1. Task Description

Team,

The sales team is considering adding some new products to Blackwell's product mix. They have shortlisted 17 that fit Blackwell's business strategy, but now they need help narrowing the list down to five. I would like to help the sales team by predicting the profitability of each of the potential new products.

I would like you to investigate this question by performing a detailed analysis using regression methods in RapidMiner. Specifically, I would like you to perform a regression analysis to predict the sales volume of each of the potential new products from which profitability can be estimated. In this analysis, our assumption is that certain attributes are associated with highly successful (current) products and, therefore, any potential new products that also have these attributes will be similarly successful, regardless of if a potential new product is similar to an existing product or not.

You will use two new methods for your regression analysis — *k*-Nearest Neighbor (KNN) and Support Vector Machine (SVM)—and you will also explore a new method called Boosting to improve the performance of decision trees. You will need to iteratively adjust the parameters of each algorithm to get the best model. You will then compare the error metrics for your optimized models to assess which one works best. After you have trained your models and determined which one is more accurate, you will apply the model to all of the potential products to predict their sales volumes. After predicting each potential new product's sales volume, you can predict the monthly profits by multiplying the predicted sales volume by the product's price and its profit margin.

Please rank all products in order of highest to lowest profit. I have already set up the data for you in the attached .zip file, which contains the three CSV files you will need.

I am looking forward to reviewing your analysis. This will be a big help to the sales team.

Thank you,

Danielle

2. Task Solution

A data set (existingProductAttributes.csv) was provided by the CTO, a quick review to ensure data is tidy and clean was conducted, several missing values were found in the attribute *Best_Sellers_Rank*. The data set contains 15 missing values out of 80 readings. This represents nearly 23% of the data; based in the best practices to handle missing values, if the missing values are higher than 10% of the data, then the use of average values is not an option. Therefore, the option for this model was to eliminate the attribute Best_Sellers_Rank from the analysis.

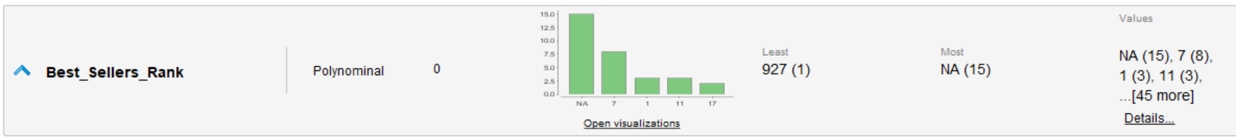


Figure 1 – Attribute "Best_Sellers_Rank" with missing values

The normalization process and the correlation analysis indicate that First-Second attributes correlations are the strongest ones for the pairwise x5Star_Reviews – Volume (1), x2Star-Reviews -x1Star_Reviews (0.952) and x4Star_Reviews – x3Star_Reviews (0.937). In order to eliminate the effects of collinearity the attributes x1Star_Reviews and x3Star_Reviews were subtracted from the analysis. Another benefit to prevent the collinearity is the reduction of machine processing time.

First Attribute	Second Attribute	Correlation ↓
x5Star_Reviews	Volume	1
x2Star_Reviews	x1Star_Reviews	0.952
x4Star_Reviews	x3Star_Reviews	0.937

Figure 2 – Pairwise correlation

One of the CTO's task is the identification of the best algorithm to predict the profitability. In order to achieve this task, two models were developed, Model 1 - K-Nearest Neighbor (KNN) and Model 2 - Support Vector Machine (SVM).

The model optimization was conducted using the operator "Optimize Parameter (Grid)". Both models will be using the same approach as is show in Figure 3. A split data was used to train/test the model, the 70/30 proportion means 70% of data for training and 30% of data for testing.

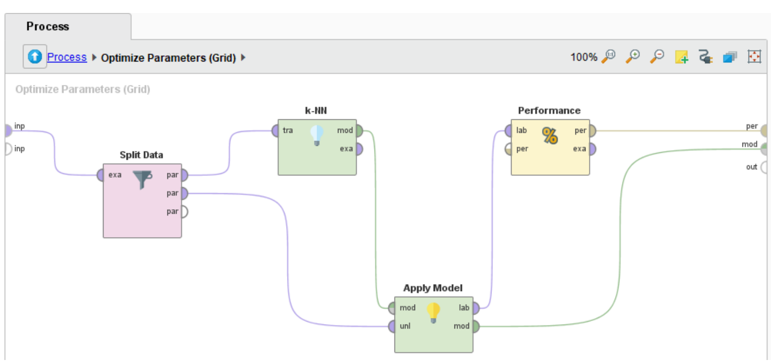


Figure 3 – Optimize parameter operators.

Model 1 - KNN: The K parameter was selected for tuning; using a K range of 1 – 100 and Step = 100, this is a total of 100 combinations (Figure 4). The K value with the lowest RMSE is K = 9 with a RMSE = 238.482 + / - 0.000 and R^2 = 0.880.

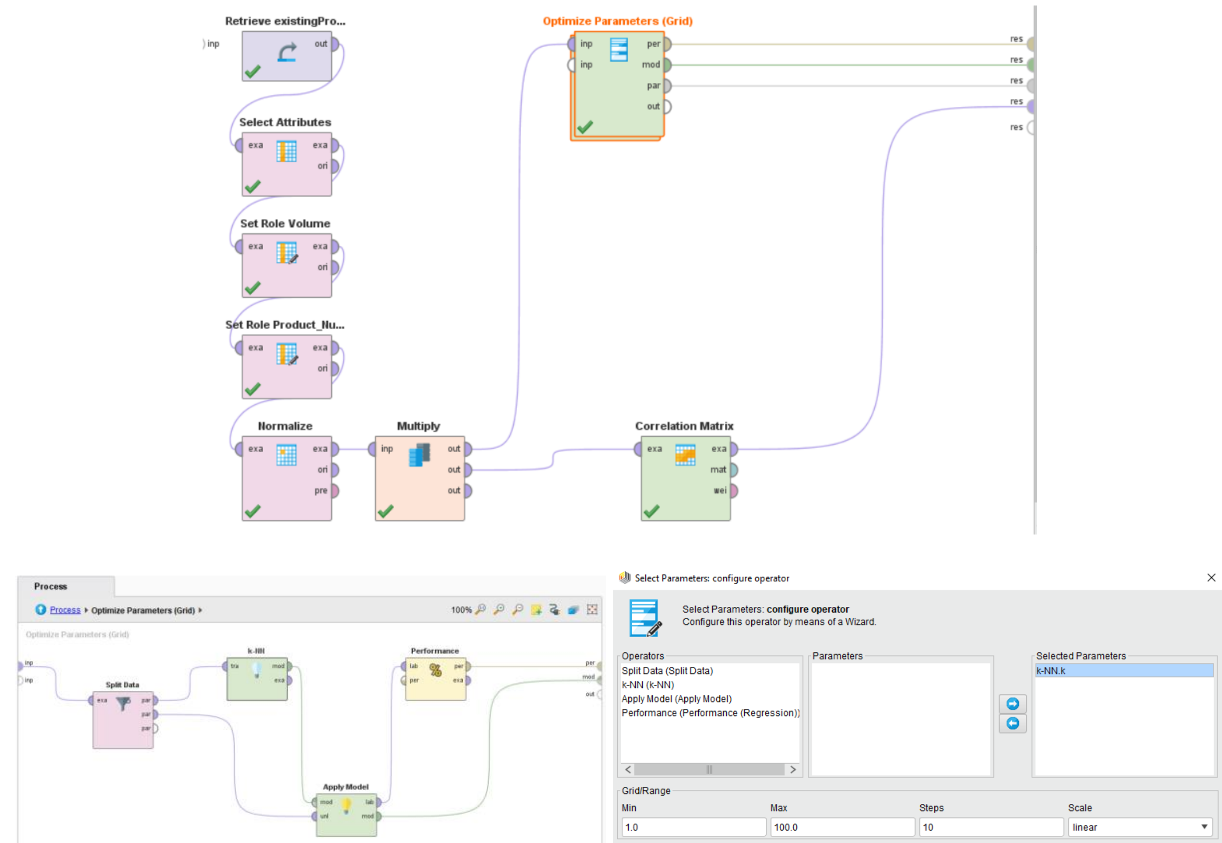


Figure 4 – Tuned parameters for KNN algorithm

Optimize Parameters (Grid) (100 rows, 3 columns)

iteration	k-NN.k	root_mean_squared_error ↑
9	9	238.482
3	3	252.747
2	2	267.848
17	17	363.640
6	6	412.603
41	41	444.163
44	44	451.462
18	18	461.798
37	37	471.018
20	20	489.413
16	16	493.619
48	48	494.634
5	5	500.784
33	33	501.509
35	35	504.841
28	28	507.162

ParameterSet

```
Parameter set:

Performance:
PerformanceVector [
-----root_mean_squared_error: 238.482 +/- 0.000
-----squared_correlation: 0.880
]
k-NN.k  = 9
```

Figure 5 – Performance vector of KNN algorithm

Typically, the k value is set to the square root of the number of records in your training set. Our training set is 80 records, then the k value should be set to sqrt (80) or ~9. Which is consistent with the results provided by the optimize operator. The K=9 showed the lowest RMMSE and the highest square correlation.

Model 2 - SVM: The C and Kernel Type parameters were selected for tuning, using a C range of 10 – 70, Step = 5 and 8 different kernel types, creating 48 combinations (Figure 6). The best values were C = 34; Kernel Type: Anova with a RMSE = 149.720 + / - 0.000 and R^2 = 0.950.

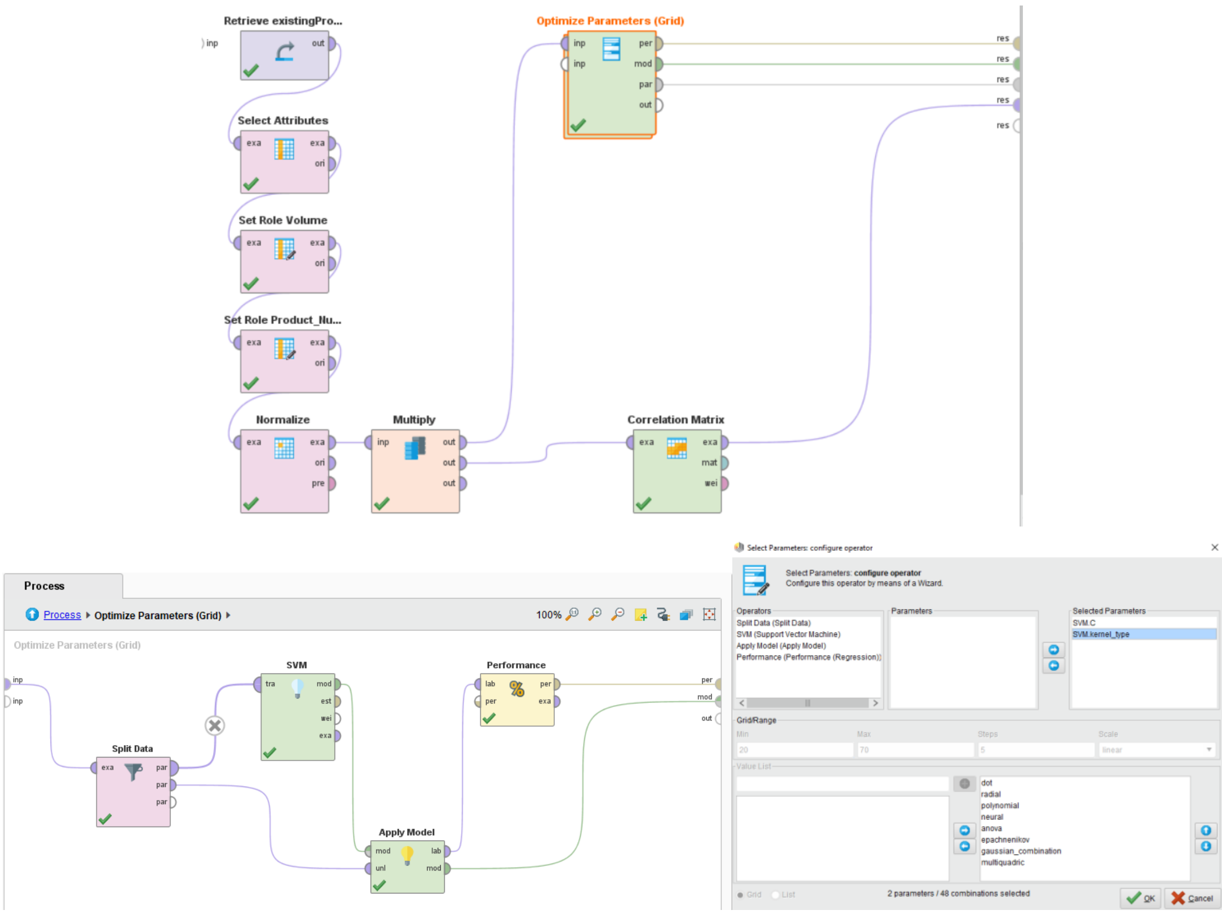


Figure 6 – Tuned parameters for SVM

Optimize Parameters (Grid) (48 rows, 4 columns)

iteration	SVM.C	SVM.ke...	root_mean_squared_error ↑
27	34	anova	149.720
2	22	dot	208.082
3	34	dot	208.846
28	46	anova	213.915
6	70	dot	419.402
18	70	polynomi...	440.506
20	22	neural	451.008
24	70	neural	505.897

ParameterSet

```
Parameter set:

Performance:
PerformanceVector [
-----root_mean_squared_error: 149.720 +/- 0.000
-----squared_correlation: 0.950
]
SVM.C    = 34.0
SVM.kernel_type = anova
```

Figure 7 – Performance vector of SVM algorithm

Using different kernels, the classifier performance changes significantly between models using the same C value, as is show in figure 7 with C = 34 and Anova / Dot kernel for these two iterations the RMSE are 149.720 and 208.846 respectively.

3. Model Selection

The model selection in this case is not so difficult since the RMSE and square correlation using the SVM algorithm are better than the ones presented with the KNN algorithm, as shows in figure 8

	KNN	SVM
Root Mean Square Error	238.482 +/- 0.000	149.720 +/- 0.000
Square Correlation	0.880	0.950

Figure 8 – Variance of RMSE in SVM Algorithm

The main reason of this selection is that a lower RMSE means a higher concentration of the data around the line of best fit. (Lower spread of the residuals); the RMSE gives a relatively high weight to large errors. As a result, lower values of RMSE indicates better fit.

In the other hand R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. In our case SVM showed a higher square correlation, closer to 1, which is good, cause shows a better correlation between the variables.

Another factor evaluated in this process, was the volume predictions, the variance between the volumes predicted by the algorithms, is high between some products, this basically shows the imperfection of the models.

The model with the lowest variance in the results is the SVM, as shows in figure 9, the volume range is 2270 – 2965 units and the volume range for KNN is 255 – 3182 units.

rwo_num	product_num	KNN Volume	SVM Volume	Variance
1	171	1541	2285	744
2	172	298	2303	2005
3	173	298	2426	2128
4	175	255	2637	2382
5	176	255	2289	2034
6	178	298	2381	2083
7	180	2811	2284	(527)
8	181	298	2381	2083
9	183	298	2965	2667
10	186	2463	2275	(188)
11	187	3182	2322	(860)
12	193	2505	2270	(235)
13	194	2962	2310	(652)
14	195	2462	2627	165
15	196	2319	2439	120
16	199	2882	2320	(562)
17	201	306	2403	2097

Figure 9 – Volume variance

It is important to highlight the machine processing time is higher for the model that uses the SVM algorithm compared with the model that uses the KNN algorithm. The elimination of attributes with high correlation values, except for the pair attribute x5Start_Review, helps to reduce the machine processing time.

4. Profitability Prediction

Using the SVM algorithm, the predicted volumes for the requested products are in the range of 2270 – 2965 units.

Row No.	Product_Nu...	Volume	prediction(V...
1	171	0	2285.831
2	172	0	2303.250
3	173	0	2426.660
4	175	0	2637.666
5	176	0	2289.248
6	178	0	2381.046
7	180	0	2284.712
8	181	0	2381.136
9	183	0	2965.751
10	186	0	2275.134
11	187	0	2322.324
12	193	0	2270.492
13	194	0	2310.653
14	195	0	2627.813
15	196	0	2439.119
16	199	0	2320.181
17	201	0	2403.824

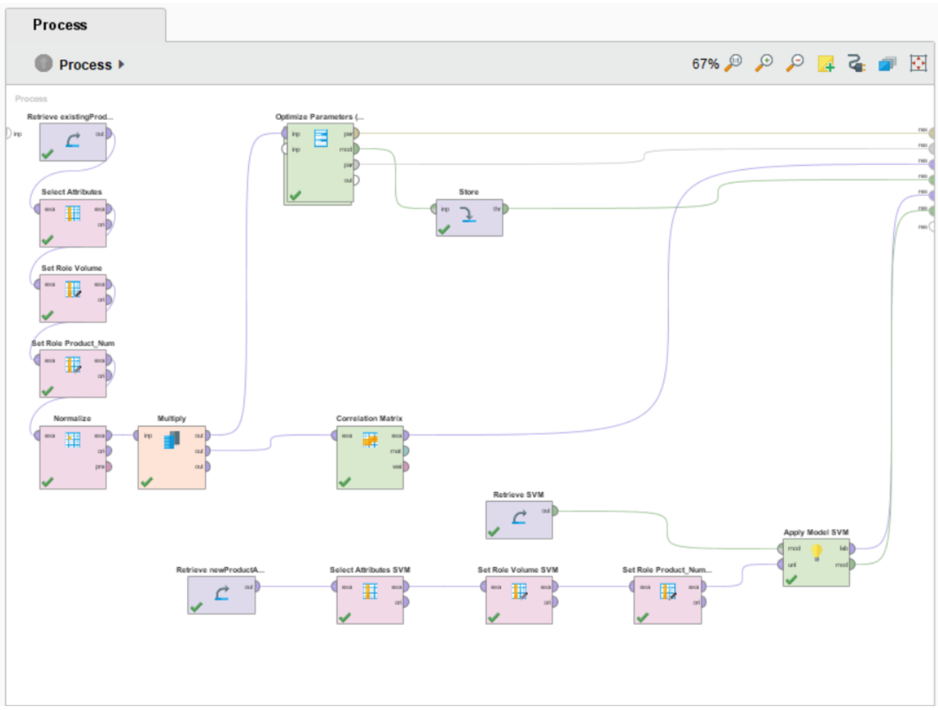
Product Type	Product	Brand Nam	Price	Profit margi	Sales Volume	Profit
Laptop	176	Razer	$1,999.00	0.23	2289	$1,052,414
Laptop	175	Toshiba	$1,199.00	0.15	2637	$474,264
PC	171	Dell	$699.00	0.25	2285	$399,304
PC	172	Dell	$860.00	0.2	2303	$396,116
Laptop	173	Apple	$1,199.00	0.1	2426	$290,877
Tablet	186	Apple	$629.00	0.1	2275	$143,098
Netbook	181	Asus	$439.00	0.11	2381	$114,978
Tablet	187	Amazon	$199.00	0.2	2322	$92,416
Netbook	183	Samsung	$330.00	0.09	2965	$88,061
Smartphone	196	Motorola	$300.00	0.11	2439	$80,487
Netbook	178	HP	$399.99	0.08	2381	$76,190
Netbook	180	Acer	$329.00	0.09	2284	$67,629
Smartphone	195	HTC	$149.00	0.15	2627	$58,713
Game Console	199	Sony	$249.99	0.09	2320	$52,198
Smartphone	193	Motorola	$199.00	0.11	2270	$49,690
Monitor	201	Asus	$140.00	0.05	2403	$16,821
Smartphone	194	Samsung	$49.00	0.12	2310	$13,583

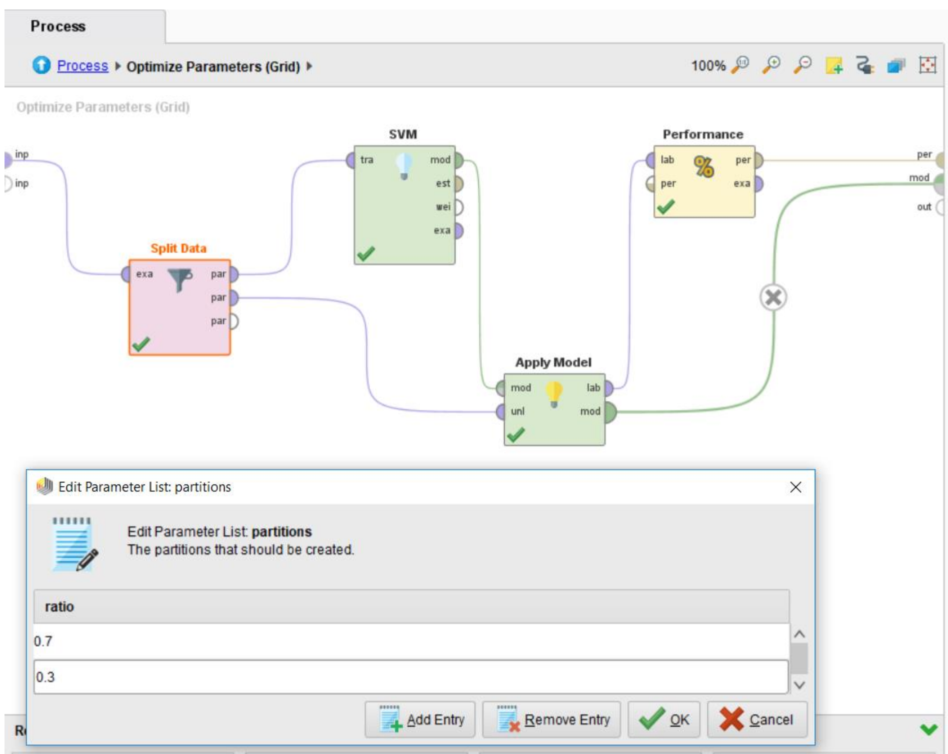
Figure 10 – Product Ranking by Profit

Based in the predicted volumes, the five products with the highest profits are 171, 172, 173 175 and 176, as is showed in figure 10.

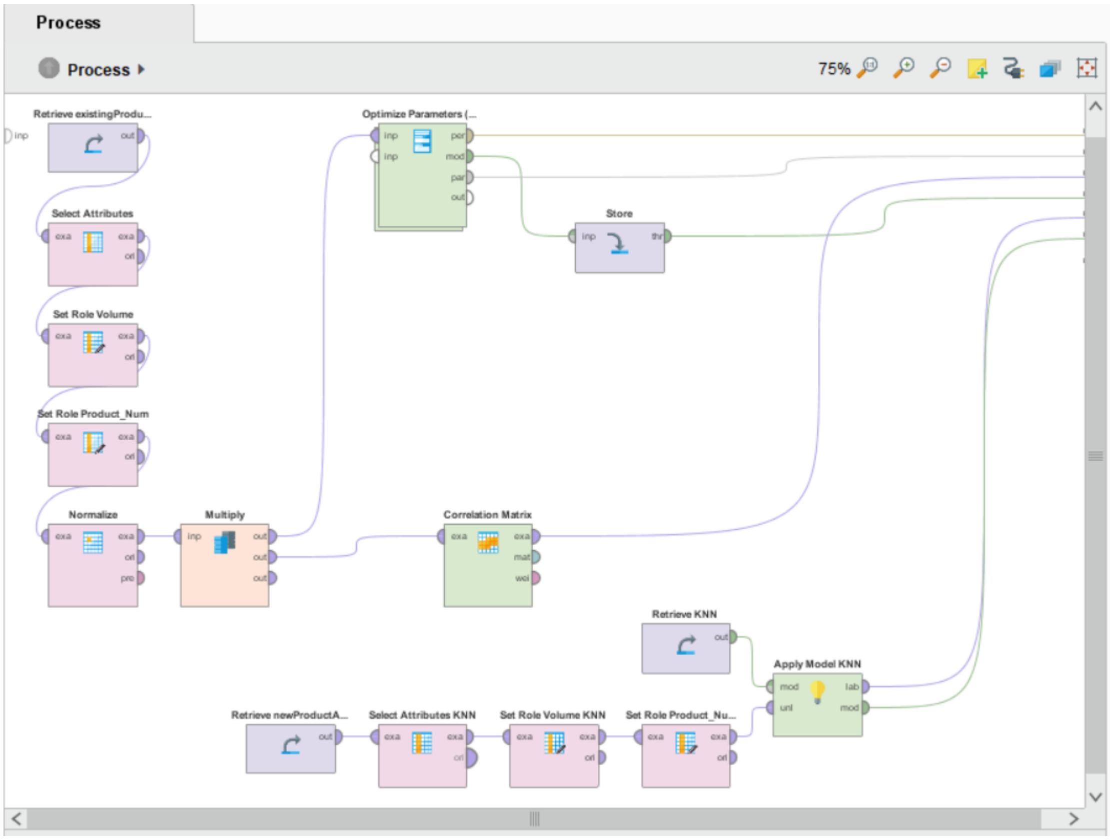
5. Models

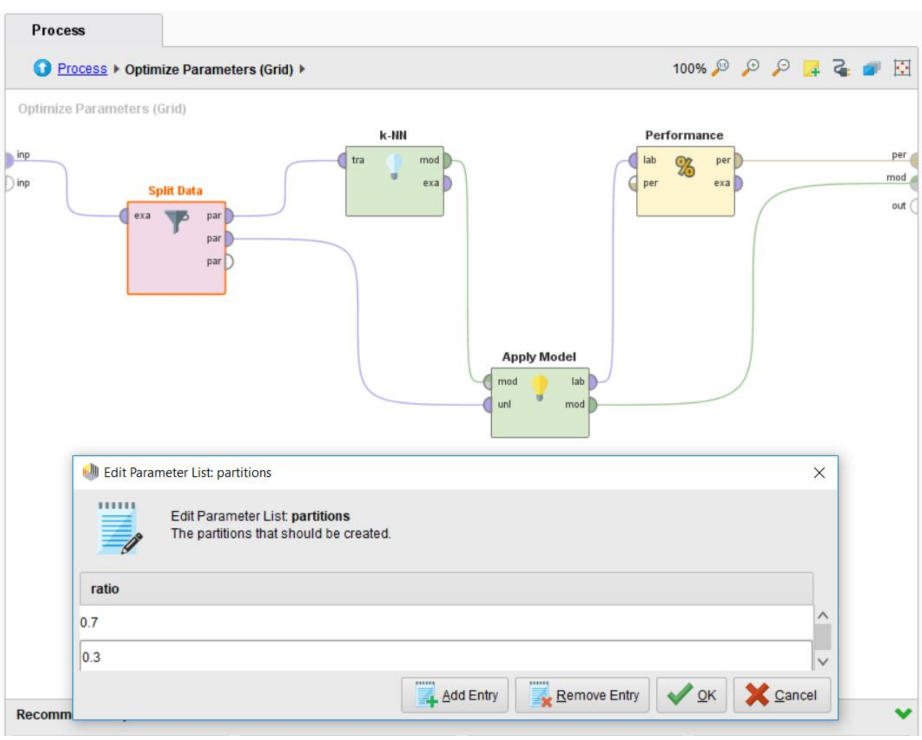
SVM


Process
Process
67%
Retrieve existingProd...
Optimize Parameters (...
Store
Select Attributes
Set Role Volume
Set Role Product_Num
Normalize
Multiply
Correlation Matrix
Retrieve SVM
Apply Model SVM
Retrieve newProductA...
Select Attributes SVM
Set Role Volume SVM
Set Role Product_Num...


Process
Process
Optimize Parameters (Grid)
100%
Optimize Parameters (Grid)
inp
inp
SVM
tra
mod
est
wei
exa
Performance
lab
per
per
exa
per
mod
out
Split Data
exa
par
par
par
Apply Model
mod
lab
unl
mod
Edit Parameter List: partitions
Edit Parameter List: partitions
The partitions that should be created.
ratio
0.7
0.3
Add Entry
Remove Entry
OK
Cancel

KNN


Process
Process
75%
Retrieve existingProdu...
Select Attributes
Set Role Volume
Set Role Product_Num
Normalize
Multiply
Optimize Parameters (...
Store
Correlation Matrix
Retrieve KNN
Apply Model KNN
Retrieve newProductA...
Select Attributes KNN
Set Role Volume KNN
Set Role Product_Nu...


Process
Process
Optimize Parameters (Grid)
100%
Optimize Parameters (Grid)
k-NN
Performance
Split Data
Apply Model
Edit Parameter List: partitions
Edit Parameter List: partitions
The partitions that should be created.
ratio
0.7
0.3
Add Entry
Remove Entry
OK
Cancel
Recomm