Task 2: **Predicting Profitability**

**FROM:** Danielle Sherman  
**Subject:** Help Sales Team Decide on Potential New Products

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

**Danielle Sherman**

Chief Technology Officer

Blackwell Electronics

www.blackwellelectronics.com

**Attachments:**

https://common.socraticarts.com/lib/images/windowsIcons/zip.gif [Product\_Analysis\_Data](https://s3.amazonaws.com/gbstool/emails/2779/Product_Analysis_Data.zip?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20200118T153719Z&X-Amz-SignedHeaders=host&X-Amz-Expires=28800&X-Amz-Credential=AKIAJBIZLMJQ2O6DKIAA%2F20200118%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=bae8f046cb7f05838879cbb47f6659bd6681a45c34fa8dd133fab45aaddb963a" \t "_blank)

# **Task 2:** Predicting Profitability

[INTRODUCTION](https://codeacademy.ubiqum.com/mc/poa?productID=6637&taskID=4242#introduction)

# **Your Task**

You have been asked by Danielle Sherman, CTO of Blackwell Electronics, to predict the overall profitability of each of a list of potential new products using regression analysis.

**This task requires you to prepare a New Product Profitability Report for Danielle Sherman. The report should be created in Excel or Word and contain the following:**

* A ranking of potential new products, ordered from highest to lowest profitability
* A brief summary of the optimized model you selected and your rationale for selecting it. Include the parameter settings for this model
* A summary of performance metrics from each individual classifier you ran

The steps in the following tabs will walk you through this process.

#### **[1. Get Started](https://codeacademy.ubiqum.com/mc/poa?productID=6637&taskID=4242" \l "collapsepoa12318)**

In Task 1 you completed a series of tasks in RapidMiner to analyze a dataset. Let’s put some structure around those tasks in the form of the series of steps that are typical of an Analytic Process.

1. **Obtaining Data**: Getting your data can be easy if it is being handed to you. Often you may need to get your own data by way of crawling the web, SQL queries, Twitter scraping, or by piecing together data from multiple sources.
2. **Cleaning the Data**: Data must often be "cleaned" before it can be successfully analyzed.
3. **Importing Data**: Prepared data is imported into analytics tools. In the importing step analysts should pay special attention to the "data types" the attributes are being assigned by the analytics tool.
4. **Initial Exploration of Data**: The longer you work as an analyst; the more emphasis you will place on this step. Using visualizations to understand the shape of your data is of huge benefit and can help determine the type of algorithms you choose for modeling. Computing summary statistics and understanding the structure of the data serve to inform potential changes needed in preprocessing.
5. **Preprocessing**: Analytic tools are powerful, but they often mis-determine the "data types" of attributes. Preprocessing involves addressing "data type" conversion like numeric to nominal and Date/Time conversions.
6. **Feature Selection and Feature Engineering**: These tasks include understanding how the features (attributes) relate to each other as well as modifying, creating and/or deleting attributes in the data set. The goal of Feature Selection and Feature Engineering is improved modeling performance.
7. **Modeling and Optimization**: This is the application of one or more algorithms to understand patterns within data sets and to make predictions about future outcomes. We can optimize our models by adjusting algorithm parameters and observing the performance metrics until we have found the best fitting model. We can also build models with different algorithms and compare the results.
8. **Making Predictions**: Applying an optimized model to a test set or unseen data to provide predictions that can help solve business or engineering issues.
9. **Reporting Your Findings**: The best analysis in the world is only as good as how well it is reported. Always consider your audience’s level of technical background when developing reports and focus on answering the key stakeholder questions you set out to investigate.

The steps enumerated above are not necessarily linear. Often you will make a discovery in a step that leads to a new understanding about the data. This may cause you to return to an earlier step to utilize this new knowledge.

#### **[2. Obtain and Prepare Data](https://codeacademy.ubiqum.com/mc/poa?productID=6637&taskID=4242" \l "collapsepoa12321)**

**Obtaining Data**

Download the data from Danielle Sherman’s email. Inside the zip folder you will find three files:

1. **existingProductAttributes.csv** – a file containing information about product features, reviews and historical sales information. You will use this file to train and test models
2. **newProductAttributes.csv** – a file containing product features and reviews, but no sales information. You will use this file and a trained model to make predictions about sales volume
3. **profitability.xlsx** – a file you will use to calculate new product profitability after you have made predictions

**Data Preparation**

RapidMiner (and other tools you will be using soon like R/RStudio) are powerful analytic tools, but they can’t overcome the old adage: Garbage In, Garbage Out.

Preparation of data before importing it into a tool is an extremely important task. Proper preparation can save hours of troubleshooting and frustration. The guidelines below cover typical tasks associated with "cleaning" your data for import.

* Attribute headers should always occupy the top row
* Numerals should not be the initial characters in the names of attribute headers and nominal attribute values. Often an "x" is placed before an attribute header to solve this problem. Example: 5StarReview -> x5StarReview
* Headers and nominal attribute values should not contain spaces between words. Analytic tools interpret spaces to mean that a new argument is to follow. Deleting space is often done as follows: Best Sellers Rank   -> BestSellersRank
* Special characters and symbols should be avoided. This includes {, }, ?, $, %, ^, &, \*, (, ),-,#, ?,,,<,>, /, |, , [ , and ]
* Any missing data values in your data set should be replaced with NA
* Any comments embedded in the data set should be deleted

**Your Tasks:**

1. Open the existingProductAttributes.csv file with Excel. Inspect the data set and make updates based on the preparation guidelines above. Check attribute headers, nominal attribute values and address missing values. Save the file when you are finished.
2. Open the newProductAttributes.csv file with Excel. Any edits you made to the existingProductAttributs.csv file must also be made identically to newProductAttributes.csv.

You are ready to import the data in the next step. Note that anything you may have missed could cause issues later. If your analysis hits a roadblock consider revisiting your data preparation as a step in troubleshooting.

#### **[3. Import and Initially Explore Data](https://codeacademy.ubiqum.com/mc/poa?productID=6637&taskID=4242" \l "collapsepoa12323)**

Using the skills you learned in Task 1, import the existingProductAttributes.csv file into RapidMiner. After upload, explore the data via the Statistics and Charts tabs.

1. Which product has the highest number of 5 Star Reviews?
2. Which product has the highest sales volume?
3. Do the data types match your expectations?
4. Use Statistics and Charts to look for other relationships that may be useful to your analysis.

#### **[4. Preprocess the data](https://codeacademy.ubiqum.com/mc/poa?productID=6637&taskID=4242" \l "collapsepoa12326)**

In Task 1 you learned that it is often necessary to apply filters to data. Review the data types on the Statistics Tab. There are three attributes that need attention: Product Type, Product Number and Best Sellers Rank.

* **Product Type** is a Polynominal attribute. Since we are working on a regression problem, this type of data can’t be measured as it has no numeric value. We will remove this attribute now, but later in the course you will learn how to include nominal attributes when working with regression.
* **Product Number** is an integer attribute. Things like product numbers and IDs are part of a special data class called unique identifiers. Unique identifiers typically do not add value to analysis. For example, how would you measure a meaningful difference between product 101 and 112? Product Number should be removed to reduce "noise" in the data. We can set this feature as an ID so it won’t be used in the model, but will still be available in the results.
* **Best Sellers Rank** contains missing data. Consult resources for methods that deal with missing data. In this task we should remove the attribute, but that isn’t the case in all situations.

Optional: Consider the remaining attributes. Are there any that do not seem relevant to predicting Sales Volume?

**Your Tasks:**

* Use the Select Attributes Operator to remove Product Type and Best Sellers Rank
* Use the Set Role Operator to set your dependent variable to "label"
* Add a second Set Role Operator and set "Product Number" as "ID"
* Connect the operators and hit run. Review the results to assure your filters have been applied correctly.

**Optional:**It's often helpful to rename your operators so that you know which attribute is being filtered. You can do this by right clicking the operator and selecting "rename operator".

**Normalization**

After inspecting the data you can see that individual features are not represented by the same units. For example: prices represented in some unit of currency, weight is represented by some unit of weight and volume represent volume, regardless of the type of unit. Whenever we are dealing with machine learning tasks like regression we should always make sure that all of our features are represented on the same scale. We do this in order to prevent any one specific feature possibly biasing the model we build.

Normalization is done with some fancy mathematics, but luckily for us there is a 'normalize' operator we can use in RapidMiner. Add it to your current process and select all of the features using the subset parameter of the operator.

**TIP:**

Should volume be considered a feature we should normalize? If normalization converts all features to be on the same scale, which is normally between 0 and 1, what will also happen to volume if we normalize it? How will this impact our predicted amounts of volume?

#### **[5. Engineer Features in the Data](https://codeacademy.ubiqum.com/mc/poa?productID=6637&taskID=4242" \l "collapsepoa12327)**

**What is Feature Engineering?**

From your studies you understand that we work with independent (x-values or predictors) and dependent variables (y-values) when building predictive models. The independent variables make up the features in your data and can directly influence the predictive models you build.

During the initial exploration and preprocessing of the data, it is critical to develop a method for understanding these features. We need to ascertain how the features relate to each other and how the features relate to the dependent variable(s).

The better we understand and prepare the features the better results we will achieve when building our predictive models. We also study features to guard against issues that might cause overfit or underfit in various types of models. For example: Colinearity exists when two independent variables are highly or even directly related to each other; this occurrence will cause regression model to overfit in most cases so one of the features will need to be removed – which one(s) will be up to you to decide as part of your exploration.

**Feature Selection with Correlation**

To understand the relationships in the data set we will use the Correlation Matrix operator. Correlation Matrix will output a grid containing correlation measures for all of the attribute relationships in the data set.

1. Click on the Design Tab. Disable the Cross Validation and Store operators by right clicking on them and deselecting "enable"
2. Find the Correlation Matrix operator and drag it into the process. Connect Set Role exa to Correlation Matrix exa. Now connect Correlation Matrix output exa and mat ports to res ports
3. Select the Set Role operator and change Volume from "label" to "regular". If you leave Volume as a label it will not appear in your correlation matrix
4. Click Run to produce the Correlation matrix and a table of the attribute weights based on the correlations of the attributes

**The Correlation Matrix**

In the results tab, you should see a correlation matrix (example shown below) where you can review the strength of the relationships between every feature. Each correlation is a number between negative one and positive one. Positive numbers imply positive associations while negative numbers imply inverse associations. The following steps will help you interpret this matrix.

1. Check for high correlations with the dependent variable. Look at the Volume column. Are there any correlations that are above 0.95? If so, note the associated attributes.
2. Using the 'pairwise table' on the left Check for collinearity between the features. Search through the table to see if any of the independent variables have correlation coefficients of 0.90 or higher with any of the other independent variables. If any of the independent variables are collinear you might need to remove one of the two of them to address the collinearity.
3. Back in the design view, use the Select Attributes operator to remove the features you identified in the preceding steps.

#### **[6. Train and Assess the Models](https://codeacademy.ubiqum.com/mc/poa?productID=6637&taskID=4242" \l "collapsepoa12329)**

In this step we will work with three algorithms:**KNN, SVM, and Gradient Boosted Trees**(check the resources for more information on each).

**Set up the modeling process**

With feature selection complete, it's time to update your process for modeling. On the design screen you need to do the following:

1. Disable and remove the Correlation Matrix operator
2. Insert a Cross Validation operator into your process
3. Open the Cross Validation operator:
   1. Leave the Training side empty for now
   2. On the Testing side add the Apply Model operator and the Performance (for Regression) operator
   3. Click on the Performance operator. On the Parameters tab place check marks next to the performance metrics "root mean squared error" and "squared correlation"
4. Make sure to connect everything in your process
5. Set Volume to "label"

**Out of the Box**

To begin our modeling, we run each algorithm with default parameter settings. This is called "Out of the Box."Later, during model optimization we will change the default settings to "tune" our models for better fit.

1. Open the Cross Validation operator. Place and connect one of the algorithms listed above. Let’s use KNN first.
2. Click "run" to run the classifier "Out of the Box".
3. Inspect and document the resulting performance metrics. Locate the PerformanceVector Tab and specifically the RMSE and R-Squared values. If the RMSE is zero and R-Squared is 1 this might be a perfect correlation, but it might also be a sign of overfitting the model. In predictive analytics very high correlations might indicate Overfit; an Overfit model is a model that has learned the training data so well that it will not generalize well to new data. You can read more about Overfit in the resources.

**Optimization**

**Find the best performance point for the model.** Your goal is to find the combination of parameter settings for the model which results in the least level of comparative error and the highest R-Squared. The optimal performance for this model with this specific data set could be the first or the last thing that you try. There is no way of knowing which will be the best before you start—it is essential that you vary the parameter values and compare the error metrics until you see no further improvement.

**To get a handle on this "optimization" process, start by experimenting with the number of neighbors (this is specific to KNN)**:

1. Open the Cross Validation Operator and click on the KNN operator. In the parameters tab, note that "Out of the Box" meant that the value of K was 1. Change this value to 2 and run the algorithm again.
2. Compare the performance metrics of 2 and 1. Is using 2 neighbors in the model better than 1 neighbor or worse? If performance is worse with 2 neighbors, then you will retain 1 as the number of neighbors and try the next setting. If 2 is better than 1, then you will try 3 neighbors and compare 3 to 2 and so on.
3. **Try tuning the next relevant parameter value and compare error metrics until you find the best setting**. Refer to the information on the classifier parameters in the ***Learning Resources*** page

**TIP:**

**Note: It is not unusual to see high error rates in a regression analysis.** When conducting a regression analysis, you will often see high error rates (in comparison to a classification analysis for example) because you are trying to predict a precise number. But that doesn’t mean that the algorithm is not getting it in the right range. The goal is to get the best result that you can give the data you have.

1. **Continue the process** until you have all the parameter values tuned as best you feel you can for this classifier.
2. **Repeat the steps** above for the remaining two models changing the values of 'C' for SVM and 'number of trees' for Gradient Boosted Trees. Consult the resources for the tuning parameters specific to each algorithm that can be used for tuning.

**Classifier Selection and Storing the Model**

You have trained and optimized three classifiers. Now, you will need to compare their performance metrics and choose the model that best "fits" the data. When making your selection, be sure to note the associated parameters. You will need this information for your final report and to make predictions for the 24 new products. After selecting the most optimal model based on your previous training you will need to add to the same process to store the model so it can be used to make predictions on the for the 24 new products.

1. Add a "Store" operator after the Cross Validation operator. The Store operator saves the last model that was built so that it can later be used for predictions. You will need name this stored file and select a location in which to save it in the "repository entry" field.
2. After you have connected the necessary inputs and outputs, click Run.

#### **[7. Use the Model to Make Predictions](https://codeacademy.ubiqum.com/mc/poa?productID=6637&taskID=4242" \l "collapsepoa12332)**

In the Modeling and Optimization step you trained and stored the classifier that best predicted the value of the dependent variable Volume. You will now use this model and the newProductAttributes.csv file to make sales volume predictions for Blackwell Electronics' 17 new products.

1. Open the Design View and disable all of your current operators. Don’t delete any as you may need to repeat earlier processes.
2. Import the newProductAttributes.csv file and then drag it into the design process. Recall that any changes you have made to the training data (existingProductAttributes.csv) you must also make identically to this file. With RapidMiner you can easily do this by copying and pasting the filters from the training set into your new process.
3. In the modeling step you added a Store operator to save the last model that you created. To access this saved model and make predictions you need to drag the model from Repository into the design process.
4. Insert the Apply Model operator into your process and connect all ports.

1. Click Run and then inspect the example set tab of the results. Find the prediction column and note the predicted Volume for all 17 new products.
2. You are now ready to write your report.

#### **[8. Report your Findings](https://codeacademy.ubiqum.com/mc/poa?productID=6637&taskID=4242" \l "collapsepoa12334)**

The following steps will help you complete your report for Danielle.

1. Open the profitability.csv file you downloaded at the beginning of this project
2. Enter the volume predictions you made in the Volume column
3. Calculate predicted profit for each potential new product with the following formula: volume x profit margin x price = profit
4. Rank the 17 products by profit from highest to lowest. Highlight the top five products.
5. Write a brief summary of the optimized model you selected and your rationale for selecting it. Include the parameter settings for this model.
6. Include a summary of performance metrics for each of the classifiers you ran
7. Save and submit your report