Paper 1014-2017

Modeling the Merchandise Return Behavior of Anonymous and Non-Anonymous Online Apparel Retail Shoppers By Using the SAS® System and SAS® Enterprise Miner™ 13.2

Sunny Lam, ANN Inc., New York, NY

ABSTRACT

This paper establishes the conceptualization of the dimension of shopping cart (or market basket) on apparel retail websites. It analyzes how the cart dimension (describing anonymous shoppers) and customer dimension (describing non-anonymous shoppers) impact merchandise return behavior. Five data-mining techniques – namely logistic regression, decision tree, neural network, gradient boosting and support vector machine – are used for predicting the likelihood of merchandise return. The target variable is a dichotomous response variable: return vs not return. The primary input variables are conceptualized as the constituents of the cart dimension, derived from engineering merchandise-related variables such as item style, item size and item color, as well as free-shipping-related thresholds. By further incorporating the constituents of the customer dimension such as tenure, loyalty membership and purchase histories, the predictive accuracy of the model built using each of the five data-mining techniques was found to improve substantially. This research also highlights the relative importance of the constituents of the cart and customer dimensions governing the likelihood of merchandise return. Recommendations on possible applications and research areas are provided.

KEYWORDS

Cart Dimension, Customer Dimension, Market Basket, Shopping Cart, Merchandise Return Behavior, Likelihood of Return, Descriptive Analysis, Predictive Analysis, Variables Engineering, Anonymous Shoppers, Non-Anonymous Shoppers

INTRODUCTION

In the retail industry, market basket analysis is a common data-mining technique utilized in analyzing multi-item purchases (Bao, X., 2007; Faron, M. and Chakraborty, G., 2012; Redlon, M., 2003). The traditional business rationale centers on the notion of "how likely a customer would buy item B if she also bought item A?". While this direction of analysis is important, it is rarely seen any past research taking the merchandise return pattern into considerations. This paper is about examining how likely a customer would return a purchased item given she had multiple items in the same basket. In fact, the return rate in the apparel industry is usually over 20%, and could even exceed 70% for certain assortment/customer segments; apparel retailers may want to minimize the return rate in order to improve net sales, and reduce re-stocking administrative costs as well. Therefore, it is imperative for apparel retailers to gain better knowledge on the merchandise return behavior. This paper serves to outline a methodological framework to obtain such knowledge, and it is to be achieved via four objectives:

- To propose a cart dimension coding scheme to portray the compositions of the shopping cart and visualize the return behavioral variations with respect to different compositions.
- To use the shopping cart dimension to build a data-mining model to predict return behavior.
- To examine if the customer dimension can help improve the model predictive accuracy.
- To investigate the relative importance of variables influencing the likelihood of return.

Finally, this paper also discusses possible areas for business applications and further research.

METHODOLOGY

The next section Data Management will describe the data source and provide a thorough illustration in compiling and processing the datasets required for our analysis. Thereafter, the Model Building section will detail the formulation and interpret the results of the descriptive as well as predictive analysis. All the data manipulation and data-mining routines are carried out using BASE SAS® via SAS Enterprise Guide® and SAS® Enterprise Miner™ 13.2. The followings outline the methodological framework.

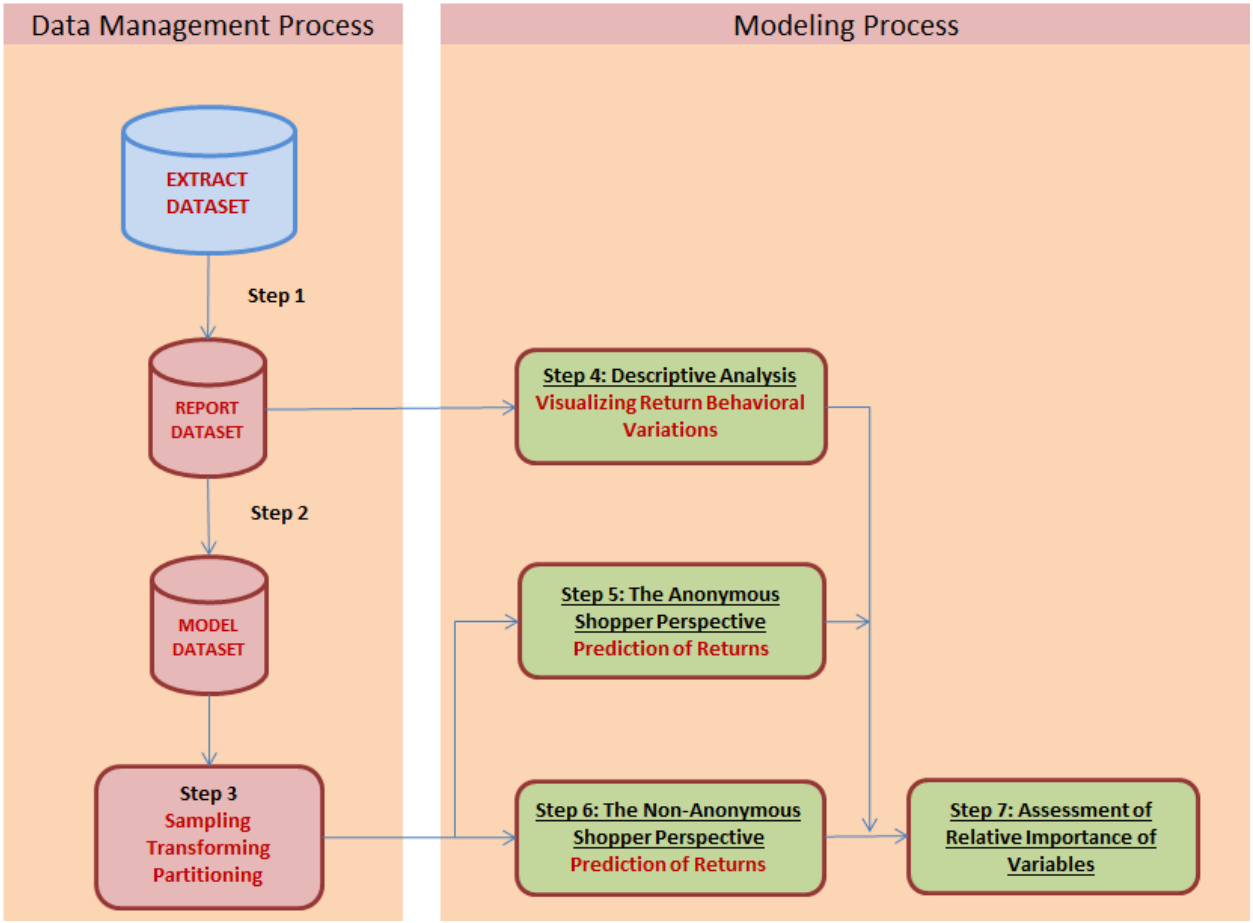


Figure 1: The Methodological Framework

In Step 1, a series of Proc SQL procedures and data steps are used to reformat a pre-extracted dataset: EXTRACT_DATASET to another dataset: REPORT_DATASET. This is a process to engineer merchandize variables; to recode them to form the cart dimension. The re-coding mechanism will be illustrated using sample observations.

In Step 2, the observations on REPORT_DATASET are filtered (based on our merchandise category under studied) and merged with customer variables (i.e. constituents of the customer dimension) to form another dataset: MODEL_DATASET. This new dataset will be loaded to SAS® Enterprise Miner™ 13.2 for subsequent processing – Step 3: data sampling, data transformation, and data partitioning.

In Step 4, the dataset: REPORT_DATASET is used to generate two summary reports, and these two reports serve to provide a descriptive analysis of the return behavioral variations across different cart dimension settings.

In Step 5, five different data-mining techniques – logistic regression, decision tree, neural network, gradient boosting and support vector machine – are utilized to predict the likelihood of return. Only the constituents of the cart dimension are used as input variables in this step.

In Step 6, we extend the process in Step 5 with the inclusion of the constituents of the customer dimension as input variables. The purpose is to see how these additional variables contribute to the model predictive accuracy.

Lastly, Step 7 examines the relative importance of all input variables in determining return likelihood.

DATA MANAGEMENT

The data is originated from the customer information data-mart of an online retailer which markets varieties of apparel merchandise categories. For confidentiality reason, all figures, dataset names, variable names and their associated values presented in this paper have been masked. We also limit our analysis on those shopping carts that checked out on a pre-determined day, and each cart contained at least one item under an undisclosed merchandise category. This paper will simply use "sweaters" as an operational description for this undisclosed category. The results and suggestions as highlighted in this paper do not imply any genuine business adaptation by this online retailer.

The retailer offers free shipping for all orders at $150 or more, and a 60-day grace period for returns at no penalty and, depending on individual circumstances, also grants credits after the 60-day grace period, to their customers. The data used in this study was extracted at a time well after 60 days following the pre-determined day, and by then all associated returns should have been captured.

Step 1: Variables Engineering for Compiling the Shopping Cart Dimension – The Description of the Anonymous Shoppers

Two data tables (a product lookup table and an item transaction table) from the customer information data-mart were pre- extracted, merged and transformed using BASE SAS via SAS Enterprise Guide® (the detailed SAS code is not shown in this paper) to form a dataset: EXTRACT_DATASET. There are 36,142 sweater related shopping carts totaling 174,843 item lines. For illustration purposes, the item lines of two shopping carts are depicted in Table 1. One shopping cart (Cart_ID = 10) consists of 2 sweater items and 4 non-sweater items[1] with a combined basket amount of $131. The other shopping cart (Cart_ID = 160) consists of 3 sweater items (with no non-sweater item) totaling $82. Both carts are not qualified for free shipping.

In Step 1, the 174,843 resulting item lines were transformed and stored using another dataset structure. This is the process for engineering merchandise variables; converting observations to be represented in the Cart Dimension format. The re-coded item lines of the two shopping carts are shown in Table 2. Appendix A displays the SAS code of such data transformation.

Refer to Cart_ID = 10 in Table 1. There are two sweater items with Product_Code = 177094 and 177320. They are of the same style (Style_Code = 48579) and same size (Size_Code = 117) but different colors (Color_Code = 1846 and 142). Therefore, we can interpret as shown in Table 2 that each of the two items is sitting in the same basket that includes "One Style" (Style_Count = 1) of sweater, and the style of sweater has only "One Size" (Size_Count = 1) but encompasses two colors (i.e. Color_Count = 2 or "Multi Color").

For Cart_ID = 160, there are three sweater items with Product_Code = 177927, 178350 and 178456. These three items are of two different styles (Style_Code = 51039 and 49483). One interpretation is that the item with Product_Code = 177927 is sitting in the basket that includes "Multi Style" (Style_Count = 2) of items; and the style (Style_Code = 51039) of this item has only "One Size" (Size_Code = 1) as well as "One Color" (Color_Count = 1). As another interpretation, the item with Product_Code = 178350 is sitting in the basket that includes "Multi Style" (Style_Count = 2) of sweater items; and the style (Style_Code = 49483) of this item has "Multi Size" (Size_Code = 117 and 113; Size_Count = 2) as well as "Multi Color" (Color_Code = 2200 and 1967; Color_Count = 2). Similar coding interpretation can be derived for the item with Product_Code = 178456.

[1] The item lines of these four non-sweater items are transformed in the same fashion as that of those sweater items, but be noticed that these non-sweater items are not our focus of analysis.

Table 1: Shopping carts including at least one sweater as an ordered item (SAS dataset: EXTRACT_DATASET)

Customer ID	Transaction ID	Product Code	Item Quantity	Item Amount	Return Quantity	Return Amount	Style Code	Color Code	Size Code	Department Code	Cart ID	Product Class	Record Number
3650000038	25355847	176115	1	$34.8	1	$34.8	51075	140	37	55	10	Non-Sweater	52
3650000038	25355847	177094	1	$24.8	1	$24.8	48579	1846	117	9	10	Sweater	53
3650000038	25355847	177320	1	$24.8	1	$24.8	48579	142	117	9	10	Sweater	54
3650000038	25355847	177940	1	$14.8	1	$14.8	51011	2200	167	26	10	Non-Sweater	55
3650000038	25355847	178391	1	$9.8	0	$0.0	51126	1633	117	21	10	Non-Sweater	56
3650000038	25355847	178762	1	$22.3	0	$0.0	50465	966	117	21	10	Non-Sweater	57
		Sweater:	2	$49.5									
		Non-Sweater:	4	$81.5									
		Total Cart:	6	$131.0									
3650000140	25355850	177927	1	$27.5	0	$0.0	51039	312	117	9	160	Sweater	792
3650000140	25355850	178350	1	$27.3	0	$0.0	49483	2200	117	9	160	Sweater	793
3650000140	25355850	178456	1	$27.3	1	$27.3	49483	1967	133	9	160	Sweater	794
		Sweater:	3	$82.0									
		Non-Sweater:	0	$0.0									
		Total Cart:	3	$82.0									

Table 2: Transformed item lines in the Cart Dimension format (SAS dataset: REPORT_DATASET)

Customer ID	Product Code	Style Count	Size Count	Color Count	Sweater Cart Qty	Non Sweater Cart Qty	Sweater Cart Amt	Non Sweater Cart Amt	STYLE	SIZE	COLOR	Record Number
3650000038	176115	4	1	1	2	4	$49.5	$81.5	Multi Style	One Size	One Color	52
3650000038	177094	1	1	2	2	4	$49.5	$81.5	One Style	One Size	Multi Color	53
3650000038	177320	1	1	2	2	4	$49.5	$81.5	One Style	One Size	Multi Color	54
3650000038	177940	4	1	1	2	4	$49.5	$81.5	Multi Style	One Size	One Color	55
3650000038	178391	4	1	1	2	4	$49.5	$81.5	Multi Style	One Size	One Color	56
3650000038	178762	4	1	1	2	4	$49.5	$81.5	Multi Style	One Size	One Color	57
3650000140	177927	2	1	1	3	0	$82.0	$0.0	Multi Style	One Size	One Color	792
3650000140	178350	2	2	2	3	0	$82.0	$0.0	Multi Style	Multi Size	Multi Color	793
3650000140	178456	2	2	2	3	0	$82.0	$0.0	Multi Style	Multi Size	Multi Color	794

The constituents of the Cart Dimension are at this point conceptualized as follows:

- STYLE – This refers to either One Style or Multi Style
- SIZE – This refers to either One Size or Multi Size
- COLOR – This refers to either One Color or Multi Color
- Cart Shipping – This refers to the qualification for free shipping (i.e. whether the total sweater plus non-sweater items amount to $150 or more)
- Sweater Shipping – This refers to the qualification for free shipping just by purely sweater items (i.e. whether the total sweater items amount to $150 or more).

These five constituents will be used to formulate our descriptive analysis (Step 4), and they will be modified (e.g. with the inclusion of the sweater item dollar amount) in Step 2 for further predictive analysis (see Table 3) in Step 5 and Step 6.

The first cart (Cart_ID = 10) involves only one Cart Dimension profile (call it cart profile for the sake of simplicity):

- One Style, One Size, Multi Color, Not Qualified for Free Shipping, Sweater Items Alone Not Qualified for Free Shipping.

The second cart (Cart_ID = 160) involves two cart profiles:

- Multi Style, One Size, One Color, Not Qualified for Free Shipping, Sweater Items Alone Not Qualified for Free Shipping.

- Multi Style, Multi Size, Multi Color, Not Qualified for Free Shipping, Sweater Items Alone Not Qualified for Free Shipping.

The above two carts altogether include 5 sweater items. They are, under the current formulation, treated as independent observations. We would like to examine how likely each item, when represented by any of the 3 cart profiles, will be returned by the (anonymous) shopper. It is important to note from the two carts that the characteristics of an observation is determined by what other items are inside the same cart, not the Product Code of that item. Two observations could be treated as having the same cart profile even their Product Codes are different, as is the case for the first cart. On the other hand, the same item (Product Code), when sitting in different carts, could be treated as having different cart profiles.

All variables in the EXTRACT_DATASET are also retained in the REPORT_DATASET, but are not displayed in Table 2 in order the fit the page. This new dataset will be used to compile two summary reports for descriptive analysis in the next section (Step 4). The descriptive analysis aims at profiling different Cart Dimension settings to explore the return behavioral variations, if any.

Step 2: Appending Customer Dimension – The Description of the Non-Anonymous Shoppers

All the sweaters related observations in the REPORT_DATASET (i.e. including Record Number 53, 54, 792, 793 and 794 in Table 2) were filtered and merged with customer variables (some are essentially customer behavioral variables) to form another dataset: MODEL_DATASET, for further data processing and modeling (see Figure 7 for the complete process flow diagram). This merged dataset primarily stores the cart and customer dimensions. Some other variables (e.g. item amount and total cart amount) have been added, and some naming conventions have been modified. The below table is a complete summary description of the dataset.

Table 3: Variable descriptions of the modeling dataset: MODEL_DATASET

Variable	Descriptions	Scale
	Cart Dimension (Input)	
Dim_Style	One Style / Multi Style	Categorical
Dim_Size	One Size / Multi Size	Categorical
Dim_Color	One Color / Multi Color	Categorical
Dim_Cart_Shipping	Free / Charged	Categorical
Dim_Sweater_Cart_Amt	Total Sweater Dolloar Amount on the Cart	Numeric
Dim_Total_Cart_Amt	Total Sweater and Non-Sweater Dollar Amount on the Cart	Numeric
Dim_Item_Amount	The Dollar Amount associated with the Item	Numeric
	Customer Dimension (Input)	
Customer_Membership	Member / Non-Member	Categorical
Customer_Tenure	The Number of Days Lapsed since the First Purchase	Numeric
RFM_12M_Frequency	The Total Number of Purchases in the Past 12 Months	Numeric
RFM_12M_Monetary	The Total Dollar Amount Spent in the Past 12 Months	Numeric
RFM_Rtn_12M_Frequency	The Total Number of Returns in the Past 12 Months	Numeric
RFM_Rtn_12M_Monetary	The Total Dollar Amount Returned in the Past 12 Months	Numeric
	Target	
Return	1 - Returned / 0 - Not Returned	Categorical
	Adminstrative Variable	
Cart_ID	Identifier of the Cart	Numeric
Customer_ID	Identifier of the Customer	Numeric
Product_Code	Product Identifier Determined by Style_Code, Size_Code and Color_Code	Numeric
Record_Number	Unique Identifier of an observation	Numeric

Step 3: Data Processing – Using SAS® Enterprise Miner™ 13.2

The dataset: MODEL_DATASET is loaded to SAS® Enterprise Miner™ 13.2, and this sub-section illustrates the subsequent application of different data processing task nodes on this dataset.

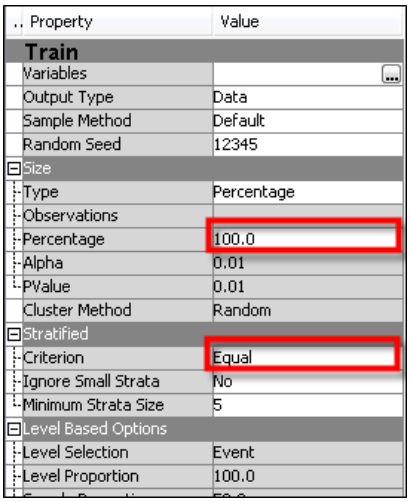
Data – When creating the data source using SAS® Enterprise Miner™ 13.2, one important step is to fill in the Column Metadata in the Data Source Wizard. Figure 2 displays the setting for the metadata. We only retain the Record Number as the observation identifier and set its Role as "ID". The other three administrative variables are dropped, by setting Drop as "Yes". All input variables have their Roles to be set as "Input". Categorical and numeric input variables have their Levels to be set as "Nominal" and "Interval" respectively. The target variable Return is set as "Target" for its Role and "Binary" for its Level.

Name	Role	Level	Drop
Cart_ID	ID	Nominal	Yes
Customer_ID	ID	Nominal	Yes
Customer_Membership	Input	Nominal	No
Customer_Tenure	Input	Interval	No
Dim_Cart_Shipping	Input	Nominal	No
Dim_COLOR	Input	Nominal	No
Dim_Item_Amount	Input	Interval	No
Dim_SIZE	Input	Nominal	No
Dim_STYLE	Input	Nominal	No
Dim_Sweater_Cart_Amt	Input	Interval	No
Dim_Total_Cart_Amt	Input	Interval	No
Product_Code	Rejected	Interval	Yes
Record_Number	ID	Interval	No
Return	Target	Binary	No
RFM_12M_Frequency	Input	Interval	No
RFM_12M_Monetary	Input	Interval	No
RFM_Rtn_12M_Frequency	Input	Interval	No
RFM_Rtn_12M_Monetary	Input	Interval	No

Figure 2: Metadata setup

Sample – Drag in the Sample node and rename it as Sample: 50/50. Figure 3 shows the property panel setting for the Sample node, and the sampling results. Since the proportion of the two classes: Returned (Return = 1) Vs Not Returned (Return = 0) is somewhat imbalanced at around 1-to-3, under-sampling[2] of the majority class (Not Returned) is used to equalize the two classes in an attempt to improve the predictive accuracy (Chawla, N., 2005). The final sample size used in this study is 15,212 + 15,212 = 30,424.

[2] The analysis focuses are on model comparison and examination of the relative importance of variables (Step 5, 6 and 7). No new observations are to be scored. As such no subsequent final probability adjustment (i.e. undo- under-sampling) is performed.



Property	Value
Train	
Variables	
Output Type	Data
Sample Method	Default
Random Seed	12345
Size	
Type	Percentage
Observations	
Percentage	100.0
Alpha	0.01
PValue	0.01
Cluster Method	Random
Stratified	
Criterion	Equal
Ignore Small Strata	No
Minimum Strata Size	5
Level Based Options	
Level Selection	Event
Level Proportion	100.0

Summary Statistics for Class Targets
(maximum 500 observations printed)

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent
Return	0	0	46250	75.2497
Return	1	1	15212	24.7503

Data=SAMPLE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent
Return	0	0	15212	50
Return	1	1	15212	50

Figure 3: Property panel of the Sample node and the sampling results

Transform Variables – Drag in the Transform Variables node and connect it with the Sample node; and then right click the Transform Variables node and choose “Edit Variables”. All interval input variables (as highlighted in Figure 4 below) are set as “Best” for the method of transformation. This “Best” method automatically tries all the available built-in transformation methods and picks the one that the input variable has the strongest R Squared relationship with the target variable. All nominal input variables as well as the target variable are set as “Default” (i.e. no transformation) for the method of transformation.

Name	Method	Number of Bins	Role	Level
Customer_Membership	**Default**	4	Input	Nominal
Customer_Tenure	**Best**	4	Input	Interval
Dim_COLOR	**Default**	4	Input	Nominal
Dim_Cart_Shipping	**Default**	4	Input	Nominal
Dim_Item_Amount	**Best**	4	Input	Interval
Dim_SIZE	**Default**	4	Input	Nominal
Dim_STYLE	**Default**	4	Input	Nominal
Dim_Sweater_Cart_Amt	**Best**	4	Input	Interval
Dim_Total_Cart_Amt	**Best**	4	Input	Interval
RFM_12M_Frequency	**Best**	4	Input	Interval
RFM_12M_Monetary	**Best**	4	Input	Interval
RFM_Rtn_12M_Frequency	**Best**	4	Input	Interval
RFM_Rtn_12M_Monetary	**Best**	4	Input	Interval
Return	**Default**	4	Target	Binary

Figure 4: Variables transformation setting

See the transformation results as displayed in Figure 5.

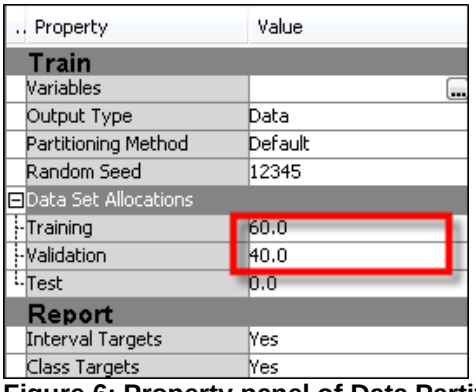
Computed Transformations
(maximum 500 observations printed)

Input Name	Role	Input Level	Name	Level	Formula
Customer_Tenure	INPUT	INTERVAL	SQRT_Customer_Tenure	INTERVAL	Sqrt(Customer_Tenure + 1)
Dim_Item_Amount	INPUT	INTERVAL	OPT_Dim_Item_Amount	NOMINAL	Optimal Binning(4)
Dim_Sweater_Cart_Amt	INPUT	INTERVAL	OPT_Dim_Sweater_Cart_Amt	NOMINAL	Optimal Binning(4)
Dim_Total_Cart_Amt	INPUT	INTERVAL	OPT_Dim_Total_Cart_Amt	NOMINAL	Optimal Binning(4)
RFM_12M_Frequency	INPUT	INTERVAL	SQRT_RFM_12M_Frequency	INTERVAL	Sqrt(RFM_12M_Frequency + 1)
RFM_12M_Monetary	INPUT	INTERVAL	SQRT_RFM_12M_Monetary	INTERVAL	Sqrt(RFM_12M_Monetary + 1)
RFM_Rtn_12M_Frequency	INPUT	INTERVAL	LG10_RFM_Rtn_12M_Frequency	INTERVAL	log10(RFM_Rtn_12M_Frequency + 1)
RFM_Rtn_12M_Monetary	INPUT	INTERVAL	LOG_RFM_Rtn_12M_Monetary	INTERVAL	log(RFM_Rtn_12M_Monetary + 1)

Figure 5: Results of variables transformation

SAS® Enterprise Miner™ 13.2 uses and arrives different transformations such as squared root, binning and logs (for base 10 and natural log) as shown above.

Data Partition – Drag in the Data Partition note and rename it as Data Partition 60/40. The dataset was split into Training and Validation, in the distribution of 60% and 40% respectively. See Figure 6 below for the property panel setting, and the partitioning results.



Summary Statistics for Class Targets

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Return	0	0	15212	50	
Return	1	1	15212	50	

Data=TRAIN

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Return	0	0	9126	49.9973	
Return	1	1	9127	50.0027	

Data=VALIDATE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Return	0	0	6086	50.0041	
Return	1	1	6085	49.9959	

Figure 6: Property panel of Data Partition, and the partitioning results

MODEL BUILDING

This section uses SAS data steps and Proc SQL procedures to construct descriptive analysis; and leverages SAS® Enterprise Miner™ 13.2 to build predictive models. The below summarizes their corresponding results and interpretations.

Step 4: Descriptive Analysis – Visualizing the Return Behavioral Variations

The dataset: REPORT_DATASET is used as the input dataset to construct Table 4. See the relevant SAS code in Appendix B. We paste the SAS output to a spreadsheet and perform some simple calculations. Not all output columns (e.g. revenue related) are reported in Table 4 in order to fit the page.

Table 4: Return rates by different cart dimension combinations

Cart	Overall Cart			Cart Dimension					Sweater Item Qty				Non Sweater Item		
Profile #	# of Cart		Cart Size	Cart Shipping	Sweater Shipping	STYLE	SIZE	COLOR	Avg. Qty	Ordered	Returned		# of Cart	% of Cart	Avg. Qty
1	24	0.1%	$94	Charged	Sweater $ < $150	Multi Style	Multi Size	Multi Color	3.44	84	16	19.0%	8	33.3%	0.50
2	23	0.1%	$97	Charged	Sweater $ < $150	Multi Style	Multi Size	One Color	3.41	78	23	29.5%	4	17.4%	0.24
3	192	0.5%	$90	Charged	Sweater $ < $150	Multi Style	One Size	Multi Color	3.23	618	112	18.1%	77	40.1%	0.57
4	3,961	11.0%	$77	Charged	Sweater $ < $150	Multi Style	One Size	One Color	2.25	8,932	1,578	17.7%	2,060	52.0%	0.85
5	116	0.3%	$74	Charged	Sweater $ < $150	One Style	Multi Size	Multi Color	2.29	266	63	23.7%	62	53.4%	1.12
6	113	0.3%	$72	Charged	Sweater $ < $150	One Style	Multi Size	One Color	2.02	230	96	41.7%	49	43.4%	0.65
7	668	1.8%	$69	Charged	Sweater $ < $150	One Style	One Size	Multi Color	2.16	1,446	216	14.9%	324	48.5%	0.87
8	11,551	32.0%	$58	Charged	Sweater $ < $150	One Style	One Size	One Color	1.01	11,610	1,709	14.7%	8,739	75.7%	1.44
9	144	0.4%	$205	Free	Sweater $ < $150	Multi Style	Multi Size	Multi Color	3.45	498	115	23.1%	144	100.0%	3.88
10	190	0.5%	$217	Free	Sweater $ < $150	Multi Style	Multi Size	One Color	3.40	647	270	41.7%	190	100.0%	4.13
11	888	2.5%	$215	Free	Sweater $ < $150	Multi Style	One Size	Multi Color	3.45	3,063	788	25.7%	888	100.0%	4.32
12	8,014	22.2%	$202	Free	Sweater $ < $150	Multi Style	One Size	One Color	2.55	20,459	5,314	26.0%	8,014	100.0%	4.34
13	198	0.5%	$211	Free	Sweater $ < $150	One Style	Multi Size	Multi Color	2.69	533	232	43.5%	198	100.0%	4.82
14	224	0.6%	$207	Free	Sweater $ < $150	One Style	Multi Size	One Color	2.05	460	271	58.9%	224	100.0%	4.75
15	907	2.5%	$204	Free	Sweater $ < $150	One Style	One Size	Multi Color	2.26	2,048	540	26.4%	907	100.0%	5.01
16	8,794	24.3%	$186	Free	Sweater $ < $150	One Style	One Size	One Color	1.01	8,890	2,406	27.1%	8,794	100.0%	5.26
17	166	0.5%	$320	Free	Sweater $ >= $150	Multi Style	Multi Size	Multi Color	6.66	1,106	464	42.0%	117	70.5%	4.20
18	186	0.5%	$313	Free	Sweater $ >= $150	Multi Style	Multi Size	One Color	6.08	1,133	583	51.5%	136	73.1%	4.28
19	510	1.4%	$297	Free	Sweater $ >= $150	Multi Style	One Size	Multi Color	5.89	3,006	1,044	34.7%	390	76.5%	4.27
20	1,365	3.8%	$284	Free	Sweater $ >= $150	Multi Style	One Size	One Color	5.35	7,295	2,445	33.5%	1,010	74.0%	4.00
21	20	0.1%	$296	Free	Sweater $ >= $150	One Style	Multi Size	Multi Color	5.93	120	43	35.8%	11	55.0%	2.67
22	4	0.0%	$168	Free	Sweater $ >= $150	One Style	Multi Size	One Color	4.33	18	4	22.2%	0	0.0%	0.00
23	24	0.1%	$246	Free	Sweater $ >= $150	One Style	One Size	Multi Color	5.83	142	20	14.1%	11	45.8%	4.00
24	1	0.0%	$170	Free	Sweater $ >= $150	One Style	One Size	One Color	2.00	3	0	0.0%	0	0.0%	0.00
Overall	36,142	105.9%	$145						1.75	63,323	15,310	24.2%	0	0.0%	3.17

There are 24 different cart profiles. The return rates (in terms of percentage of quantity returned) range from 0.0% to 58.9%. Consider Profile 12 for a general interpretation. Out of the 36,142 sweaters related checked out shopping carts, 8,014 (or 22.2%) of them include the combination of merchandise items that can be described by Profile 12. A sweater item sitting on this cart has at least one other sweater item (but in a different style) sitting together with it on the same cart. For each of them, the style has only one size and one color. All sweaters total less than $150 (and so sweaters alone is not qualified for free shipping) but the overall cart is qualified for free shipping. This implies that the cart must include non-sweater items. That is: 100% of the 8,014 carts have non-sweater items (see the last but second and third column). The average cart size is $202. The average quantities of sweaters and non-sweater items are 2.55 and 4.34 respectively. The return rate for this profile is 26.0%. The followings attempt to illustrate how the different combinations of cart dimension impact return.

Consider Profile 4 and Profile 12. The two cart profiles only differ in Cart Shipping (Profile 4 does not qualify for free shipping). The return rates of them are quite different (17.7% and 26.0% respectively). Similarly, Profile 8 and Profile 16 also only differ in Cart Shipping (Profile 8 does not qualify for free shipping), their corresponding return rates are also very different (14.7% and 27.1% respectively). These may suggest that free shipping would lead to higher return rate.

Consider Profile 12 and Profile 20. They only differ in Sweater Shipping (sweater items alone total $150 or more for Profile 20) but the return rates of them are somewhat different (26.0% and 33.5% respectively). Similarly, Profile 10 and Profile 18 only differ in Sweater Shipping (sweater items alone

total $150 or more for Profile 18). The return rates of them are also somewhat different (41.7% and 51.5% respectively). These may imply that when sweater items alone are enough for the basket to be qualified for free shipping (in-regardless of whether there is any non-sweater item in the cart), we would expect to see a higher return rate.

Consider Profile 7. It only differs from Profile 8 in terms of COLOR (Multi Color for Profile 7). Its return rate of 14.9% is similar to that of Profile 8 (14.7%). Similarly, Profile 15 is of Multi Color while Profile 16 is not, the return rate of them are also similar (26.4% and 27.1% respectively). These may imply that COLOR gives very minimal impact on the return rate.

Compare Profile 4 with Profile 8. They only differ in STYLE (Multi Style and One Style respectively), and the one in Multi Style records slightly higher return rate (17.7% versus 14.7%). Profile 12 and Profile 16 also differ in STYLE only, and in this case their return rates are very close (26.0% and 27.1% respectively). These may suggest that STYLE gives relatively lower impact, if any, on the return rate.

Compare Profile 14 and Profile 16. They only differ in SIZE (Multi Size and One Size respectively), but the return rates are substantially different (58.9% versus 27.1%). Similarly, Profile 6 and Profile 8 also only differ in SIZE, and substantial difference in return rates (41.7% versus 14.7%) are observed. These tend to suggest that SIZE impacts return substantially.

Table 5 below attempts to contrast the variable-level return rates for each of the 5 cart dimension variables, one at a time. The measure Diff% represents the percentage difference of the higher-level return rate versus the lower-level return rate for a given variable. One may intuitively imagine that a large Diff% value would suggest that the corresponding variable is important in determining the likelihood of return. The Diff% values are in the following descending orders:

Cart Shipping > Sweater Shipping > SIZE > STYLE > COLOR

Apparently, this sequence agrees with the descriptions that have just been illustrated with Table 4. More discussion on variable importance will be presented in Step 7.

Table 5: Return rates by aggregated cart dimension

Dim	Overall Cart			Cart Dimension		Sweater Item Qty				
Profile #	# of Cart		Cart Size	Variable	Level	Avg. Qty	Ordered	Returned		Diff%
1	16,424	45.4%	$80	Cart_Shipping	Charged	1.37	22,542	3,818	16.9%	66%
2	19,718	54.6%	$199		Free	2.07	40,781	11,492	28.2%	
3	34,638	95.8%	$139	Sweater_Shipping	Sweater $ < $150	1.60	55,250	12,435	22.5%	58%
4	1,504	4.2%	$269		Sweater $ >= $150	5.37	8,073	2,876	35.6%	
5	13,520	37.4%	$177	STYLE	Multi Style	2.78	37,558	9,709	25.9%	19%
6	22,622	62.6%	$126		One Style	1.14	25,765	5,601	21.7%	
7	1,364	3.8%	$197	SIZE	Multi Size	3.51	4,782	1,621	33.9%	52%
8	35,388	97.9%	$135		One Size	1.73	61,309	13,690	22.3%	
9	3,822	10.6%	$184	COLOR	Multi Color	3.32	12,693	2,500	19.7%	12%
10	34,116	94.4%	$135		One Color	1.71	58,324	12,810	22.0%	
Overall	36,142		$145			1.75	63,323	15,310	24.2%	

*The SAS code for compiling this table is similar to that was used for compiling Table 4 and is not to be replicated in this paper.

Apart from the two free-shipping-related variables, Table 5 implies that SIZE is most impactful to return. When a customer bought two different sizes for the same style of sweater, the chance of return would be increased by 52% as compared with that of another customer who bought only one size. One explanation is that a customer of this kind did not know her size well and as such put two different sizes of the same style to the shopping cart, and thereafter returned the one that less fit her. This behavior may even become motivational if adding one more sweater to the cart would make the shipping to be free. Cart Profile 14 could be a representative example of this kind and it has the highest return rate of 58.9%. Fortunately, it is shown in Table 5 that only 3.8% of the checked out shopping carts involved Multi Size for the same style of sweater.

Step 5 & 6: Predictive Analysis – Anonymous and Non-Anonymous Shoppers

The modeling process is developed using SAS® Enterprise Miner™ 13.2. The process flow diagram is shown in Figure 7. Two modeling scenarios are built. The first scenario (Step 5) only utilizes the shopping cart dimension to analyze the return behavior of online shoppers. In the absence of any individual customer information, these online shoppers are considered to be anonymous shoppers (e.g. new site visitors). The second scenario (Step 6) models the return behavior of the same group of customers. But when appended with additional individual-level background and behavioral variables (i.e. customer dimension), these customers are well recognized and can be regarded as non-anonymous (or known) shoppers (e.g. existing customers identified by web cookie).

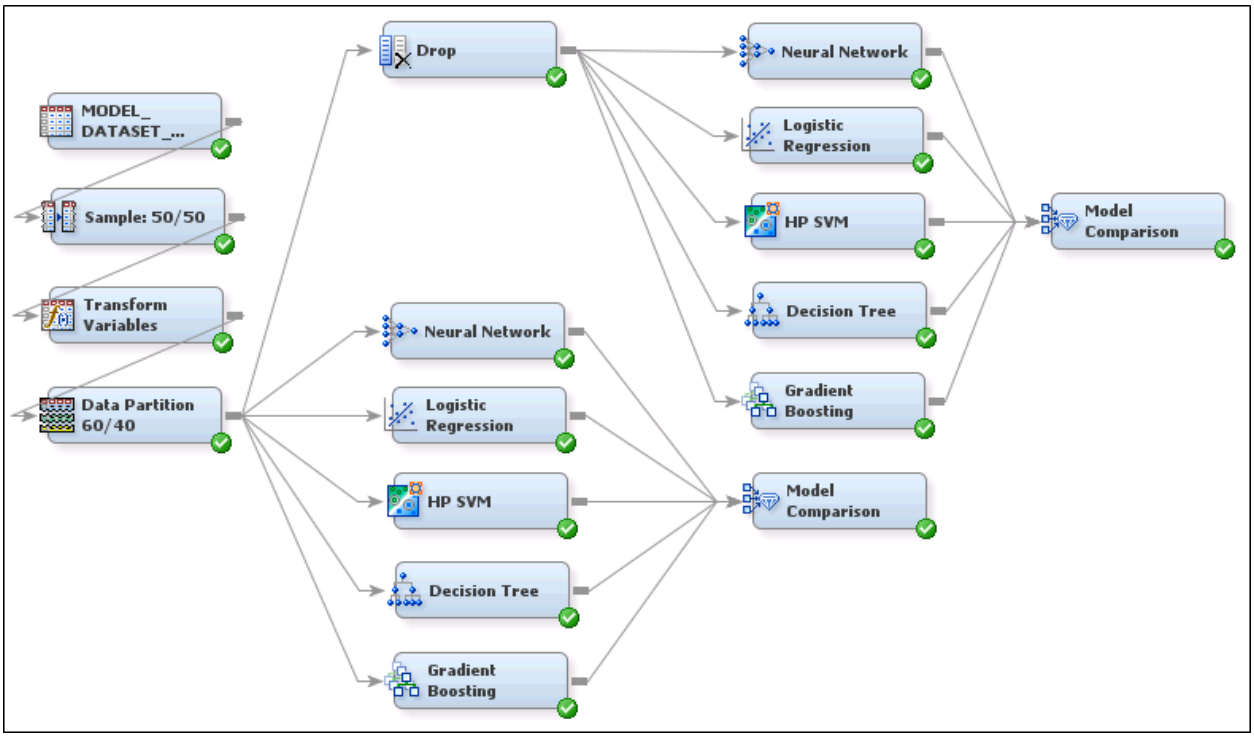


Figure 7: A Process Flow Diagram for Data Processing and Predictive Analysis using SAS® Enterprise Miner™ 13.2

The panel setting of each data-mining modeling node (i.e. Neural Network, Logistic Regression, Support Vector Machine, Decision Tree and Gradient Boosting), as well as that of the Drop node and Model Comparison node, are outlined in Appendix C. In Step 5, some modeling node parameter values are tested for different values (e.g. the weight decay constant in Neural Network and penalty constant in Support Vector Machine) to reach better predictive accuracy performance (primarily based on misclassification rate). No vigorous attempt is made to optimize all parameter settings in this study. In order to attain a fair comparison, however, all panel settings (non-optimized) of the five data-mining modeling nodes in Step 6 are simply set as the same as those in Step 5. Figure 8 and Figure 9 display the comparative results obtained from the two Model Comparison nodes.

In terms of misclassification rate, Neural Network performed the best and Support Vector Machine performed the worst in both scenarios. In terms of the top 10% cumulative lift, Neural Network performed the best and Gradient Boosting performed the worst in the anonymous shopper scenario; Logistic Regression performed the best and Decision Tree performed the worst in the non-anonymous shopper scenario.

It is important to note that the best model in the anonymous shopper scenario (Neural Network with misclassification rate = 0.403 and cumulative lift = 1.482) is still relatively inferior as compared with the worst models (non-optimized) in the non-anonymous shopper scenario (Support Vector Machine with

misclassification rate = 0.359 and Decision Tree with cumulative lift = 1.521). This suggests that while a better predictive performance may be achieved by digging from a list of data-mining technique candidates, the room for performance improvement can still be limited when important input variables are missing, as is evident in this case when the customer dimension is absent. These individual customer background and behavioral variables have proven their valuable contribution to the likelihood prediction of merchandise return.

Model Description	Selection Criterion: Valid: Misclassification Rate	Valid: Cumulative Lift
Neural Network	0.40229	1.482142
Decision Tree	0.402512	1.456874
Logistic Regression	0.404735	1.443438
Gradient Boosting	0.404958	1.418939
HP SVM	0.4243	1.46755

Figure 8: Model Accuracy Comparison by Misclassification Rate and Cumulative Lift (depth=10%) for Anonymous Shoppers (Step 5)

Model Description	Selection Criterion: Valid: Misclassification Rate	Valid: Cumulative Lift
Neural Network	0.341485	1.635556
Decision Tree	0.347265	1.520508
Gradient Boosting	0.349822	1.551111
Logistic Regression	0.351267	1.648889
HP SVM	0.359493	1.618889

Figure 9: Model Accuracy Comparison by Misclassification Rate and Cumulative Lift (depth=10%) for Non-Anonymous Shoppers (Step 6)

The two classes (Returned Vs Not Returned) re-sampled in this study are balanced by under-sampling the majority class. We would expect to see the misclassification rate to be close to the baseline of 50% if the prediction was made by random. However, the current misclassification rates for the anonymous and non-anonymous shopper scenarios are found to be lower than the baseline and are in the proximities of 40% and 35% respectively. This suggests that all models developed in either scenario do have value in predicting the likelihood of return.

Step 7: Relative Importance of Variables – Anonymous and Non-Anonymous Shoppers

The relative importance of variables has just been briefly assessed in our descriptive analysis in Step 4. This sub-section attempts to replicate the assessment (with more variables added) using Decision Tree in each of the Anonymous and Non-anonymous shopper predictive analysis scenario. Decision Tree offers high model interpretability and it is free of linearity assumption. A direct assessment routine is available in SAS® Enterprise Miner™ 13.2, and the results depend on the setting of the property panel. We simply use the same setting as we have used to build the trees in either scenario (see Figure 16 in Appendix C) in Step 5 and 6. The results as displayed in Figure 10 and 11 are mostly self-explanatory. Some interpretations are highlighted in the next two paragraphs.

In the Anonymous Shopper scenario, the 4-binned transformed OPT_Dim_Total_Cart_Amt (3 split points) is of top importance while Dim_Cart_Shipping (1 split point), which ranked top in the analysis in Step 4, becomes unimportant. This suggests that OPT_Dim_Total_Cart_Amt has a better split point (e.g. leading to higher reduction in Entropy index) in comparison with that of Dim_Cart_Shipping. As moving down the tree, the variability of the target explained by OPT_Dim_Total_Cart Amt (and Dim_Cart_Shipping as well) diminishes while that explained by other variables such as Dim_Item_Amount and Dim_SIZE remain strong. The branch splitting continues via other variables until the tree reaches the pre-set maximum depth (=10) or optimal level (measured by mis-classification), at which Dim_Cart_Shipping is still not being used for splitting. This variable then appears to have no contribution in impacting returns.

In the Non-Anonymous shopper scenario, the variable RFM_Rtn_12M_Monetary followed by RFM_12M_Monetary, come to the top in variable importance. This suggests that individualized historical return and purchase activities are more important than any other Cart Dimension related variables in determining the likelihood of return.

Variable Name	Validation Importance ▼
OPT_Dim_Total_Cart_Amt	1.0000
OPT_Dim_Item_Amount	0.6288
Dim_SIZE	0.5593
OPT_Dim_Sweater_Cart_Amt	0.2095
Dim_COLOR	0.1848
Dim_STYLE	0.1439
Dim_Cart_Shipping	0.0000

Figure 10: Relative Importance of Variables for the Anonymous Shopper scenario

Variable Name	Validation Importance ▼
LOG_RFM_Rtn_12M_Monetary	1.0000
SQRT_RFM_12M_Monetary	0.5276
OPT_Dim_Total_Cart_Amt	0.5238
OPT_Dim_Item_Amount	0.4048
Dim_SIZE	0.3109
OPT_Dim_Sweater_Cart_Amt	0.2383
SQRT_Customer_Tenure	0.2171
SQRT_RFM_12M_Frequency	0.1392
LG10_RFM_Rtn_12M_Frequency	0.0719
Customer_Membership	0.0585
Dim_STYLE	0.0000
Dim_Cart_Shipping	0.0000
Dim_COLOR	0.0000

Figure 11: Relative Importance of Variables for the Non-Anonymous Shopper scenario

CONCLUSION

This paper has demonstrated a coding scheme to portray the composition of the shopping cart of an online apparel retailer. A shopping cart can be described with five cart dimensional variables, or represented by one (or a few) of the twenty-four different cart profiles. In the absence of customer dimension, it was evident from the descriptive analysis that the return likelihood of an item on a checked out shopping cart varies with its associating cart profile. In general, free shipping impacts return likelihood the most followed by whether the cart contains another size of the same style of item; and whether the cart contains another color of the same style of item gives negligible impact. Our predictive analysis utilized and compared five data-mining techniques, and Neural Network was found to be the most superior technique in predicting the likelihood of return. In the presence of the customer dimension, historical purchase and return activities have become most determinant; and the predictive superiority of Neural Network (and other comparative techniques as well) has improved substantially.

BUSINESS APPLICATIONS AND FURTHER RESEARCH

The shopping cart dimension built and merchandise return behavior analyzed in this paper were based on variables engineering performed within one single merchandise category, and in the style-level; whether a cart contains multiple styles of sweaters, and whether each style has multiple sizes (and/or multiple colors) in the cart. While this within-category focused analysis may be interesting to the merchandising manager responsible for the category, a more senior management executive may want to see a bigger picture from a higher merchandise-hierarchy-level point of view. This may be achieved by redefining the shopping cart dimension, for an instance, whether a cart contains multiple categories of merchandises, and whether each category has multiple styles in the cart. The same rationale applies if the retailer is interested in a more micro-level merchandise view, and this can be achieved by shifting the analysis focus to one merchandise-hierarchy-level down.

The heterogeneity of the merchandise return behavior uncovered through the presented methodology may have business implications for the general online apparel retailers. Retailers may incorporate these behavioral variations in their website design and enhancement of their online product recommendation engines.

- For the same style of item, ordering multiple sizes would lead to high likelihood of return. Retailers may provide better intelligence (e.g. via third-party size/fit discovery platform vendors) to online customers to help them understand their true sizes.
- Given the fact that buying multiple colors of the same style of item would give minimal impact on the return likelihood, retailers may encourage the customer to pick another color of the same style of item once she puts one to the cart.

- One challenge is about boosting purchases of multiple styles while avoiding them to be returned by the customer. When building recommendation engines, retailers may contemplate to set up a style-based bundle pricing structure (e.g. offer discounts for additional style purchased) along with certain return restriction policy. To achieve this, a deeper understanding of the inherent behavioral economics (e.g. the net effect of incremental profit margin minus the discounts offered) is needed. This would naturally require accurately predicting the customer's return likelihood of the style she has put to the shopping cart. The prediction accuracies, when measured amongst the anonymous and non-anonymous shoppers, have been evident to be different. Recommendation engine designers may find the presented methodological framework to be useful in solving the needed complexity for the prediction of return likelihood.

ACKNOWLEDGEMENTS

The author is grateful to Carole Jesse for the initial review of this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Sunny Lam
ANN Inc.
Address: 7 Times Square, New York, NY 10036
Email: sunny_lam@anninc.com or sunnylam_us@yahoo.com
LinkedIn: https://www.linkedin.com/in/sunnycylam

REFERENCES

Bao, X. (2007), "Mining Transaction/Order Data Using in SAS® Enterprise Miner™ Association Node". Proceedings of the SAS Global Forum 2007, Cary, NC: SAS Institute Inc.

Chawla, N. V. (2005), "Data Mining for Imbalanced Datasets: An Overview", Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, Springer, pp. 853-867.

Faron, M. and G. Chakraborty (2012), "Easily Add Significance Testing to your Market Basket Analysis in SAS® Enterprise Miner™". Proceedings of the SAS Global Forum 2012, Cary, NC: SAS Institute Inc.

Redlon, M. (2003), "A SAS® Market Basket Analysis Macro: The 'Poor Man's Recommendation Engine'". Proceedings of the Twenty-Eighth Annual SAS Users Group International Conference April 2003.

RECOMMENDED READING

Sarma, K. S. (2013). Predictive Modeling with SAS® Enterprise Miner™: Practical Solution for Business Applications, Second Edition. Cary, NC: SAS Institute Inc.

APPENDIX A

```
/* 1. PROFILING */

Proc SQL;
     Create Table Cart_1 as  /* table b --- size & color for each style for sweater */
         SELECT              Cart_ID,
                             Style_Code, COUNT(DISTINCT Size_Code) as Size_Count,
                             COUNT(DISTINCT Color_Code) as Color_Count
             FROM            EXTRACT_DATASET
             WHERE           Product_Class = "Sweater"
             GROUP BY        Cart_ID, Style_Code
                 UNION
         SELECT              Cart_ID, /* size & color for each style for non-sweater */
                             Style_Code, COUNT(DISTINCT Size_Code) as Size_Count,
                             COUNT(DISTINCT Color_Code) as Color_Count
             FROM            EXTRACT_DATASET
             WHERE           Product_Class = "Non-Sweater"
             GROUP BY        Cart_ID, Style_Code;
     Create Table Cart_2 as  /* table c --- style for sweater and non-sweater */
         SELECT              Cart_ID, Product_Class,
                             COUNT(Distinct Style_Code) as Style_Count
             FROM            EXTRACT_DATASET
             GROUP BY        Cart_ID, Product_Class;
    Create Table Cart_3 as  /* table d --- quantity & amount for the whole cart */
         SELECT              Cart_ID,
                             SUM(Case When Product_Class="Sweater     " Then Item_Quantity
                                      Else 0 End) as Sweater_Cart_Qty,
                             SUM(Case When Product_Class="Non-Sweater" Then Item_Quantity
                                      Else 0 End) as NonSweater_Cart_Qty,
                             SUM(Case When Product_Class="Sweater     " Then Item_Amount
                                      Else 0 End) as Sweater_Cart_Amt,
                             SUM(Case When Product_Class="Non-Sweater" Then Item_Amount
                                      Else 0 End) as NonSweater_Cart_Amt
             FROM            EXTRACT_DATASET
             GROUP BY        Cart_ID;
    Create Table REPORT_DATASET as
         SELECT              a.*,
                             c.Style_Count, b.Size_Count, b.Color_Count,
                             d.Sweater_Cart_Qty, d.NonSweater_Cart_Qty, d.Sweater_Cart_Amt,
                             d.NonSweater_Cart_Amt
             FROM            EXTRACT_DATASET as a
             LEFT JOIN       Cart_1 as b
             ON              a.Cart_ID = b.Cart_ID AND
                             a.Style_Code = b.Style_Code
             LEFT JOIN       Cart_2 as c
             ON              a.Cart_ID = c.Cart_ID AND
                             a.Product_Class = c.Product_Class
             LEFT JOIN       Cart_3 as d
             ON              a.Cart_ID = d.Cart_ID
             ORDER BY        a.Record_Number;
QUIT;
Data REPORT_DATASET;
     Set REPORT_DATASET;
         If      Style_Count = 1 Then STYLE = "One Style  ";
         Else                         STYLE = "Multi Style";

         If      Size_Count  = 1 Then SIZE  = "One Size   ";
         Else                         SIZE  = "Multi Size ";

         If      Color_Count = 1 Then COLOR = "One Color  ";
         Else                         COLOR = "Multi Color";
 Run;
```

APPENDIX B

```
 /* 2. REPORTING */

Proc SQL;
     Create Table Cart_Profile_1 as
         SELECT      DISTINCT Cart_ID, STYLE, SIZE, COLOR
         FROM        REPORT_DATASET
         WHERE       Product_Class="Sweater";
     Create Table Cart_Profile_2 as
         SELECT      Cart_ID, Sweater_Cart_Qty, NonSweater_Cart_Qty,
                     Sweater_Cart_Amt, NonSweater_Cart_Amt,
                     Case When Sweater_Cart_Amt + NonSweater_Cart_Amt < 150
                              Then "Charged"
                              Else "Free    " End as Cart_Shipping,
                     Case When Sweater_Cart_Amt < 150 Then "Sweater $ < $150 "
                              Else "Sweater $ >= $150" End as Sweater_Shipping,
                     Case When NonSweater_Cart_Qty > 0 Then 1 Else 0 End as NonSweater_Cart,
                     SUM(Case When Product_Class="Sweater     " Then Return_Quantity Else 0 End)
                              as Rtn_Qty_Sweater,
                     SUM(Case When Product_Class="Non-Sweater" Then Return_Quantity Else 0 End)
                              as Rtn_Qty_NonSweater,
                     SUM(Case When Product_Class="Sweater     " Then Return_Amount Else 0 End)
                              as Rtn_Amt_Sweater,
                     SUM(Case When Product_Class="Non-Sweater" Then Return_Amount Else 0 End)
                              as Rtn_Amt_NonSweater
         FROM        REPORT_DATASET
         GROUP BY    Cart_ID, Sweater_Cart_Qty, NonSweater_Cart_Qty, Sweater_Cart_Amt,
                     NonSweater_Cart_Amt, Calculated Cart_Shipping,
                     Calculated Sweater_Shipping, Calculated NonSweater_Cart;
 QUIT;
Data Cart_Profile_1_2;
     Merge       Cart_Profile_1 Cart_Profile_2;
     By          Cart_ID;
 Run;
Proc SQL;
     Create Table Report_1 as
         SELECT      Cart_Shipping, Sweater_Shipping,
                     STYLE, SIZE, COLOR,
                     COUNT(DISTINCT Cart_ID)      as Distinct_Cart,
                     SUM(NonSweater_Cart)         as NonSweater_Cart,
                     SUM(Sweater_Cart_Qty)        as Cart_Qty_Sweater,
                     SUM(NonSweater_Cart_Qty)     as Cart_Qty_NonSweater,
                     SUM(Rtn_Qty_Sweater)         as RT_Qty_Sweater,
                     SUM(Rtn_Qty_NonSweater)      as RT_Qty_NonSweater,
                     SUM(Sweater_Cart_Amt)        as Cart_Amt_Sweater,
                     SUM(NonSweater_Cart_Amt)     as Cart_Amt_NonSweater,
                     SUM(Rtn_Amt_Sweater)         as RT_Amt_Sweater,
                     SUM(Rtn_Amt_NonSweater)      as RT_Amt_NonSweater
         FROM        Cart_Profile_1_2
         GROUP BY    Cart_Shipping, Sweater_Shipping,
                     STYLE, SIZE, COLOR;
    Create Table Report_2 as
        SELECT      COUNT(DISTINCT Cart_ID)      as Distinct_Cart,
                    SUM(NonSweater_Cart)         as NonSweater_Cart,
                    SUM(Sweater_Cart_Qty)        as Cart_Qty_Sweater,
                    SUM(NonSweater_Cart_Qty)     as Cart_Qty_NonSweater,
                    SUM(Rtn_Qty_Sweater)         as RT_Qty_Sweater,
                    SUM(Rtn_Qty_NonSweater)      as RT_Qty_NonSweater,
                    SUM(Sweater_Cart_Amt)        as Cart_Amt_Sweater,
                    SUM(NonSweater_Cart_Amt)     as Cart_Amt_NonSweater,
                    SUM(Rtn_Amt_Sweater)         as RT_Amt_Sweater,
                    SUM(Rtn_Amt_NonSweater)      as RT_Amt_NonSweater
        FROM        Cart_Profile_2;
QUIT;
```

APPENDIX C

Drop – For the sake of convenience, the whole MODEL_DATASET was loaded to SAS® Enterprise Miner™ 13.2 in Step 3. When analyzing anonymous customers, we do not need any variables under the customer dimension (See the dataset structure in Table 3). Right click the Drop node, and choose Drop = "Yes" for all customer dimension related variables. See Figure 12 below.

Name	Drop	Role	Level
Customer_Membership	**Yes**	Input	Nominal
Dim_COLOR	**Default**	Input	Nominal
Dim_Cart_Shipping	**Default**	Input	Nominal
Dim_SIZE	**Default**	Input	Nominal
Dim_STYLE	**Default**	Input	Nominal
LG10_RFM_Rtn_12M_Frequency	**Yes**	Input	Interval
LOG_RFM_Rtn_12M_Monetary	**Yes**	Input	Interval
OPT_Dim_Item_Amount	**Default**	Input	Nominal
OPT_Dim_Sweater_Cart_Amt	**Default**	Input	Nominal
OPT_Dim_Total_Cart_Amt	**Default**	Input	Nominal
Record_Number	**Default**	ID	Interval
Return	**Default**	Target	Binary
SQRT_Customer_Tenure	**Yes**	Input	Interval
SQRT_RFM_12M_Frequency	**Yes**	Input	Interval
SQRT_RFM_12M_Monetary	**Yes**	Input	Interval

Figure 12: Dropping the Customer Dimension

Neural Network – Drag in the Neural Network node and connect it with the Drop node. Set the Model Selection Criterion to be "Misclassification" in the Property panel, as shown in Figure 13 below. Click the ellipsis besides Network. A Network property panel will be displayed.

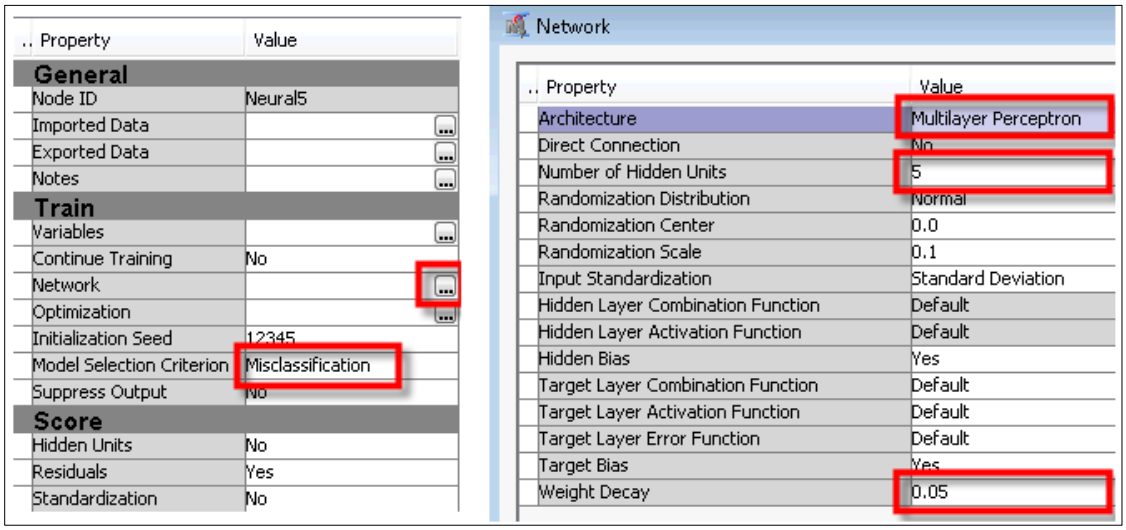


Figure 13: Property panel setting for Neural Network

Let the Architecture be at the default setting “Multilayer Perceptron”. Various trial and error settings were experimented on the Number of Hidden Units and Weight Decay, and they were finally set as 5 and 0.05 respectively.

Logistic Regression – Drag in the Regression node (renamed as Logistic Regression) and connect it with the Drop node. Under Model Selection, set Selection Model = “Stepwise” and Selection Criteria = “Validation Misclassification”. See Figure 14 below.

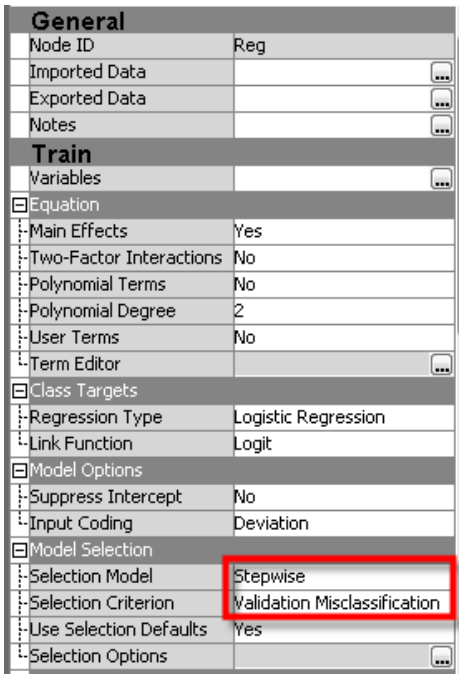


Figure 14: Property panel setting for Logistic Regression

Support Vector Machine – Drag in the HP SVM node and connect it with the Drop node. Click the ellipsis besides the Interior Point Options and let the default Kernel be “Linear”. A better misclassification rate was seen for Penalty = 0.25. See Figure 15.

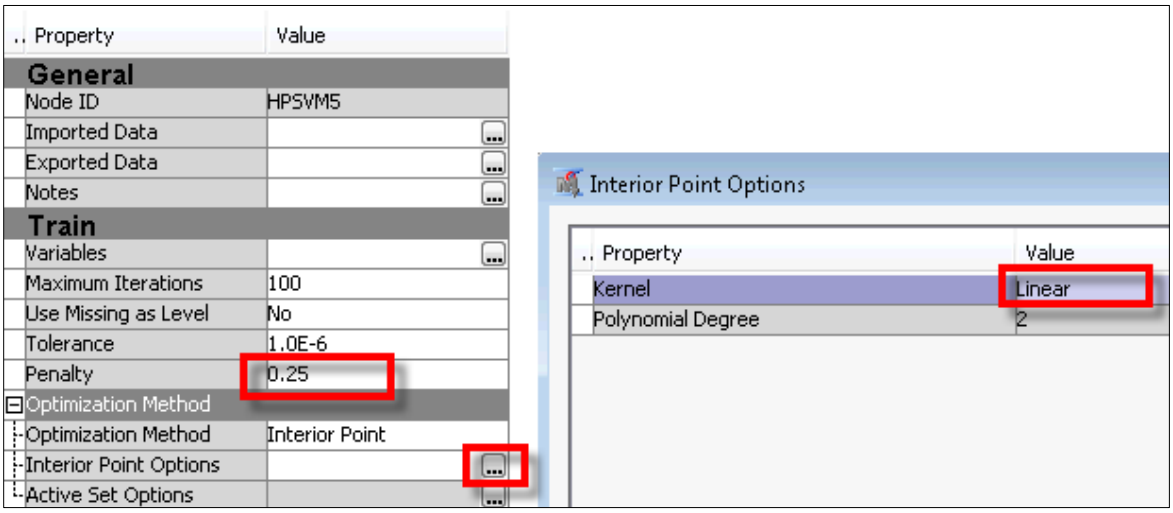


Figure 15: Property panel setting for Support Vector Machine

Decision Tree – Drag in the Decision Tree node and connect it with the Drop node. Set the Nominal Target Criterion = "Entropy", Maximum Depth = 10 and Assessment Measure = "Misclassification". See Figure 16 below.

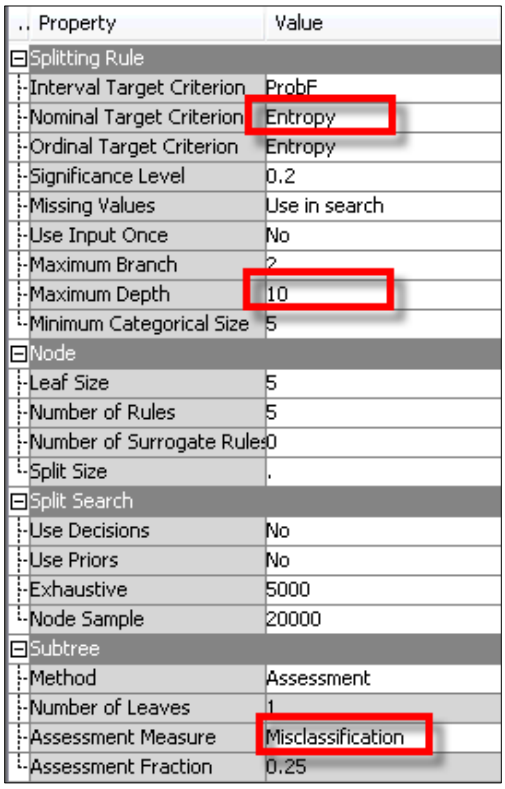


Figure 16: Property panel setting for Decision Tree

Gradient Boosting – Drag in the Gradient Boosting node and connect it with the Drop node. By using "Misclassification" as the Assessment Measure, a few different combinations of Shrinkage and Maximum Depth have been tested. Finally, a better misclassification measure was seen for Shrinkage = 0.85 and Maximum Depth = 2, with N Iterations = 100 to be sufficient.

Property	Value
Series Options	
N Iterations	100
Seed	12345
Shrinkage	0.85
Train Proportion	60
Splitting Rule	
Huber M-Regression	No
Maximum Branch	2
Maximum Depth	2
Minimum Categorical Size	5
Reuse Variable	1
Categorical Bins	30
Interval Bins	100
Missing Values	Use in search
Performance	Disk
Node	
Leaf Fraction	0.1
Number of Surrogate Rules	0
Split Size	.
Split Search	
Exhaustive	5000
Node Sample	20000
Subtree	
Assessment Measure	Misclassification

Figure 17: Property panel setting for Gradient Boosting

Model Comparison – The primary model comparison metric used was Misclassification Rate, and was assessed with the use of the validation dataset (40% of the observations in the overall dataset).

Property	Value
Assessment Reports	
Number of Bins	10
ROC Chart	Yes
Recompute	No
Model Selection	
Selection Data	Default
Selection Statistic	Misclassification Rate
Grid Selection Statistic	Default
Selection Table	Validation
Selection Depth	10

Figure 18: Property panel setting for Model Comparison