# **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

# **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.

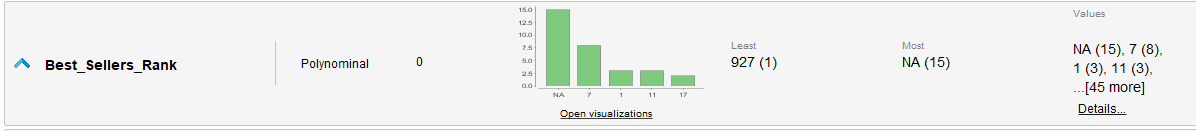


Figure 1 – Attribute “Best\_Sellers\_Rank” with missing values

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.



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.

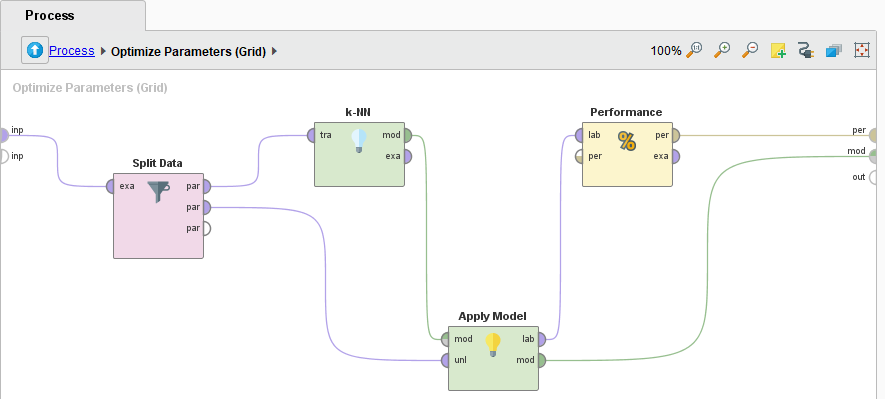
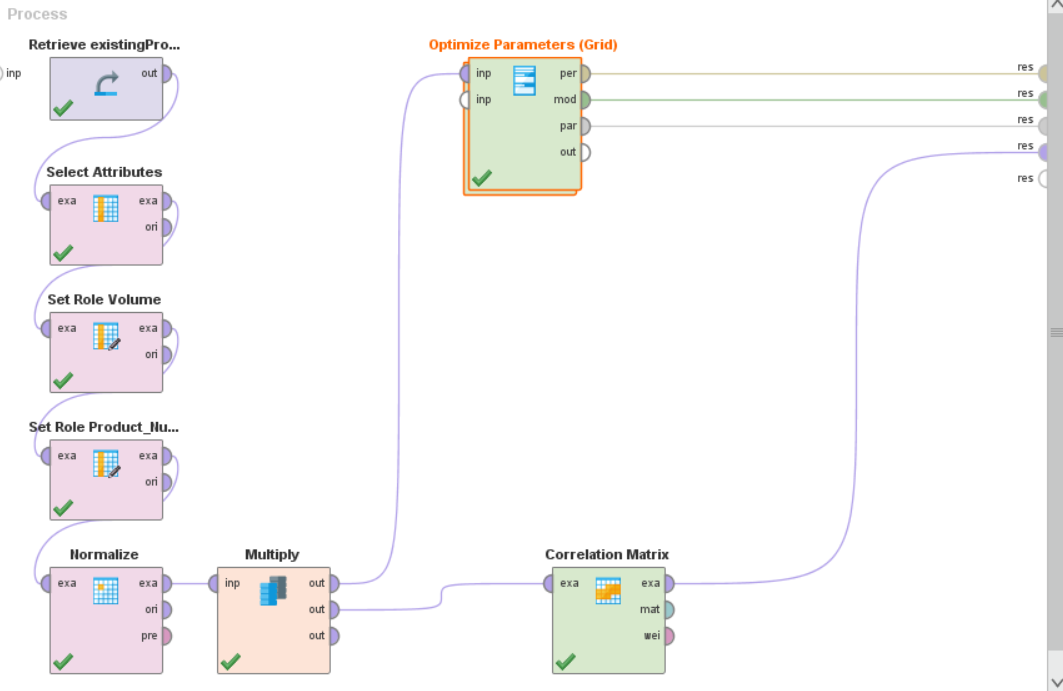


Figure 3 – Optimize parameter operators.

**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 R2 = 0.880.



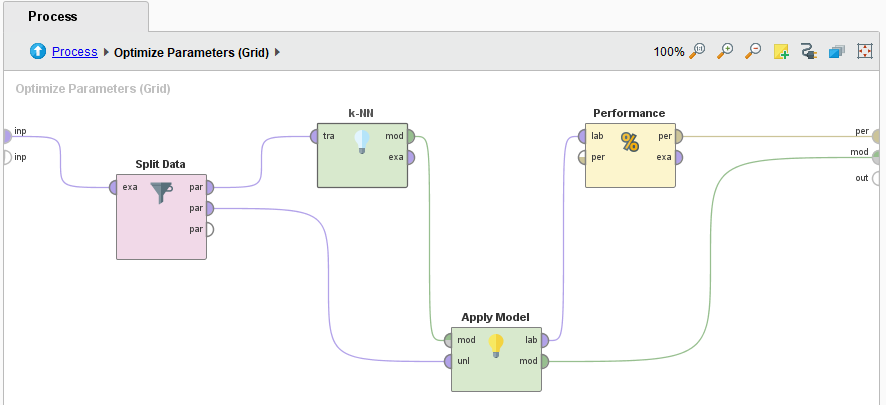
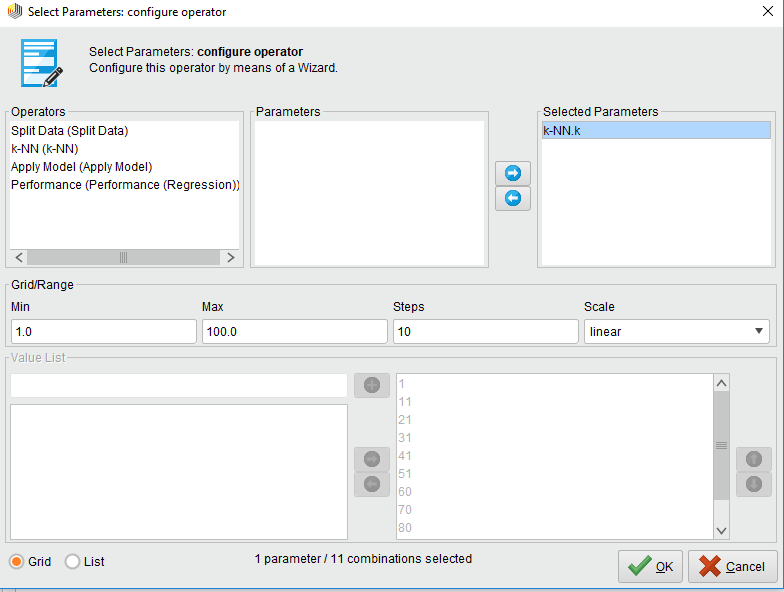
 

Figure 4 – Tuned parameters for KNN algorithm

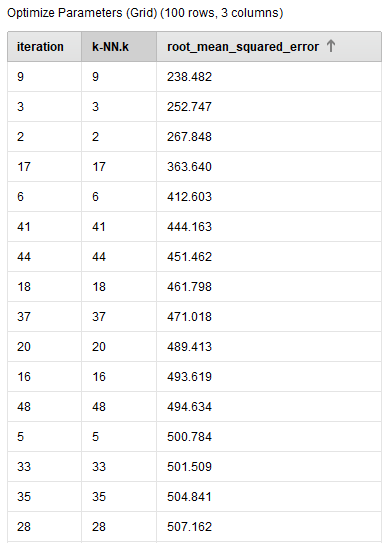
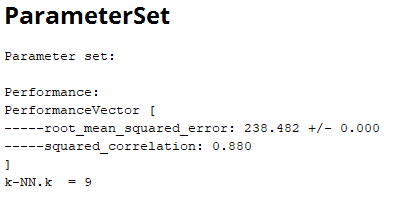
 

Figure 5 – Performance vector of KNN algorithm

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 R2 = 0.950.

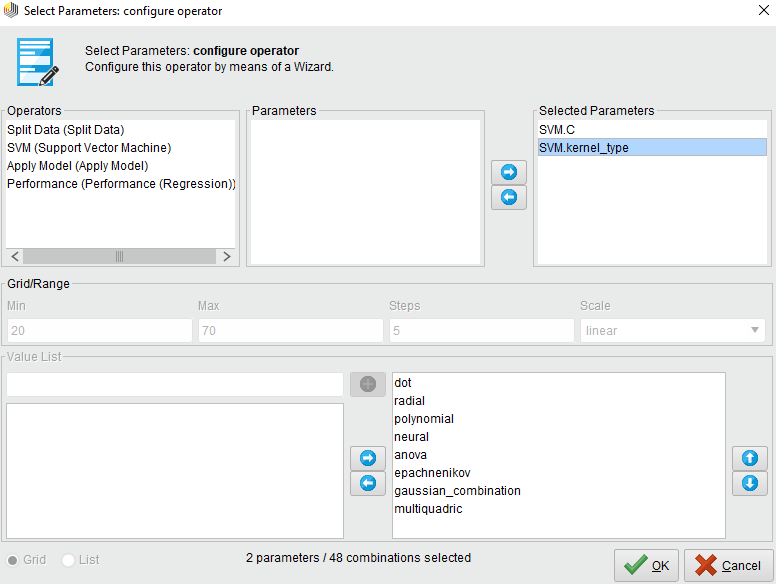
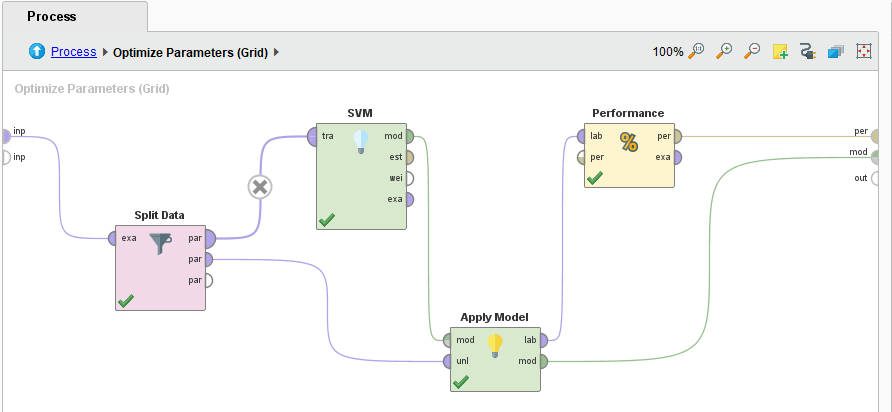
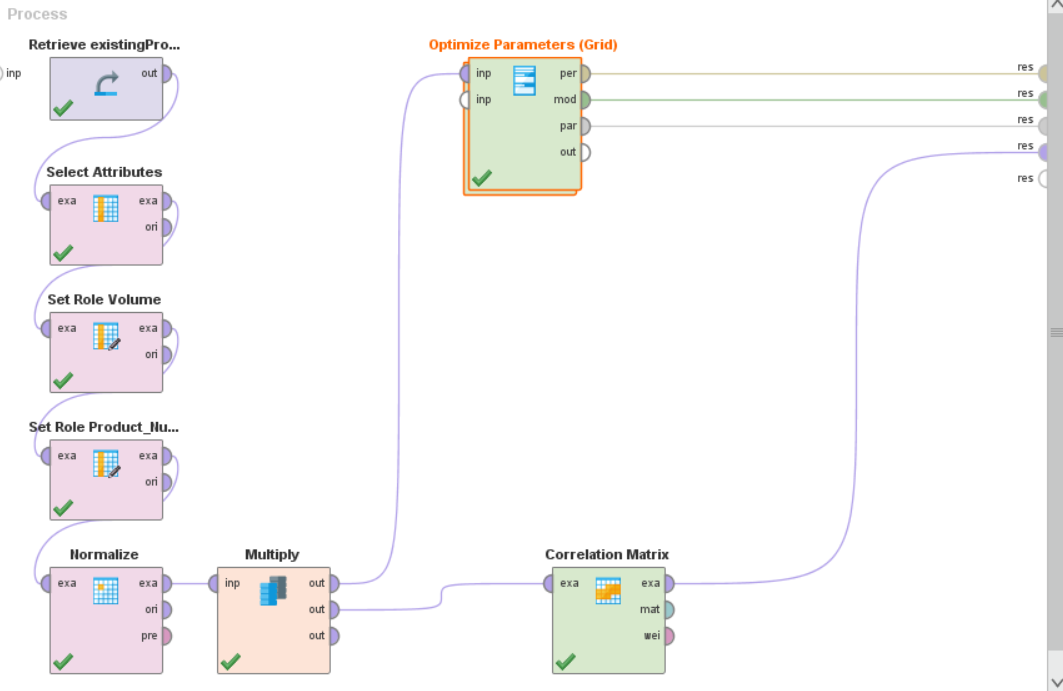


Figure 6 – Tuned parameters for SVM

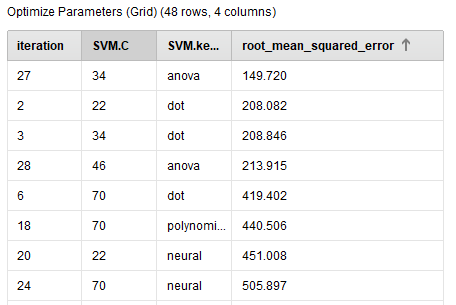
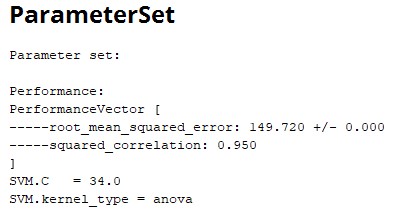
 

Figure 7 – Performance vector of SVM algorithm

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.

# **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



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.



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.

# **Profitability Prediction**

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

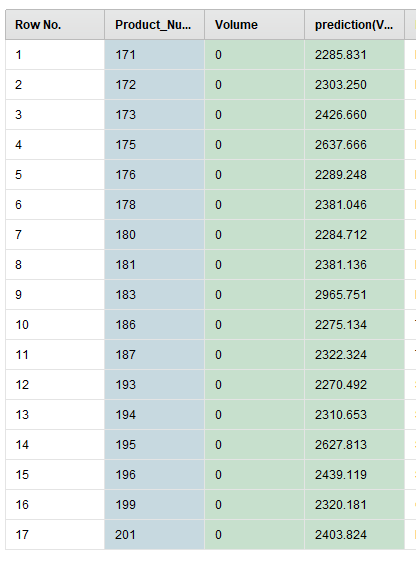
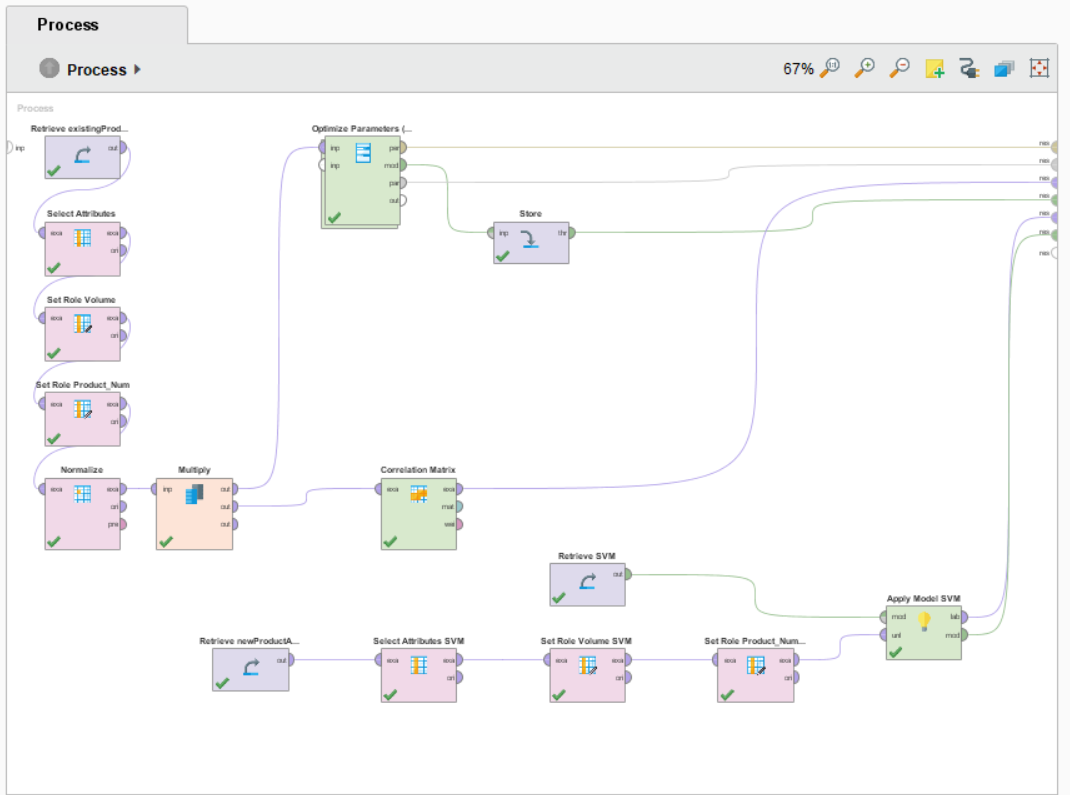
 

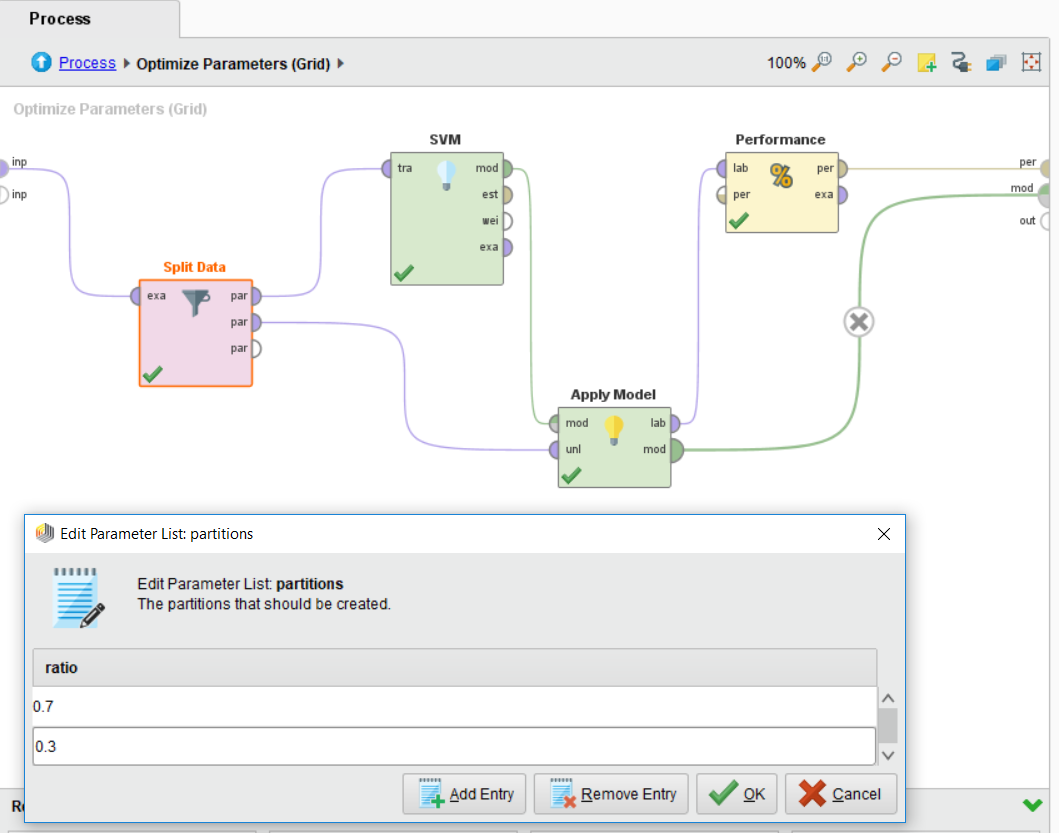
Figure 10 – Product Ranking by Profit

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.

# **Models**

**SVM**





**KNN**

