EFFECTIVE PREMIUM - CUSTOMER TARGETING USING CLASSIFICATION METHODS

- Increase number of purchases of high margin products using classification methods

TEAM B2

Akshat Narain (61310333)

Aman Rathi (61310645)

Arpita Ray (61310120)

Jayagowri Hoblidar (61310353)

Vismay Shah (61310258)

TEAMPROJECT : BUSINESS ANALYTICS USING DATAMINING

EFFECTIVE PREMIUM - CUSTOMER TARGETING USING CLASSIFICATION METHODS

Table of **Contents**

1. Executive summary

Hypermarts frequently use promotions via mail-in-rebate coupons, bulk buy discount offers etc. to influence customers to purchase greater number of products from their stores. Keeping this in mind, the potential benefit to the Hypermart can be significantly increased if the right promotions are targeted to the right customers - more specifically, identifying a **new customer** as a potential **high margin customer** and targeting him/her with promotions related to high margin products for greater sales turnover of such products.

New Customer: A customer who has purchased **exactly once** in the Hypermart

High Margin Customer ('H'): A customer who purchases high margin products more than 50% of the time.

The data mining problem is to 'classify' a new customer as either a high margin or a low margin customer using the supervised learning techniques.

We used Logistic Regression to classify customers in the high or low margin category using the following predictors obtained on the first basket purchase: # of SubDepts, Quantity Sold, Price of Basket, Age, Sex, Day of week. The predictors were selected using stepwise regression method and selecting the best subset of predictors.

After experimenting with the Logistic Regression and CART classification methods on partitioned data (50%: training, 30%: validation, 20%: testing), we compared the accuracy of the models using the confusion matrix (cutoff probability for a high margin customer = 0.5) results on the test data. We also performed an ensembles analysis on the results of the two models and noticed that the error rate on the (21%) is higher than the logistic regression model. Logistic regression gives us the best accuracy (error rate on 'H' prediction: 19%) and CART gives us the lowest accuracy (23%).

We recommend our Hypermart client to implement the logistic regression model in real-time while the new customer is checking out. Based on the prediction of the model, the customer may be offered promotions related to high margin products leading to increased return on marketing spend.

<u>Assumptions:</u> In order to determine products in-store as high margin vs. low margin, the team looked at various sources online and did secondary research on typical margins on products in Hypermart chains. We are also assuming a cutoff probability of 0.5 in our models for classifying a customer as a high margin customer. This may need to be modified based on the cost-benefit analysis of incorrect predictions.

2. Problem description

Our client is a large format hyper market that sells food, fashion and electronics. It is a highly customer centric company with a loyalty membership base of more than a lakh of customers. They use data driven insights and analytics to better understand their customer shopping behavior and drive higher sales and profitability.

2.1 Business goal: The business goal would be to increase the sales of the high margin products in the hypermart. We first try to predict the future shopping behavior of first time shoppers at hypermart; whether they would be high margin or low margin customers. Based on this classification we can send targeted promotions and discount offers to these potential high margin customers and increase the sales of high margin products. This would in turn result in increased profits and create an opportunity to maximize the return on advertisement (ROA). The shortcoming that we are trying to address is inability to push off the high margin products off the shelf and clear the inventory in regular basis.

2.2 Data mining goal: Identify the potential high margin customers from the set of first time buyers at hyper mart using data mining methods for classification. It is a supervised form of learning wherein the input and output attributes would be as follows:

Input variables	UniqueCount(Sub_Deparm	Sum(Quantity_Sold)	Sum(Extended_Price)	Age	Min(SEX)_M	Min(CLEAN_EMAIL_FLAG)	Day_4	Day_6
Output variable	Margin_H_L							

The output variable is categorical - whether the customer is a potential buyer of high margin products or not. Therefore it is a predictive form of analytics and also forward looking.

3. Data

3.1 Source and Key characteristics: We used the Hyper mart datasets (transaction_1005 - Food) with the transaction data of the customers from 2011 - 2012 and merged it with the customer dataset which contained customer demographics.

3.2 Data preparation

Using retail industry domain knowledge, we attached retailer margin to each of the transaction and then classified them as either high margin or low margin transaction. Thereafter, prepared customer level data by rolling up each of the customer baskets into a single record and identified the customer as - High margin or Low margin depending on whether the average value of all the margins of the customer transaction was greater than the threshold (we chose it as 0.5)

3.3 Picking the relevant data: We wanted to identify the customers who purchased more than one basket, and hence out of the total customer records we filtered 8616 customers with more than 1 basket. Also, the basket data associated with that customer was the first basket that he/she had purchased since we wanted to predict the new customer (assuming a new customer to be one with only 1 basket purchase) - High / Low margin depending only on the first basket that he/she had purchased. We then partitioned this dataset for our purpose.

The following predictor attributes were used :

UniqueCount(Sub_Department)	Sum(Quantity_Sold)	Sum(Extended_Price)	Age	Min(SEX)_M	Min(CLEAN_EMAIL_FLAG)_N	Day_4	Day_6

We added the columns :

- Avg(Margin_bin)
- Margin
- Margin_H_L

3.4 Missing data/Dummy values: We used average age to populate the missing values and the categorical data was handled using "N/A" , dummies were created for categorical variables.

Figure 3.1 - Final Partitioned dataset used for running models

3.5 Data Exploration and visualization

We used the TIBCO Spotfire tool to visualize data and study relationships between the predictors. **Figure 3.2** gives the relationship between how the margins depend on the age,sex and marital status of customers.

4. Data mining solution

Once the data was prepared we partitioned the data (50% training, 30% validation and 20% testing). Since the objective was classification, we decided to use predictive techniques such as logistic regression and CART. We decided to compare the outputs of the two models, and selecting the best model with the least error while classifying an actual high margin customer as a predicted high margin customer. We also decided to perform ensemble analysis on the data by averaging the probability predictions across both the models and comparing the errors with the standalone models.

4.1 Performance of Models and key findings

In order to begin analyzing the best implementation of the models, we first looked for 'illegal' predictors - predictors which may have a lot of explanatory power but cannot be used to predict because their information is not available at run-time while predicting. We stripped out predictors from our consideration which had already been converted into dummy variables or which displayed no variation in values across all the records.

The cut-off probabilities we used for all classification thresholds = 0.5

4. 1.1 Logistic Regression

We started with a total of 17 predictors for running both the logistic regression and CART models. The logistic regression was run using stepwise selection technique of the best subsets. As seen in figure 3.3 (refer appendix), the best subset were obtained for both the 9 as well as 17 predictors where the Cp ~ # of predictors in the subset. In order to keep the model simple with limited number of predictors we decided to re-run the logistic regression using subset 9 which included the following predictors.

The Regression Model

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	0.50860626	0.14276753	0.00036737	*
UniqueCount(Sub_Deparmen	-0.06632935	0.02002351	0.00092442	0.93582261
Sum(Quantity_Sold)	-0.01077881	0.00142052	0	0.98927909
Sum(Extended_Price)	0.00033176	0.0000233	0	1.00033176
Age	-0.01201673	0.00306918	0.0000903	0.98805517
Min(SEX)_M	0.29751062	0.073025	0.00004619	1.34650266
Min(CLEAN_EMAIL_FLAG)_N	-0.20087712	0.07406903	0.00668734	0.81801295
Day_4	-0.21556935	0.09954209	0.030341	0.80608237
Day_6	-0.33779919	0.10347726	0.00109666	0.71333849

Residual df	4299
Residual Dev.	5442.638184
% Success in training data	58.0547818
# Iterations used	8
Multiple R-squared	0.07120106

Figure 4.1

We also made sure we are not including predictors with high p-values, and so finalized on the logistic regression model above. The multi-variable logistic regression output would be:

logit = 0.509 - 0.066*(unique_count_sub_dept) - 0.0108*(sum_qty_sold) + 0.0003*(sum_extended_price) - 0.012*(Age) + 0.298*(min_sex_male) - 0.2009*(min_clean_email_flag) - 0.216*(Day_4) - 0.338*(Day_6)

where min_sex_male = 1 if male, 0 otherwise
min_clean_email_flag = 1 if email has been provided, 0 otherwsie
Day_4 = 1, if purchase date = Thursday, 0 otherwise
Day_6 = 1, if purchase date = Saturday, 0 otherwise

Result of the model on test data Appendix - **Figure 4.2**

4.1.2 CART

For the CART model we used the same set of 17 predictors and analyzed the key predictors using the full tree as well as the best prune models (refer appendix **Figure 4.3**). The key predictors identified were: sum_extended_price, sum_quantity_sold and unique_count_sku_number

4.1.3 Ensembles

Using a simple average of the probability predictions from the two models we computed a resulting probability prediction and compared to the threshold of 0.5 to manually classify the customer record as high-margin Vs. low-margin. A pivot table of the results is given alongside.

Error Report			
Class	# Cases	# Errors	% Error
H	1029	218	21.19
L	694	407	58.65
Overall	1723	625	36.27

Figure 4.4 Ensemble Result

4.2 Benchmark

We implemented the Naive Rule for benchmarking and found it as - error percentage less than 40% (Naïve Rule error for H→ L)

4.3 Model Selection

Since error rate for H classified as L on logistic regression was the lowest, we selected the logistic regression model as the best model to recommend to our client. With some help from the client we could improve our model selection process by assigning cost factors to the error predictions and adjusting the cutoff threshold accordingly.

5. Conclusions

The proposed model provides a good prediction of the customers (with only a record of 1 basket purchase) who are potential buyers of high margin products in the hypermarket. The advantage of this model is that it is simple and provides prediction within the acceptance standards (as per industry bench marks).

The model can be improved by improving the prediction by collecting better predictors like income levels their residential location information and behavioral data if could be captured from social media sites especially for the electronic items. Also, getting numbers on costs for incorrectly classifications to develop accurate confusion matrices and then determining the cost of sending / not sending promotions to customers.

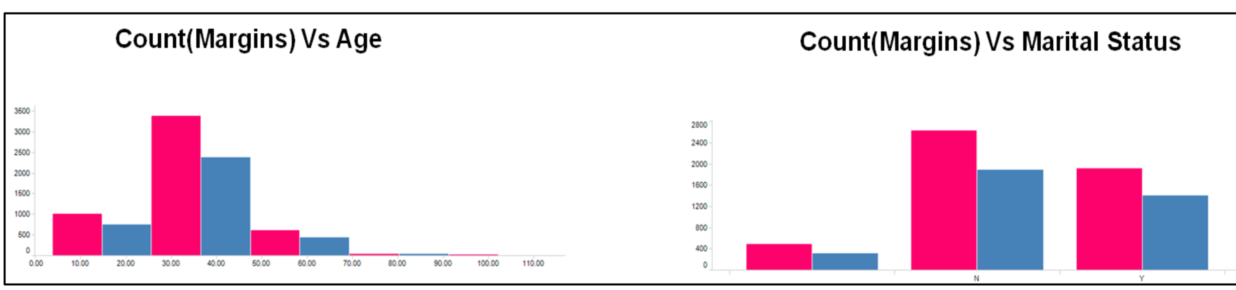
We Recommend that the hypermarket focuses on the following groups as they are more likely to be high margin customers -Males (3 times more likely),Age group b/w 25-45 (3 times more likely) amd Unmarried (1.3 times more likely). Also, execute the model in real time as the customer checks out, and give coupons if he/she is a high margin first time buyer. Don't let the model become stale. Continue to collect data periodically to refine the model in future.

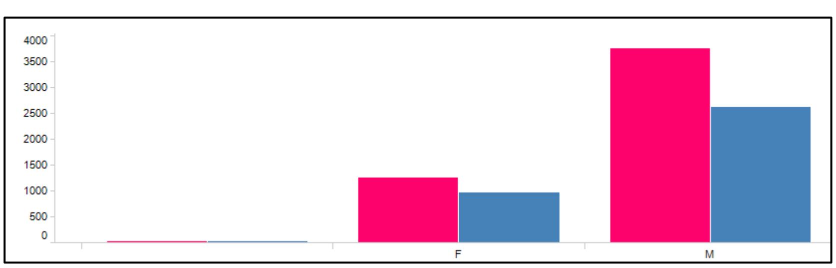
Appendix :

Figure 3.1 Final Partitioned dataset used for running models

Row Id.	Customer_No	First(CES_HEADER_KEY)	Avg(Margin_bin)	Min(DOB)	Min(STATUS)	Min(ENROLLMENT_SALE	Min(ENROLLMENT_STOR	Min(CLEAN_MOBILE_FL	Min(Transaction_Date)	UniqueCount(CES_HEA	UniqueCount(Sub_Dep	UniqueCount(Sku_Nu	Sum(Quantity_Sold)	Sum(Extended_Price)	Margin	Margin_H_L	Age
	Selected variables																
1	3000051676	02|1|20120618	0.5	23449	A	39950	1001	Y	6/18/2012	2	8	29	95.32	4001.1	1	H	48.78611111
4	3000117543	01|1|20120517	0.69	29930	A	40063	1005	Y	5/17/2012	2	5	26	77.66	3556.98	1	H	31.04444444
5	3000117717	|7|0|20120108	0.71	31962	A	40077	1005	Y	1/8/2012	2	8	17	49.96	4241.43	1	H	25.47777778
9	3000118137	|1|0|20110914	0.51	27064	A	40076	1005	Y	9/14/2011	2	7	18	126	5622	1	H	38.89444444
10	3000118335	03|0|20120429	0.81	29839	A	40076	1005	Y	4/29/2012	2	7	24	90	6613.38	1	H	31.29444444
12	3000118871	02|1|20120623	0.77	29438	A	40073	1005	Y	6/23/2012	2	5	17	57	5869.35	1	H	32.39166667
17	3000120042	11|0|20111213	0.6	24318	A	40091	1005	Y	12/13/2011	2	1	2	6	507	1	H	46.40555556
18	3000120117	04|0|20120304	0.5	26682	A	40087	1005	Y	3/4/2012	2	6	28	102	3830.88	1	H	39.93888889
21	3000120802	|1|0|20111106	0.73	28444	A	40098	1005	Y	11/6/2011	2	4	13	39	5023.47	1	H	35.11388889
26	3000121719	04|0|20120602	0.53	33198	A	40090	1005	Y	6/2/2012	2	3	9	33	1748.91	1	H	22.09722222
29	3000127989	|1|0|20111022	0.37	26212	A	40112	1005	Y	10/22/2011	2	4	7	22.45	606.36	0	L	41.22222222
30	3000128052	|1|0|20110803	0.58	26622	A	40100	1005	Y	8/3/2011	2	7	28	117	7762.44	1	H	40.10277778
35	3000128797	|2|0|20111020	0.39	23374	A	40107	1005	Y	10/20/2011	2	7	22	68.92	5484.51	0	L	48.99166667
36	3000129290	|6|0|20110907	0.67	22255	A	40106	1005	Y	9/7/2011	2	5	14	45	4328.94	1	H	52.05833333
37	3000129340	02|0|20120306	0.68	24455	A	40108	1005	Y	3/6/2012	2	10	46	165	13581.33	1	H	46.03333333
38	3000129407	03|0|20120324	0.21	25200	A	40108	1005	Y	3/24/2012	2	2	5	15	1480.5	0	L	43.99444444
39	3000129423	|1|0|20110919	0.08	15493	A	40109	1005	Y	9/19/2011	2	2	2	6.83	901.56	0	L	70.56944444
41	3000129472	|4|0|20111113	0.92	30351	A	40109	1005	Y	11/13/2011	2	4	11	33	1420.44	1	H	29.89444444
43	3000129548	05|0|20120612	0.64	26450	A	40111	1005	Y	6/12/2012	2	1	5	15	1009.5	1	H	40.57222222
44	3000129720	|2|0|20111120	0.68	26075	A	40118	1005	Y	11/20/2011	2	8	25	87	5801.94	1	H	41.59444444
45	3000129910	|3|0|20111104	0	24082	A	40102	1005	Y	11/4/2011	2	2	5	18	795	0	L	47.05555556
48	3000130348	03|1|20120520	0.53	33284	A	40125	1005	Y	5/20/2012	2	10	78	246.95	17722.68	1	H	21.86388889
49	3000130447	02|1|20120701	0.61	28122	A	40118	1005	Y	7/1/2012	2	4	8	57	3193.44	1	H	35.99444444
51	3000130611	|2|0|20111123	0.64	35739	A	40121	1005	Y	11/23/2011	2	9	24	78.85	4131.24	1	H	15.14166667
52	3000130702	|2|0|20110902	0.79	28324	A	40122	1005	Y	9/2/2011	2	7	20	79.72	3925.59	1	H	35.43888889
54	3000131189	11|0|20120106	0.44	27411	A	40126	1005	Y	1/6/2012	2	1	3	9	1275	0	L	37.94166667
55	3000131205	|3|0|20110901	0.73	29342	A	40125	1005	Y	9/1/2011	2	8	58	180.98	11217.6	1	H	32.65277778
56	3000131429	|4|0|20120120	0.77	25384	A	40125	1005	Y	1/20/2012	2	2	5	15	1704	1	H	43.48888889
58	3000131817	|5|0|20120113	0.46	19586	A	40121	1005	Y	1/13/2012	2	6	10	42	1560	0	L	59.36388889

Figure 3.2 Data visualization





Count(Margin) Vs Sex

Figure 3.3: Stepwise-run subset results

Best subset selection

	#Coeffs	RSS	Cp	Probability	Model (Constant present in all models)									
					1	2	3	4	5	6	7	8	9	10
Choose Subset	2	4444.780762	141.8166962	0	Constant	Sum(Extended_Price)	*	*	*	*	*	*	*	*
Choose Subset	3	4364.35791	63.37509918	0	Constant	Sum(Quantity_Sold)	Sum(Extended_Price)	*	*	*	*	*	*	*
Choose Subset	4	4346.550293	47.5633316	0.00000037	Constant	Sum(Quantity_Sold)	Sum(Extended_Price)	Min(SEX)_M	*	*	*	*	*	*
Choose Subset	5	4329.144043	32.15302658	0.00011896	Constant	Sum(Quantity_Sold)	Sum(Extended_Price)	Age	Min(SEX)_M	*	*	*	*	*
Choose Subset	6	4319.526367	24.53310776	0.00200443	Constant	Count(Sub_Deparment)	Sum(Quantity_Sold)	tended_Price)	Age	Min(SEX)_M	*	*	*	*
Choose Subset	7	4310.807617	17.81232643	0.02267708	Constant	Count(Sub_Deparment)	Sum(Quantity_Sold)	tended_Price)	Age	Min(SEX)_M	Day_6	*	*	*
Choose Subset	8	4303.441895	12.44488621	0.14313717	Constant	Count(Sub_Deparment)	Sum(Quantity_Sold)	tended_Price)	Age	Min(SEX)_M	AIL_FLAG)_N	Day_6	*	*
Choose Subset	9	4298.783203	9.78510952	0.36085314	Constant	Count(Sub_Deparment)	Sum(Quantity_Sold)	tended_Price)	Age	Min(SEX)_M	AIL_FLAG)_N	Day_4	Day_6	*
Choose Subset	10	4294.486816	7.48772097	0.72398841	Constant	Count(Sub_Deparment)	Sum(Quantity_Sold)	tended_Price)	Age	Min(SEX)_M	STATUS)_NA	AIL_FLAG)_N	Day_4	Day_6
Choose Subset	11	4291.832031	6.83231735	0.93373984	Constant	Count(Sub_Deparment)	Sum(Quantity_Sold)	tended_Price)	Age	Min(SEX)_M	STATUS)_NA	AIL_FLAG)_N	Day_1	Day_4
Choose Subset	12	4299.866211	16.86837006	0.07841841	Constant	Count(Sub_Deparment)	ueCount(Sku_Number)	Quantity_Sold)	tended_Price)	Age	Min(SEX)_F	Min(SEX)_M	_STATUS)_N	STATUS)_NA
Choose Subset	13	4299.140137	18.14212608	0.05920067	Constant	Count(Sub_Deparment)	ueCount(Sku_Number)	Quantity_Sold)	tended_Price)	Age	Min(SEX)_F	Min(SEX)_M	_STATUS)_N	STATUS)_NA
Choose Subset	14	4298.945801	19.94774437	0.03097295	Constant	Count(Sub_Deparment)	ueCount(Sku_Number)	Quantity_Sold)	tended_Price)	Age	Min(SEX)_F	Min(SEX)_M	_STATUS)_N	STATUS)_NA
Choose Subset	15	4298.042969	21.04470253	0.01849823	Constant	Count(Sub_Deparment)	ueCount(Sku_Number)	Quantity_Sold)	tended_Price)	Age	Min(SEX)_F	Min(SEX)_M	_STATUS)_N	STATUS)_NA
Choose Subset	16	4297.977539	22.97925758	0.00492223	Constant	Count(Sub_Deparment)	ueCount(Sku_Number)	Quantity_Sold)	tended_Price)	Age	Min(SEX)_F	Min(SEX)_M	_STATUS)_N	STATUS)_NA
Choose Subset	17	4290	16.99985886	1	Constant	Count(Sub_Deparment)	ueCount(Sku_Number)	Quantity_Sold)	tended_Price)	Age	Min(SEX)_F	Min(SEX)_M	_STATUS)_N	STATUS)_NA

Figure 4.2 : Result of multi-variable logistic regression

Figure 4.3.1 CART Results

Test Data scoring - Summary Report

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

Classification Confusion Matrix		
	Predicted Class	
Actual Class	H	L
H	826	203
L	422	272

Error Report			
Class	# Cases	# Errors	% Error
H	1029	203	19.73
L	694	422	60.81
Overall	1723	625	36.27

Test Data scoring - Summary Report (Using Best Pruned T

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

Classification Confusion Matrix		
	Predicted Class	
Actual Class	H	L
H	736	293
L	342	352

Error Report			
Class	# Cases	# Errors	% Error
H	1029	293	28.47
L	694	342	49.28
Overall	1723	635	36.85

Figure 4.3.2 CART Results