IBM **Services**

Modeling and scoring response based on a retail marketing campaign

Contents

Part of any successful targeted marketing campaign is identifying those customers most likely to respond, so that you can focus your efforts where they can bear fruit. Using the visual modeling capabilities in IBM® Watson® Studio Desktop and Studio Cloud, you can build a predictive model that uses data on past customer behavior to help identify where to focus in the future.

The figures in this solution brief[1] as described in the following steps are an application to model the potential response to a retail-targeted marketing campaign. The examples can help you identify the relationship between recency of purchase and propensity to respond.

This example records the response or non-response of a sample of customers targeted with a specific campaign. Included are factors associated with the response to this campaign based on various customer value profiles. The example illustrates a method of scoring customers for their propensity to respond to similar future campaigns.

Prior to applying modeling methods in the flow, the following preprocessing steps occur:

- Calculating customer spending from transaction data
- Merging customer database and geodemographic data Creating RFM (recency, frequency, monetary) responsiveness scores.
- Calculating customer purchase patterns and spending summaries
- Preparing data for response modeling

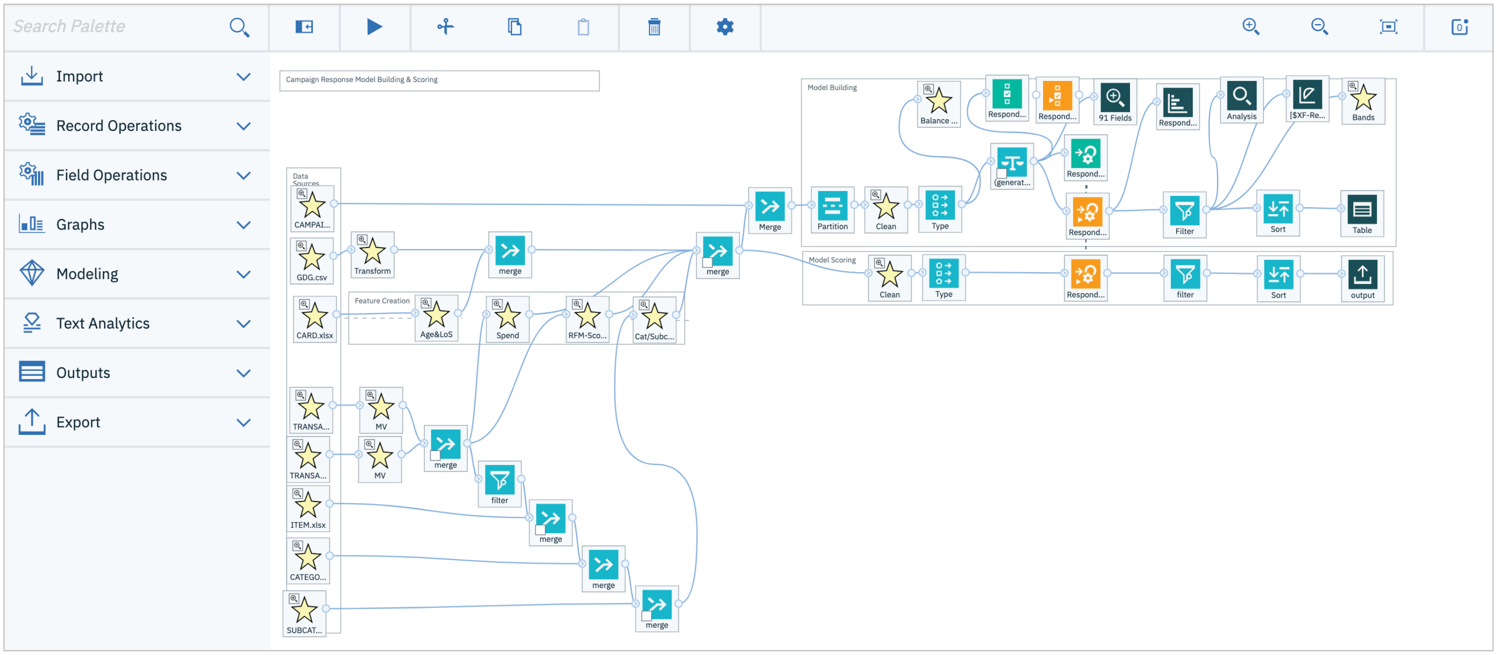


Figure 1. Flow as shown in the SPSS Modeler flow editor after opening file named ResponseModelBuildScore.str

Raw data sets	The function of each raw data set
CARD.xlsx	Customer information based on a preferred card or loyalty card
GDG.csv	Geodemographic information keyed on postal code
TRANSACTION.xlsx	Transaction data at a “basket” level—one transaction may include several purchases
TRANSACTION_ITEM.xlsx	Transaction data at the level of individual purchases
ITEM.csv	Table of all items on sale, joining item codes in the transaction data with product category, subcategory and brand data
CATEGORY.csv	Mapping of product category codes to descriptions
SUBCATEGORY.csv	Mapping of product subcategory codes to descriptions
CAMPAIGN.csv	List of customers targeted for a previous campaign with response or non-response flagged

Table 1. The following raw data sets are available for download.

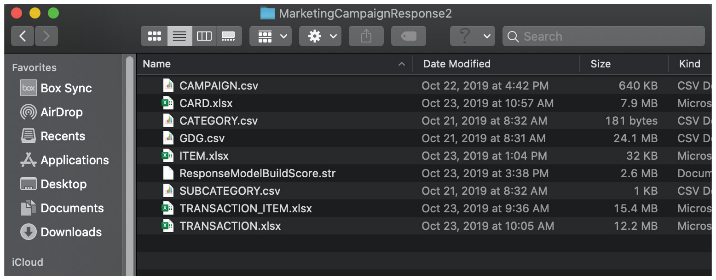
The next step is the application and evaluation of modeling the imbalanced campaign response using the prepared data. The flow consists of two branches, starting with an upper branch used to create a model based on campaign results. The lower branch takes these results to illustrate the scoring of the following raw customer data sets. These preprocessing and modeling steps work directly from the following raw data files.

Data sets

Shown in the left side of the flow in the previous chart, the raw data sets are available for download.

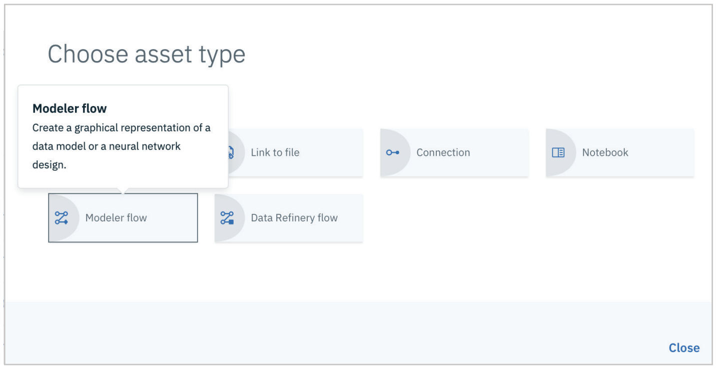
Take the following steps to plug data sources into the flow from a project:

1. Upload all assets into a compressed file to your project by typing *Browse* under the load tab on right of the project screen and uploading the uncompressed version of the files.



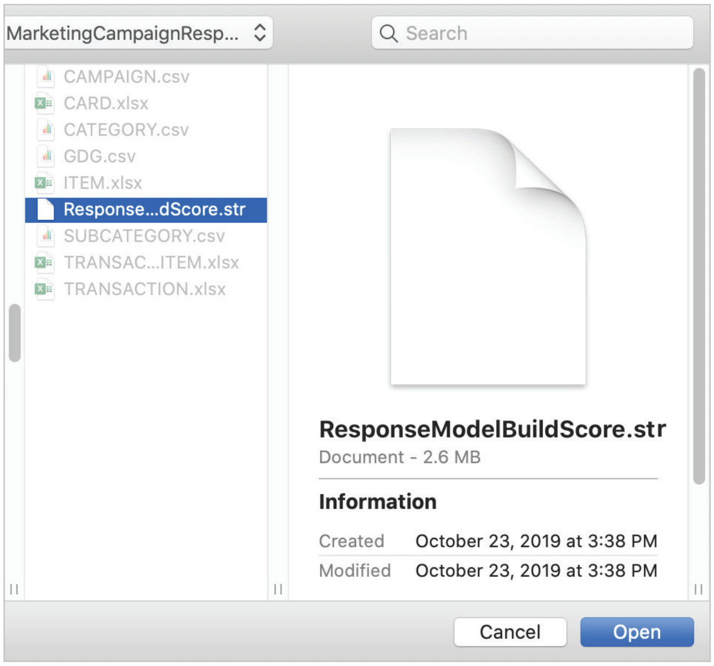
Name	Date Modified	Size	Kind
CAMPAIGN.csv	Oct 22, 2019 at 4:42 PM	640 KB	CSV D
CARD.xlsx	Oct 23, 2019 at 10:57 AM	7.9 MB	Micros
CATEGORY.csv	Oct 21, 2019 at 8:32 AM	181 bytes	CSV D
GDG.csv	Oct 21, 2019 at 8:31 AM	24.1 MB	CSV D
ITEM.xlsx	Oct 23, 2019 at 1:04 PM	32 KB	Micros
ResponseModelBuildScore.str	Oct 23, 2019 at 3:38 PM	2.6 MB	Docum
SUBCATEGORY.csv	Oct 21, 2019 at 8:32 AM	1 KB	CSV D
TRANSACTION_ITEM.xlsx	Oct 23, 2019 at 9:36 AM	15.4 MB	Micros
TRANSACTION.xlsx	Oct 23, 2019 at 10:05 AM	12.2 MB	Micros

2. From the project screen, select *Add to project.*
3. Choose *Modeler flow* from the menu.


Choose asset type

Modeler flow
Create a graphical representation of a data model or a neural network design.

4. On the *From file* tab, select *Browse.*
5. Choose the .str file from your file browser.

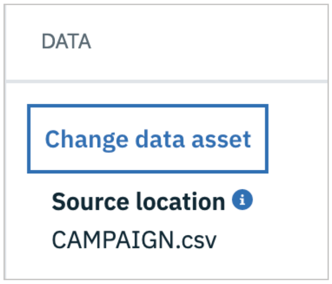


ResponseModelBuildScore.str
Document - 2.6 MB

Information

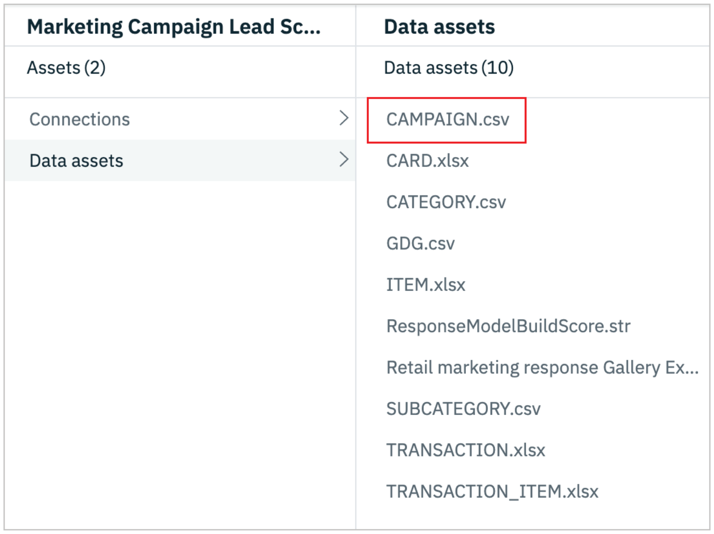
Created October 23, 2019 at 3:38 PM
Modified October 23, 2019 at 3:38 PM

6. From the project screen, launch the modeler flow.
7. Under the *Data sources* section, view each SuperNode and open the resulting data source node.



Source location
CAMPAIGN.csv

8. Choose *Change data asset* on the right and point to the corresponding data asset of the same name found in your project.



9. Click *OK*.
10. Repeat for each data source file.

SuperNodes

The flow includes SuperNodes, which conceal node operations that would otherwise clutter up the flow. The SuperNodes focus on a specific task and are grouped together. The primary SuperNodes of interest are those that create features used for predictive modeling. These SuperNodes are grouped in the box labeled “Feature Creation” in the flow.

For example, one creates RFM scores based on transaction data pertaining to recency, frequency and monetary value. Another SuperNode uses the transaction data to create a “basket-style” purchase pattern for each customer, where the number of items purchased in each product category and subcategory is noted. The Age&LoS SuperNode builds a consolidated customer description including demographic and geodemographic data keyed on postal code. The Spend SuperNode works with spending information based on the customer card database and transaction information in summary form.

The following descriptions apply to additional SuperNodes:

Age&LoS SuperNode
In this SuperNode, the card data is enhanced with an Age field calculated from the date of birth using a Derive node and an LoS or length of service field calculated from card start date.

Spend SuperNode
The TRANSACTION data containing the card ID merges with the TRANSACTION_ITEM data containing the amount spent and aggregates to produce the following four fields:

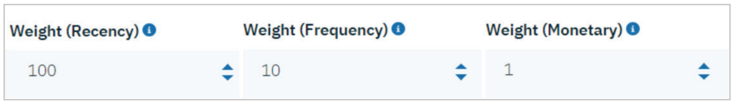
– Total spend
– Average spend on a single item
– Maximum spend on a single item
– Number of items purchased

The merged card and GDG data then merges with the aggregated transaction data using CardID as a key field. This merge is set to include incomplete records so that customers with no spending can be identified.

The enhanced card data in Age&LoS is then merged with the geodemographic table GDG.dat, using postal codes as the key field.

The Clean SuperNode discards customers absent from the card data from the results and fills spend-related fields with zeros for customers with no purchases.

RFM SuperNode
This SuperNode uses the RFM nodes to assist in the prediction of responsiveness to campaigns. RFM scoring combines customers' recency in making purchases, their frequency in doing so and the monetary amount they spend to produce a score often found to relate to responsiveness. The SuperNode keeps the value of each of these three aggregated fields separated and allows for adjustable weighting of each influence. Some users choose recency for the highest weighting and monetary for the lowest. The default values provided by IBM Watson Studio Desktop have an order of magnitude difference between the three of 100, 10 and 1 as shown.



Weight (Recency)	Weight (Frequency)	Weight (Monetary)
100	10	1

Figure 2. The RFM score field allows for adjustable weighting for recency, frequency and monetary activities of customers.

The result in this formulation of RFM is a single numeric score, the RFM score field.

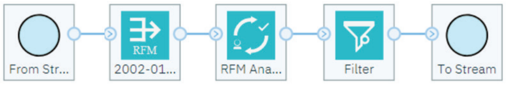


Figure 3. The RFM score SuperNode includes these processes.

The RFM nodes include the RFM Aggregate node and the RFM Analysis node.

Category and subcategory counts and Clean SuperNodes
These SuperNodes create purchase profiles for each customer by indicating the number of items purchased in each category and subcategory. This activity occurs to describe customers' behavior before the campaign under analysis, so that you can analyze the relationship between customers' prior behavior and their propensity to respond.

For each customer, a purchase profile shows how many items of each product category and subcategory were bought. This stream produces a profile through a series of data manipulations. For each item purchased, a merge attaches the category and subcategory descriptions, then a set-to-flag operation turns these descriptions into fields, one for each category and subcategory. These new fields are then populated with a count, generated by aggregation. Finally, the Clean SuperNode discards records for unrecognized cards, and zeros appear for all card or time-period combinations with no purchases. On a more detailed level, the following steps occur to produce a record for each item purchased.

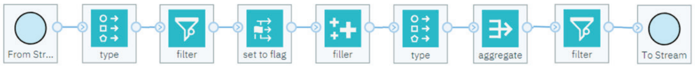


Figure 4. The category and subcategory counts SuperNode, also known as "cats and subcats SuperNode," works with these steps.

Balance SuperNode

As shown in Figure 5, this SuperNode creates the ratio to insert in the Balance node to correct the imbalance in the populations of the two classes: Respond = T & F. There are many methods for creating an “even” balance between classes and over-sampling, as used in this flow, is only one. You can experiment with under-sampling of the majority class or a mixture of both over- and under-sampling. The Synthetic Minority Over-sampling Technique or SMOTE node is another approach to create synthetic data near the border of the classes. Alternatively, you can use individual nodes that accommodate imbalanced data directly with certain parameter settings, such as XGBoost and random trees.

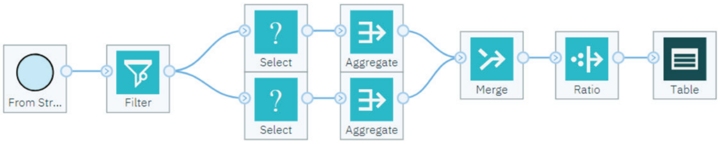


Figure 5. The Balance SuperNode creates the ratio as shown in this process.

Transform SuperNode

Using the Transform node with the Data Audit node, this SuperNode indicates which transformations are normalizing as measured by reductions in skewness and kurtosis. Kurtosis involves the tails of the graph of a frequency distribution. These results are important for some multivariate statistical modeling methods.[1] Figure 6 illustrates how this process occurs.

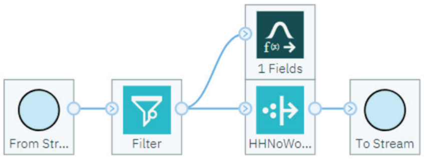


Figure 6. The Transform SuperNode works with these steps.

Note that SuperNodes must begin with a single node connection and end with a single node connection or a terminal node. For this reason, you can use a Filter node without any setting changes as the initial node in a SuperNode, which immediately branches to one or more nodes.

Model building branch

The uppermost branch of the flow shows the model building section, which begins with the campaign data from a previously obtained marketing response exercise that recorded responders and non-responders. If only responders were recorded, then this exercise wouldn’t be possible, since there would be no criteria by which a model could discern between responders and non-responders.

Immediately after the campaign data merges with the features engineered in the bottom of the flow, a Partition node can train a model with a portion of the data. In Figure 7, 70 percent of the data can train a model and the holdout data or remaining 30 percent can test its performance.

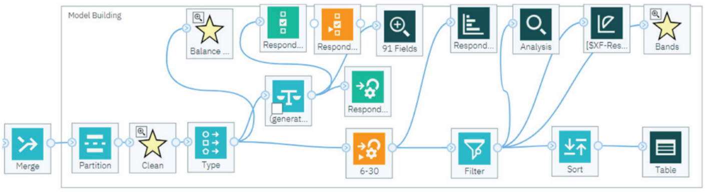
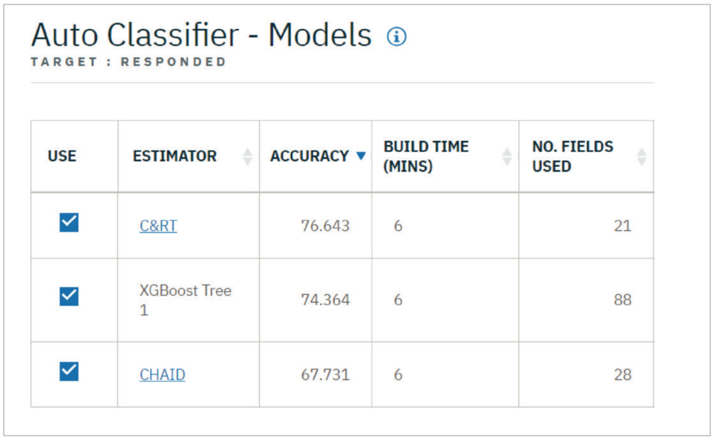


Figure 7. The model building section of the flow is set up to train the model using 70 percent of the data merged from the campaign and feature engineered data.

A Feature Selection node appears with the created model or nugget to verify that no highly ranked fields will show up in any of the fields representative of the target field. This example is only included as a reminder that this quick and simple step requires no time to build a model.

Next, a Data Audit node helps you obtain an immediate readout of the number of fields that are included. The node subtracts three fields—Partition, CardID and Response—to find the total number of predictors. Some modeling algorithms, such as random forests and neural networks, use the square root of the number of predictors to start assigning the number of hidden fields. Random forests use the square root for the random number of fields for each tree, while neural networks use the square root for the number of nodes to use in the first hidden layer.

With default settings, the Auto Classifier node created an ensemble in the following example of the top three models selected using the Expert section. The output of the model indicates its rankings of fields by accuracy.



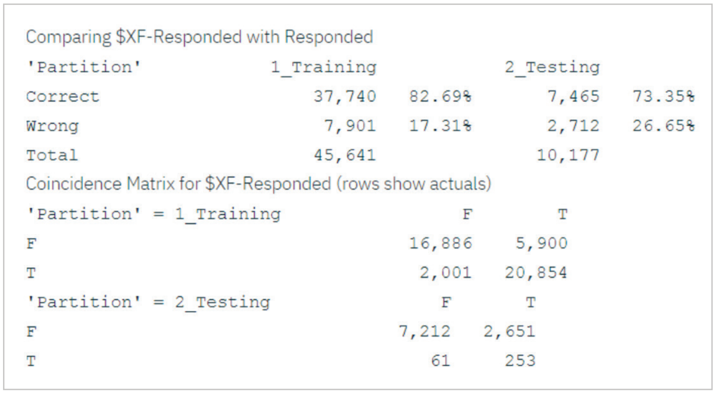
Auto Classifier - Models ⓘ

TARGET : RESPONDED

USE	ESTIMATOR	ACCURACY ▼	BUILD TIME (MINS)	NO. FIELDS USED
☑	C&RT	76.643	6	21
☑	XGBoost Tree 1	74.364	6	88
☑	CHAID	67.731	6	28

Figure 8. The Auto Classifier node shows the top three models selected.

For the testing partition, the true positive rate is fairly good at (253/314)* 100% = 80 percent. However, a sizable number of false positives yielded a true negative rate of 27 percent, as shown in Figure 9.



```
Comparing $XF-Responded with Responded
'Partition'             1_Training              2_Testing
Correct                     37,740   82.69%         7,465   73.35%
Wrong                        7,901   17.31%         2,712   26.65%
Total                       45,641                 10,177
Coincidence Matrix for $XF-Responded (rows show actuals)
'Partition' = 1_Training                        F        T
F                                          16,886    5,900
T                                           2,001   20,854
'Partition' = 2_Testing                       F        T
F                                         7,212    2,651
T                                            61      253
```

Figure 9. Testing results from the Analysis Node show the true negative rate.

Gains Bands SuperNode

The Gains node allows you to visualize the proportion of total hits that occurs in each quantile. Gains are computed as the proportion of hits in each increment relative to the total number of hits, using the equation: (hits in increment / total number of hits) multiplied by 100 percent. If you have a limited marketing budget for a campaign, you can sample from your population and leverage its prediction. For instance, you can market to the top 20 percent of the population and thereby include almost two-thirds of your predicted responders, as shown in Figure 10.

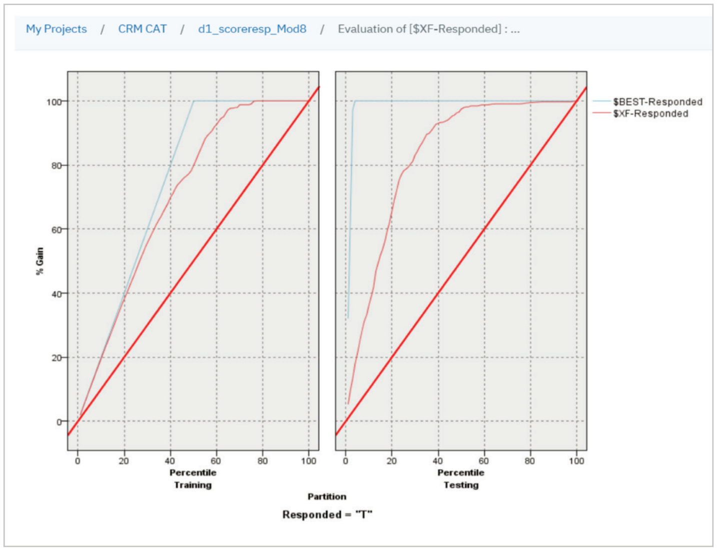


Figure 10. A visualization of the gains can be seen from the Evaluation of [$XF-Responded] node.

You can select these top 20% customers by using a Select node for the top two decile-binned (using a Binning node) on the prediction for the testing or holdout partition: 'XF-Test-Confidence_TILE10' = 9 or 'XF-Test-Confidence_TILE10' = 10. For the prior campaign, this would have resulted in campaigning to 2,178 customers.

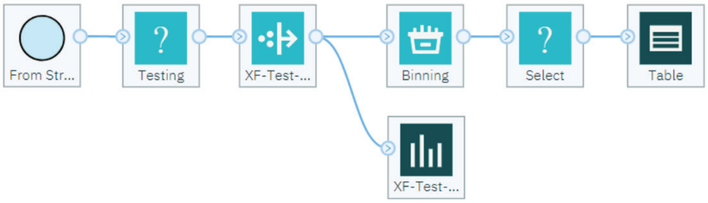


Figure 11. The Bands SuperNode follows this process.

For more information, see ibm.com/support/knowledgecenter/SSBFT6_1.1.0/wsd/nodes/evaluation.html

Scoring your data

After creating your predictive model for responders to a marketing campaign, you can score your customers by using the same data preparation that you already performed for the model building. This particular case doesn't exclude the customers that were in the previous marketing campaign, but this could be performed with an anti-join with the CAMPAIGN data. The results are written to an external file, as shown in Figure 12.

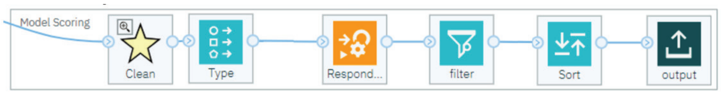


Figure 12. The model scoring section of the flow follows this pattern.

Summary

Many features can be created for predicting customer response to a marketing campaign. You can explore many other features and modeling methods with IBM Watson Studio Desktop.

About the author

Steve Barbee is Offering Manager Algorithms for IBM Cloud and Cognitive Software.

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
January 2020

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

[1] This example uses content from the Clementine® Application Template for Customer Relationship Management 7.0, Copyright © 2002 by Integral Solutions Limited. IBM acknowledges and appreciates the efforts of Tom Khabaza and the rest of the team that created the original content. Disclaimer: The data provided with the CRM CAT is based on a fictitious retail company selling consumer electronics. However, the techniques are applicable to a wide range of industries. The data is entirely synthetic and bears no relation to any real company.