Confidential

Missing Marital Status Prediction for Hypermarkets

Team B5

i. 61410610 Sankalp Gaur
ii. 61410105 Vineet Jain
iii. 61410700 Sonali Gadekar
iv. 61410425 Harshita Jujjuru
v. 61410715 Tushna Mistry

EXECUTIVE SUMMARY

Business Problem

The customer database contains a field called "MARITAL_STATUS". This is an important field for business. It can help the marketing department to segment the customers and target marketing and promotional initiatives accordingly.

Currently, around 13% of the customers have not reported their marital status. Also we found that the ones who have reported status as single exhibit purchase behavior similar to those of married customers. Through our analysis, we intend to segregate customers into family and non-family customers.

Data

The following data is currently available based on which we need to predict.

Customer Information: Sex , Age, Enrollment date

Customer Purchases: Comprehensive information related to his purchases i.e. The exact items purchased, their price, quantity, class, sub class, sku number etc.

This data can be aggregated at Customer level/ Basket level/ Item Level/ Class level etc. as per the requirement.

Analytics Solution

The analytics objective was to be able to build a model for successful prediction of marital status (in case the same is missing). This was a supervised predictive task, and both forward-looking and retrospective task as new and old records would fall under its purview.

We tried out different data analytics approaches such as KNN (at transaction level and customer level), Classification trees, Association, Logistic Regression, etc. And we finally used ensemble to combine the predictions of the best four models into one single prediction.

Recommendations

Based on our results tested on the test data, we realize that the error rate is much lower when we predict married status Vs the unmarried status. This is because the users have not updated their status accurately.

(i) We recommend this approach to be used for identifying the marital status of the **existing customers**

(ii) Use this approach to also predict the marital status of customers who have not filled it.

(iii) Direct your marketing and promotional activities accordingly.

PROBLEM DESCRIPTION

Business Goal

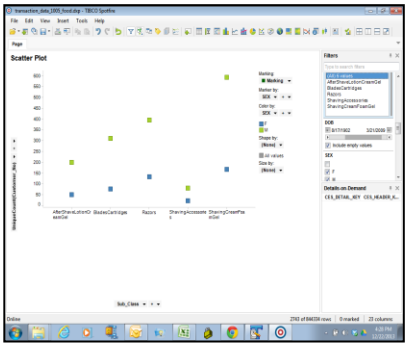
As described above, the business goal we are trying to achieve is to identify a segment towards which we should target bulk offers. In effect, we would like to predict whether a certain member belongs to a household where there is at least one married couple, so that they can target their promotional activities accordingly.

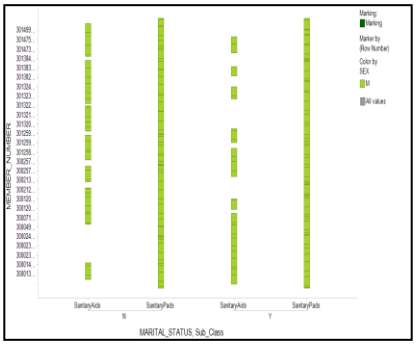
Data Mining Goal

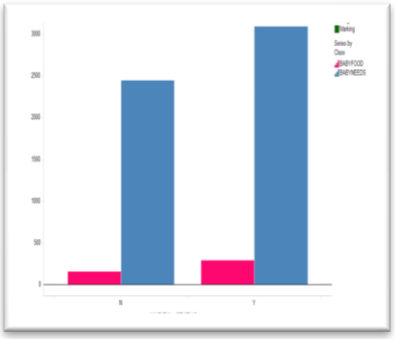
The analytics goal is to predict whether a certain member emulates married buying behavior or not and accordingly classify them as "Married" and "Unmarried". Initially we opted the Naïve Bayes method due to the highly classification nature of the problem. But later we encountered the following issues

(i) Purchasing behavior for baby products (a differentiator) amidst both the segments is similar

(ii) There was hardly any differentiating behavior in the purchase of uniquely female(sanitary pads) and uniquely male(razor blades) products for unmarried customers.







This could happen owing to a number of reasons.

(i) Customers took the membership before marriage and never updated their marital status in the card later.

(ii) Customers are unmarried but stay in a family. Hence, their buying behavior emulates that of families.

At this point, we decided to try multiple approaches, collect their statistics and then run an ensemble of the results. We ran statistics on the below 4 models - (i) KNN (at transaction level and customer level data) (ii) Classification Trees (iii) Association Rules

Data Description and Preparation

We used the customer data joined with transaction data. It had the maximum allowed number of rows for partitioning -65000 and 25 columns. We replaced the numerical NULL values by 0 and removed the column “Children”. We added a new derived field “Age”. We also removed all the records that had NULL values for ‘Marital Status’ as that was of no use for the model. Finally this data was partitioned in 3 sets: Training, Validation and Test. Whereever needed, we aggregated this data at the basket level or Customer level.

Data Mining Solution

(i) KNN

This method involved working on transaction level data. The data was partitioned in Training, validation and test set.

- Inputs: SKU#, Age (derived field), Qty Sold, Extended Price
- Output: Marital Status ; Cut Off: 0.5 ; Benchmark: Error percentage < 50% ; Best K: 12

Validation Data Scoring

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

Test Data Scoring

Classification Confusion Matrix		
	Predicted Class	
Actual Class	Y	N
Y	8166	2690
N	4970	6176

Error Report			
Class	# Cases	# Errors	% Error
Y	10856	2690	24.78
N	11146	4970	44.59
Overall	22002	7660	34.82

This output was for transaction level and we consolidated it to customer level. Logic used for this was to take the marital status for a customer as the predicted most number for times for each of his/her transactions.

Association Rules

We classified the data into classes purchased at each basket level and joined the basket to the customer data to run associations for each segment. Now, We shortlisted only those associations from married customers whose combination DO NOT OCCUR in the unmarried customers.

MARRIED UNMARRIED

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

We ran these associations for each basket and predicted whether the customer who purchased the basket is married or not. Next, we aggregated all the customer basket predictions and took a majority vote to predict whether he is married or not. The results from the Training and Test Data are as below.

TRAINING TEST

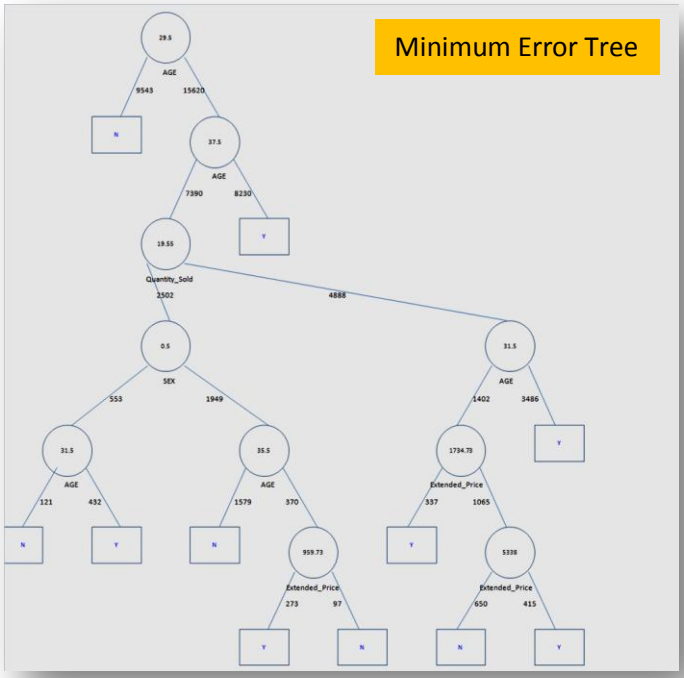
Test Error Report	
Row Labels	Count of Predicted
N	**1534**
N	1189
Y	345
Y	**1184**
N	863
Y	321
Grand Total	**2718**

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

Classification Trees

This method bins people into **Married** or **Unmarried** categories based on predictor variables.
The ones we have used are:

- Age of Customer
- Quantity Sold
- Sex of Customer
- Extended Price (Basket Price)



The model was deployed at a Basket Level. Later, the predictions will be aggregated to get a Customer Level vote (based on a Cutoff Value). The number of levels chosen for the Tree is 7.

As shown alongside, the main predictors are Customer Age and Quantity Sold. Sex weighs in only once, while Extended Price is useful for ultimate separation into bins.

Caveat: The accuracy of this prediction does depend largely on Age, and running the same model without Age as a predictor drops the accuracy for 'Married' customers by ~20%.

When the model is deployed on New Data, the error rate for predicting **Married** customers is 31.1% while that of **Unmarried** is slightly higher, driving up overall error to 34.3%.

CART	Predicted			
Actual	**0**	**1**	**Grand Total**	**Error%**
0	969	565	1534	36.8%
1	368	816	1184	31.1%
Grand Total	**1337**	**1381**	**2718**	**34.3%**

Conclusion / Ensemble

To generate the final prediction, we used a holdout set (The test data of the initial partition) and scored it on the four best models. The error rates were different for the models, but we tried to improve the error rates by combining the different models in an ensemble to create a final prediction. The ensemble is shown below:

					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
3000069629	1	1	0	0	0.5	0	1	N	Y
3000079917	1	1	1	0	0.75	1	0	Y	N
3000117550	0	1	1	0	0.5	0	1	N	Y
3000117857	1	1	1	0	0.75	1	1	Y	Y
3000117865	1	0	1	1	0.75	1	1	Y	Y
3000117899	1	1	1	0	0.75	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

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%

Notice that when we take a simple average of the votes, though the overall error percentage is quite good (36%), the error of our desired category (married prediction) is quite high at 49%. But with some trial and error and using a cutoff of 0.4 while taking the average, we are able to optimize this error to 18.7% with only a small increase in overall error. Hence, we predict this as our final model.

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%

Implications

Our original business problem was to predict which consumers were exhibiting the same buying behaviors as married couples and families. Once a store has a confident prediction about such customers, the management can target them with bulk buying deals and other promotions.

The cost associated with predicting an unmarried customer as married is not very high because as long as the promotion is for bulk buying deals and other general promotions, the customer is not likely to be offended. The store will only lose the costs associated with printing the coupons and promotions. On the other hand, the benefits from gaining customer loyalty and giving the right offers to the right consumers will be greater. Therefore, it would be better to get the error rate for predicting the married customers as low as possible while keeping the error rate for predicting unmarried customers within reasonable limits.

Recommendations

- Based on this information, the management could create promotions for products used regularly by married people and families.
- Other actions could be during certain Indian holidays that are associated more with married people.
- Finally, at a more broad level, such married and unmarried customer information could be used to cross sell products related to the each lifestyle.