度を確認しましょう。学習データでの交差検証の結果がA、検証データの結果がBです。Aではaccuracyが90.62%±3.55となっており、Bの結果がその範囲内に収まっていることが確認できます。

A

accuracy: 90.62% +/- 3.55% (micro average: 90.62%)

	true negative	true positive	class precision
pred. negative	386	55	87.53%
pred. positive	20	339	94.43%
class recall	95.07%	86.04%	

B

accuracy: 87.50%

	true negative	true positive	class precision
pred. negative	99	22	81.82%
pred. positive	3	76	96.20%
class recall	97.06%	77.55%	

おわりに

今回は”AutomaticFeatureEngineering”オペレータを使った特徴量エンジニアリングをご紹介しました。このオペレータを活用いただければ、特徴量エンジニアリングを簡単に行うことが出来、モデルのさらなる向上に役立ちます。

更にRapidMiner Studioの有償版機能AutoModelにも自動特徴量エンジニアリング機能が付いております。こちらは今回ご紹介したプロセスよりも遥かに早く簡単に特徴量エンジニアリングを実施することが出来ますので、AutoModelをご利用中の方は是非、お試しください。