Confidential

Missing Marital Status Prediction for Hypermarkets

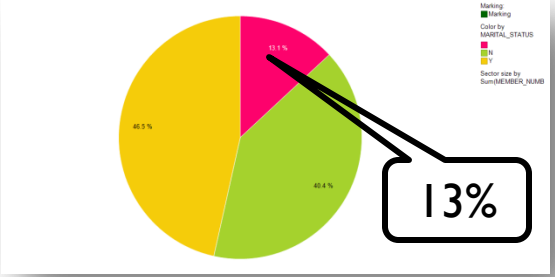
Project Presentation

BADM Team B-5: Sankalp Gaur, Sonali Gadekar, Harshita Jujjuru, Tushna Mistry, Vineet Jain

Business Problem

MEMBER_NUMBER	DOB	SEX	ENROLLMENT_SALE_DATE	MARITAL_STATUS	Transaction_Date	Sku_Number	Quantity_Sold
3000489694	10/15/1970	M	2/4/2010	Y	4/4/2012	[illegible]	3
3000489694	10/15/1970	M	2/4/2010	Y	4/4/2012	100003227	3
3000496319	9/20/1980	M	4/4/2010		4/4/2012	100382156	6
3000496319	9/20/1980	M	4/4/2010		4/4/2012	100021016	6
3000710511	8/31/1990	M	2/9/2011	N	4/4/2012	100011449	3
3000710511	8/31/1990	M	2/9/2011	N	4/4/2012	100006185	3
3000710511	8/31/1990	M	2/9/2011	N	4/4/2012	100006185	3
3000710511	8/31/1990	M	2/9/2011	N	4/4/2012	100394655	3

Missing values for 'Marital Status'



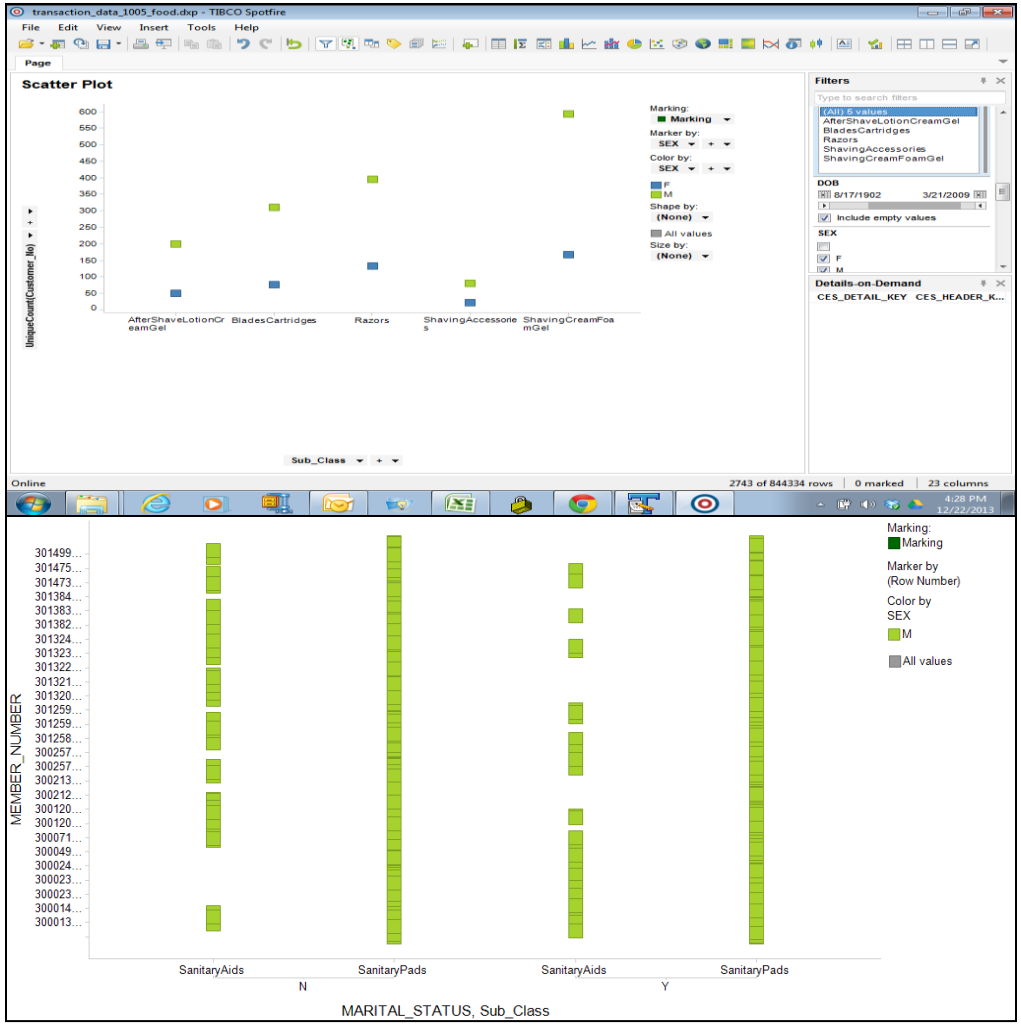
Stakeholder
- Marketing team of the supermarket could be the client

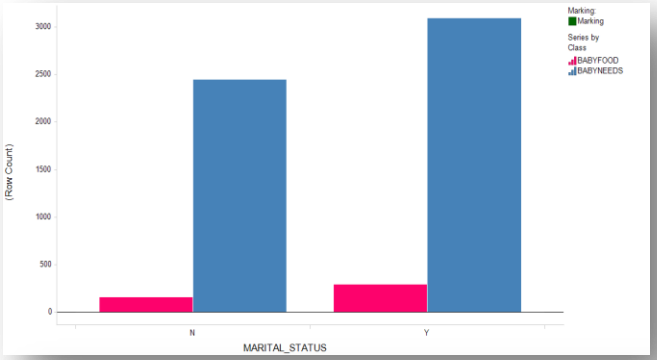
Use Case
- Targeting family bulk shopping offers to family customers

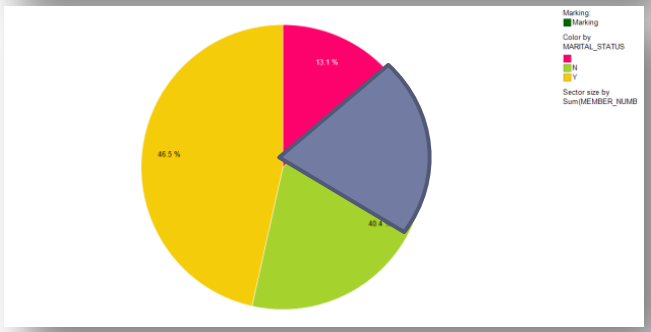
Objective
- To identify married customers in the customer data set.

Benefit
- Correct grasp of the marital status for customer segmentation.

Data Mining Problem







Objective

- To identify FAMILY customers in the customer data set.

Data Description

Customer Data

MEMBER_ NUMBER	DOB	SEX	STATUS	ENROLLMENT_ SALE_DATE	ENROLLMENT _STORE	MARITAL _STATUS	CHILDREN
3000000046	9/19/1968	M	A	4/12/2009	1001	Y	
3000000053	7/2/1958	M	A	4/12/2009	1001	Y	
3000000061	7/18/1964	F	A	4/12/2009	1001	Y	
3000000079	8/29/1967	M	A	4/12/2009	1001	Y	
3000000087	3/17/1969	F	A	4/12/2009	1001	Y	
3000000095	2/27/1962	F	A	4/12/2009	1001	Y	
3000000137	10/16/1962	M	A	4/12/2009	1001	Y	
3000000145	4/2/1956	F	A	4/12/2009	1001	Y	
3000000152	7/1/1957	M	A	4/12/2009	1001	Y	

Transaction Data

Transaction_ Date	Sku_Number	Quantity Sold	Extended_ Price	Item_Description	Sub_Department	Class	Sub_ Class
8/11/2012	100397653					POS/	FacialT
8/11/2012	100005941						
8/11/2012	100184169						
8/11/2012	100205998	6					
8/11/2012	100011791	3					
8/11/2012	100296358	3					

Transaction Level

- KNN (SKU#, Age (derived field), Qty Sold, Extended Price)

Basket Level

- Classification Trees (Age, Qty Sold, Sex, Extended Price)
- Association Rules (Classes within a basket)

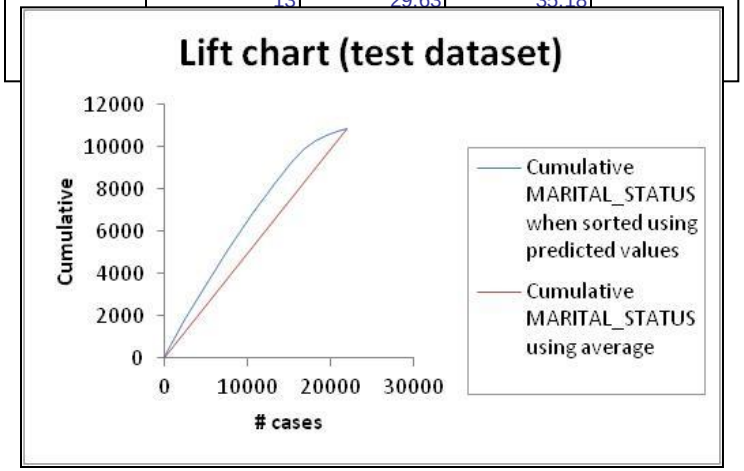
Customer Level

- KNN (Frequency of Class/Subclass, Age, Dummy Sex)
- Logistic Regression

KNN (Transaction Level Data)

Validation error log for different k

Value of k	% Error Training	% Error Validation	
1	2.14	38.53	
2	19.17	37.98	
3	19.82	37.06	
4	23.96	36.07	
5	24.52	36.22	
6	26.21	35.07	
7	26.61	35.59	
8	27.59	34.99	
9	27.98	35.29	
10	28.66	34.81	
11	28.93	35.04	
12	29.42	34.71	<--- Best k
13	29.63	35.18	



Training Data scoring - Summary Report (for k=12)

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	Y	N
Y	3951	1008
N	1933	3105

Error Report			
Class	# Cases	# Errors	% Error
Y	4959	1008	20.33
N	5038	1933	38.37
Overall	9997	2941	29.42

Validation Data scoring - Summary Report (for k=12)

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	Y	N
Y	12244	4199
N	7257	9301

Error Report			
Class	# Cases	# Errors	% Error
Y	16443	4199	25.54
N	16558	7257	43.83
Overall	33001	11456	34.71

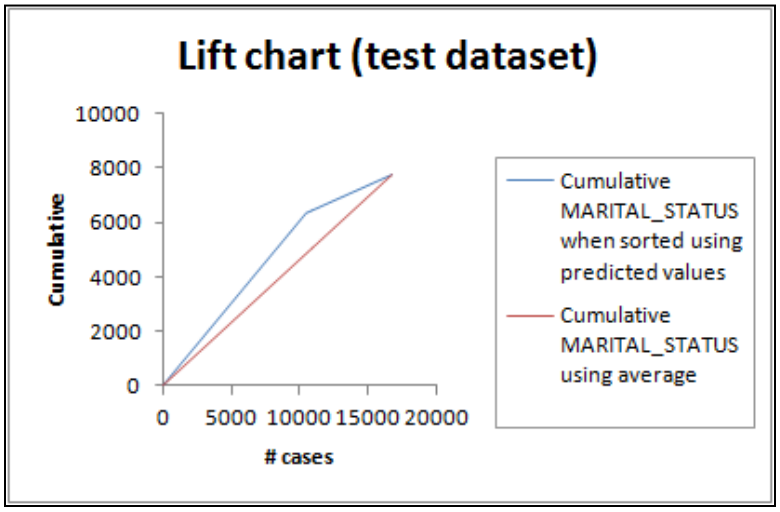
KNN – Customer Level Aggregation

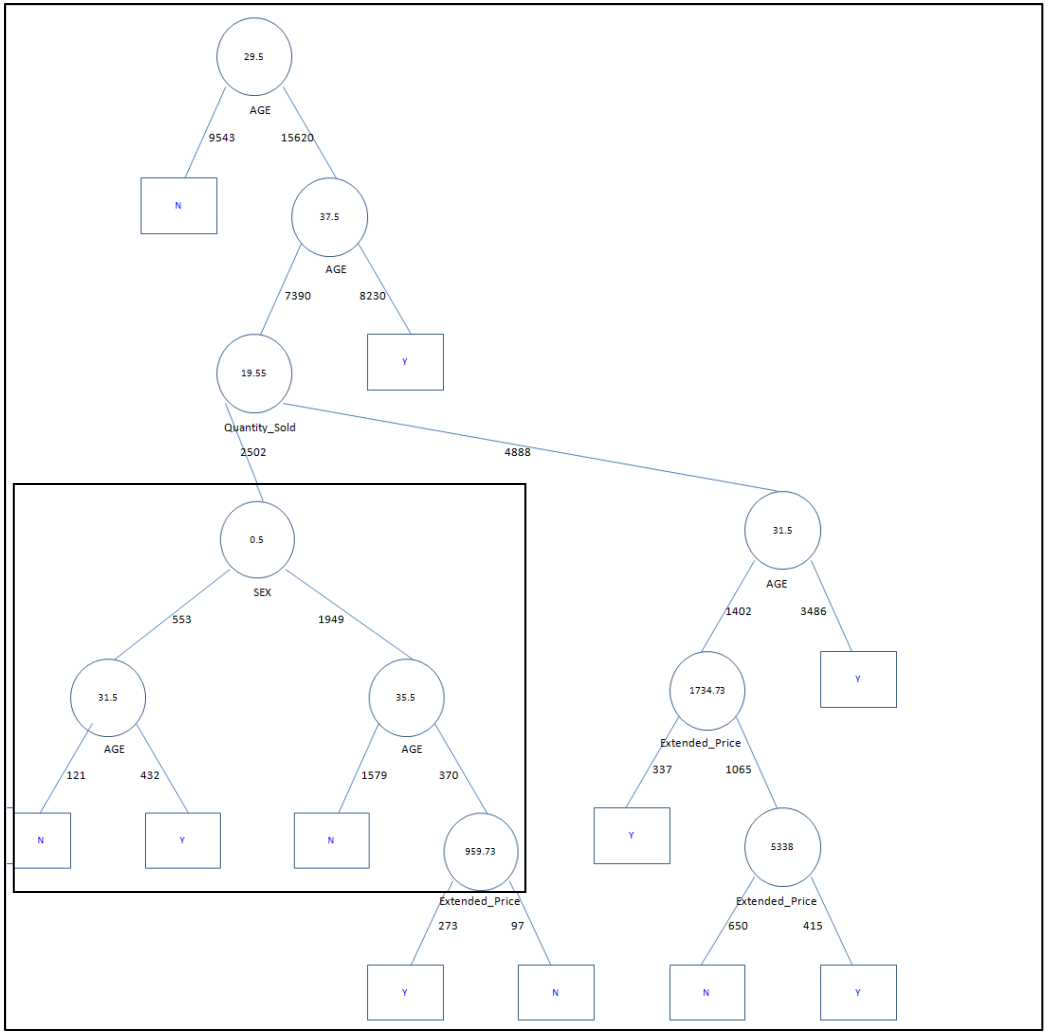
Row Id.	Predicted Class	Actual Class	Prob. for Y (success)	Actual #Nearest Neighbors	Sku_Number	Quantity_Sold	Extended_Price	Age	USER IDs
5	Y	Y	0.583333	12	100183807	3	165	43	3000489694
7	Y	Y	0.666667	12	100107679	9	315	43	3000489694
13	Y	Y	0.833333	12	100008984	3	42	43	3000489694
15	N	Y	0.083333	12	100007346	3	264	43	3000489694
16	Y	Y	0.833333	12	100037662	3	207	43	3000489694
19	Y	Y	0.5	12	100004403	3	108	43	3000489694
21	Y	Y	0.5	12	100208363	3	957	43	3000489694

Row Labels	N	Y	Total Votes	Final Predicted Class	Actual Class
3000489694	3	21	24	Y	Y
3000117550		1	1	Y	Y
3000117857	4	2	6	N	Y
3000117865	21	42	63	Y	Y
3000117899	3	13	16	Y	Y
3000117956	2	15	17	Y	Y
3000118194	5		5	N	N
3000118335	7	3	10	N	Y
3000118699	3	15	18	Y	Y

Classification Tree (Basket Level Data)

Cut off Prob.Val. for Success (Updatable)			0.5
Classification Confusion Matrix			
	Predicted Class		
Actual Class	Y	N	
Y	6309	1408	
N	4172	4887	
Error Report			
Class	# Cases	# Errors	% Error
Y	7717	1408	18.24543216
N	9059	4172	46.05364831
Overall	16776	5580	33.26180258





Association Rules (Basket Level Data)

Rule No.	Antecedent (a)	Consequent (c)	Lift Ratio		Unmarried Consequent (c)	Unmarried Lift Ratio		Difference in Lift Ratio
1	DETERGENTS	HOUSEHOLDCLEANING	2.734715		BISCUITS	1.883152		0.851563
2	HOUSEHOLDCLEANING	DETERGENTS	2.734715		NULL	NULL		NULL
3	PULSES	SPICESMASALAS	2.655388		CONFECTIONERY	1.605463		1.049925
4	PULSES	EDIBLEOILS	2.653584		BISCUITS	1.663776		0.989808
5	BISCUITS, PERSONAL HYGIENE	HOUSEHOLDCLEANING	2.603407		SAVORIES	1.908845		0.694562
6	HOUSEHOLDNEEDS	HOUSEHOLDCLEANING	2.537743		CONFECTIONERY	1.605463		0.93228
7	ORALCARE	PERSONAL HYGIENE	2.458185		BISCUITS	1.663776		0.794409
8	DETERGENTS	EDIBLEOILS	2.408741		NULL	NULL		NULL
9	BISCUITS, HOUSEHOLDCLEANING	PERSONAL HYGIENE	2.394673		NULL	NULL		NULL
10	EDIBLEOILS	SPICESMASALAS	2.388987		NULL	NULL		NULL

Training Error Report

Row Labels	Count of Predicted
N	**1756**
N	1074
Y	682
Y	**1301**
N	722
Y	579
Grand Total	**3057**

Test Error Report

Row Labels	Count of Predicted
N	**1534**
N	1189
Y	345
Y	**1184**
N	863
Y	321
Grand Total	**2718**

KNN (Customer Level Data)

Predictors

- Frequency of Class (in transaction level data)
- Age, Dummy variable for Sex

Training Data scoring - Summary Report (for k=15)

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	1481	1525
0	909	4289

Error Report			
Class	# Cases	# Errors	% Error
1	3006	1525	50.73
0	5198	909	17.49
Overall	8204	2434	29.67

Validation error log for different k

Value of k	% Error Training	% Error Validation	
1	0.73	40.88	
2	19.09	39.85	
3	20.93	39.70	
4	24.32	38.12	
5	24.65	37.56	<--- Best k
6	25.84	37.82	
7	26.61	38.78	
8	27.06	39.04	
9	27.97	39.29	
10	28.38	39.18	

Validation Data scoring - Summary Report (for k=15)

Cut off Prob.Val. for Success (Updatable)	0.3

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	925	259
0	750	784

Error Report			
Class	# Cases	# Errors	% Error
1	1184	259	21.88
0	1534	750	48.89
Overall	2718	1009	37.12

Logistic Regression (Customer Level Data)

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-2.74345732	0.1016434	0	-
AGE	0.06231643	0.00253889	0	1.06429911
BABYNEEDS	0.08727679	0.01971586	0.00000957	1.09119868
CEREALS	0.08065249	0.02149971	0.0001759	1.08399416
READYTOFRY	0.05502384	0.01520647	0.00029638	1.05656576
IS_M?	-0.19071524	0.05824867	0.00105978	0.82636786
PERSONAL HYGIENE	-0.03015647	0.01004258	0.00267455	0.9702937
SKIN CARE	-0.04058183	0.01405276	0.00387921	0.96023059
CONFECTIONERY	0.01360664	0.00517822	0.00859722	1.01369965
PACKAGED FOODS	-0.07694651	0.03097989	0.01300046	0.92593938
OTHERFOODS	0.13291447	0.05750981	0.02082393	1.14215231
EGGS	0.22186564	0.09924499	0.02538225	1.24840367
OTHER DAIRY	-0.10177834	0.04575041	0.02610519	0.90322971
JAMSSPREADS	-0.074089	0.03339112	0.02649873	0.92858906
PERISHABLES	-0.21029806	0.09781259	0.03155441	0.81034267
Frozen Veg	0.02776345	0.01305884	0.033501	1.02815247
BISCUITS & COOKIES	-0.0525604	0.02488948	0.03470779	0.94879699
PANEER	-0.10216051	0.05000343	0.04104551	0.9028846
EDIBLEOILS	0.03312676	0.01688433	0.04976448	1.03368151
SHOECARE	0.08761712	0.04471721	0.05007061	1.09157014
Frozen Non Veg	-0.01732613	0.00901711	0.05467219	0.98282307
DAHI & YOGURT	0.05010561	0.03137866	0.11030915	1.05138218
SAUCES	0.02679485	0.01724466	0.12023079	1.02715707
FRESH MILK	-0.04447171	0.03066757	0.14702451	0.95650268
HOUSEHOLDNEEDS	-0.01766296	0.01300819	0.17451642	0.98249209
OTC	0.04465905	0.03452531	0.19583255	1.04567122
HOUSEHOLDCLEANING	0.0179157	0.01385175	0.1958767	1.01807722
BAG	[illegible]	[illegible]	[illegible]	[illegible]
ORALCARE	0.01551995	0.01411404	0.27150169	1.01564097
STAPLES	-0.1670147	0.15601736	0.28439975	0.84618717
BISCUITS	0.00383814	0.00363896	0.29154658	1.00384545
FRESHENERSPESTICIDES	0.0185983	0.01794145	0.29991719	1.01877236
FRUITS	0.00631237	0.00617168	0.30640355	1.0063324
DRINKSJUICES	-0.00405824	0.00398564	0.30857506	0.99594998
CHEESE	0.02170736	0.02168845	0.31688878	1.02194464
SPICESMASALAS	-0.00849055	0.00953569	0.3732526	0.99154538
VIENOISSARY	0.08450485	0.09611169	0.37927341	1.08817816
DISPOSABLES	0.0158936	0.0188664	0.3995479	1.01602054
SANITARYNEEDS	0.01988628	0.02361422	0.39971438	1.02008533
PICKLESCHUTNEYS	-0.03030115	0.03706766	0.41366842	0.97015333
BEVERAGES	0.18421629	0.23815095	0.43921033	1.20227587
HAIRCARE	0.0101694	0.01507898	0.50005162	1.01022124
TEACOFFEE	-0.00899268	0.01342442	0.50293761	0.99104762
SAVORIES	0.00441298	0.00678996	0.51573867	1.00442278
ICECREAMS & GELATO	0.02200767	0.03509916	0.5306499	1.02225161
DETERGENTS	-0.00914698	0.0147819	0.53605068	0.99089473
BUTTER	-0.01700054	0.02777619	0.54050189	0.98314316
FRAGRANCES & DEOS	-0.01206079	0.02003146	0.54711246	0.98801166
READY MEASLS	0.84653908	1.42730975	0.55311346	2.33156347
VEGETABLES	0.00145303	0.00261206	0.57802069	1.00145411
DESSERTS	0.01443248	0.02786805	0.60453773	1.0145371
CIGARETTES	0.03978355	0.07891113	0.61415148	1.04058552
CAKES & PASTRIES	0.01374127	0.03021831	0.64930123	1.01383615
BREADS BUNS N ROLLS	0.00626451	0.01379632	0.64977777	1.00628412
COSMETICS	0.01483446	0.0402584	0.71251458	1.01494503
PULSES	0.00450194	0.01309341	0.73097241	1.00451207
HEALTHDIETFOODS	0.01078645	0.03166864	0.73340207	1.01084483
NOODLESSOUPS	-0.0021026	0.00730741	0.77354938	0.99789959
MENS GROOMING	-0.00621387	0.0267597	0.81637484	0.99380541
READYTOEAT	0.00295611	0.012795	0.81728673	1.00296044
CANNEDFOOD	-0.01181292	0.06790499	0.86189497	0.98825657
FRESH CHICKEN	0.20131375	1.2997359	0.87690943	1.22300839
HEALTHDRINKS	-0.00346201	0.02992665	0.90790349	0.99654394
HERBAL COSMETICS	0.00267376	0.02373337	0.91030139	1.00267732
THERMALBAGS	0.03254323	0.34443678	0.92472571	1.03307856
BREAKFASTFOOD	0.00065505	0.01711194	0.969464	1.00065529
FLOURS	-0.00079511	0.02247385	0.97177726	0.99920523
FRESH PORK	-17.2596626	900.400818	0.9847064	0.00000003
PACKAGEDFOODS	-15.338397	1338.39783	0.99085623	0.00000022
FRESH MUTTON	-13.8045616	1336.84778	0.99176103	0.00000101

Training Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	977	2029
0	666	4532

Error Report			
Class	# Cases	# Errors	% Error
1	3006	2029	67.50
0	5198	666	12.81
Overall	8204	2695	32.85

Validation Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.3

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	876	308
0	641	893

Error Report			
Class	# Cases	# Errors	% Error
1	1184	308	26.01
0	1534	641	41.79
Overall	2718	949	34.92

Ensemble

					Cut-Off	0.5			
Customer ID	C-Trees	K-NN_Trans	K-NN_Cust	Association	Average	Predicted	Actual	Predicted	Actual
3000039911	1	1	1	0	0.75	1	1	Y	Y
3000065270	1	1	1	0	0.75	1	1	Y	Y
3000117899	1	1	1	1	1	1	1	Y	Y
3000117956	1	1	1	1	1	1	1	Y	Y
3000118103	1	1	1	0	0.75	1	1	Y	Y

Cut-off = 0.5	Column Labels			
Row Labels	N	Y	Grand Total	Error%
N	1133	401	1534	26.1%
Y	584	600	1184	49.3%
Grand Total	**1717**	**1001**	**2718**	**36.2%**

Cut-off = 0.4	Column Labels			
Row Labels	N	Y	Grand Total	Error%
N	726	808	1534	52.7%
Y	221	963	1184	18.7%
Grand Total	**947**	**1771**	**2718**	**37.9%**

C-Trees	Column Labels			
Row Labels	0	1	Grand Total	Error%
0	969	565	1534	36.8%
1	368	816	1184	31.1%
Grand Total	**1337**	**1381**	**2718**	**34.3%**

K-NN Trans	Column Labels			
Row Labels	0	1	Grand Total	Error%
0	514	1020	1534	66.5%
1	371	813	1184	31.3%
Grand Total	**885**	**1833**	**2718**	**51.2%**

K-NN Cust	Predicted Class			
Actual Class	1	0		Error%
1	925	259	1184	21.9%
0	750	784	1534	48.9%
Grand Total	**1675**	**1043**	**2718**	**37.1%**

Association	Column Labels			
Row Labels	0	1	Grand Total	Error%
0	1251	283	1534	18.4%
1	930	254	1184	78.5%
Grand Total	**2181**	**537**	**2718**	**44.6%**

Thank You